Real-Time Focus Range Sensor

Shree K. Nayar, Masahiro Watanabe, Minori Noguchi

CUCS-028-94

Department of Computer Science
Columbia University
New York, N.Y. 10027, USA

nayar@cs.columbia.edu
nabe@cs.columbia.edu
noguchi@cs.columbia.edu

Abstract

Structures of dynamic scenes can only be recovered using a real-time range sensor. Depth from defocus offers an elegant solution to fast and dense range estimation. It is computational efficient as it circumvents the correspondence problem faced by stereo and feature tracking in structure from motion. However, accurate depth estimation requires theoretical and practical solutions to a variety of problems including recovery of textureless surfaces, precise blur estimation, and magnification variations caused by defocusing. Both textured and textureless surfaces are recovered using an illumination pattern that is projected via the same optical path used to acquire images. The illumination pattern is optimized to ensure maximum accuracy and spatial resolution in computed depth. The relative blurring in two images is computed using a narrow-band linear operator that is designed by considering all the optical, sensing, and computational elements of the depth from defocus system. Defocus invariant magnification is achieved by the use of an additional aperture in the imaging optics. A prototype focus range sensor has been developed that produces upto 512x480 depth estimates at 30 Hz with an accuracy of 0.3%. Several experimental results are included to demonstrate the performance of the sensor. Applications to well-known vision problems as well as automatic CAD model generation and visualization are discussed.

1 Introduction

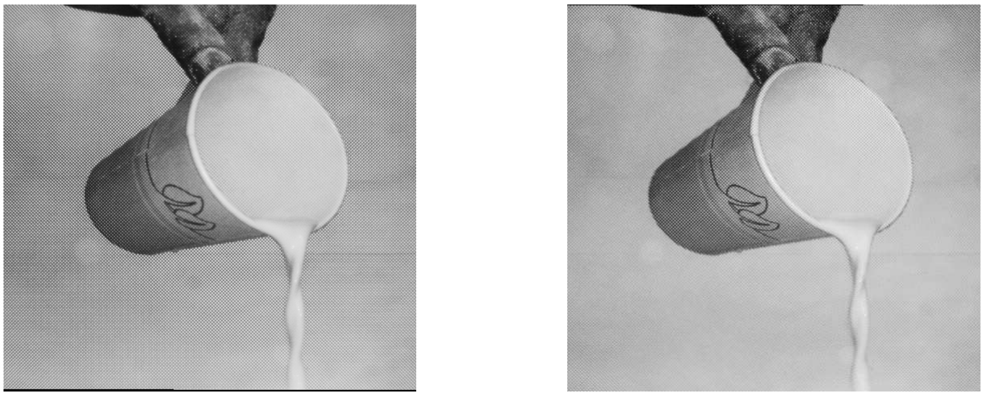
A pertinent problem in computational vision is the recovery of three-dimensional scene structure from two-dimensional images. Of all problems studied in vision, the above has by far attracted the most attention. This has resulted in a panoply of sensors and algorithms [Jarvis-1983] [Besl-1988] that can be broadly classified into two categories; passive and active. Passive techniques such as shape from shading and shape from texture attempt to extract structure from a single image. These algorithms are still under investigation and, given the assumptions they are forced to invoke, it is expected they will prove complementary to other techniques but not serve as stand-alone strategies. Other passive methods such as stereo and structure from motion use multiple views to resolve shape ambiguities inherent in a single image. The primary bottleneck for these methods has proved to be correspondence and feature tracking. In addition, passive algorithms have yet to demonstrate the accuracy and robustness required for high-level perception tasks such as object recognition and pose estimation.

Hitherto, high quality depth maps have resulted only from the use of active sensors based on time of flight or light striping [Jarvis-1983]. From a practical perspective, light stripe range finding has emerged as a clear winner. In structured environments, where active radiation of a scene is feasible, it offers a robust yet inexpensive solution to a variety of problems. However, it has suffered from one inherent drawback, namely, speed. To achieve depth maps with sufficient spatial resolution, a large number (say, N) of closely spaced stripes are used. If all stripes are projected simultaneously it is impossible to associate a unique stripe with any given image point, a process that is necessary to compute depth by triangulation. The classical approach is to obtain N images, one for each stripe. If T_f is the time required to sense and digitize an image, the scanning of N stripes takes at least $N \,.\, T_f$. Substantial improvements can be made by assigning gray codes to the stripes and scanning the entire collection of stripes in sets [Inokuchi *et al.*-1984]. All the information needed is then acquired in $log_2(N)\,T_f$, a significant improvement. An alternative approach uses color-coded stripe patterns [Boyer and Kak-1987]; this however is practical only in a gray-world that reflects all wavelengths of light. New hope for light stripe range finding has been instilled by advances in VLSI. Based on the notion of cell parallelism [Kanade *et al.*-1991], a computational sensor is developed where each sensor element records a stripe detection time-stamp as a single laser stripe sweeps the scene at high speed. Depth maps are produced in as little as 1 msec, though present day silicon packaging limits the total number of cells, and hence spatial depth resolution, to 28x32 [Gruss *et al.*-1993]. Future advances in VLSI are expected to yield high-resolution depth maps at unprecedented speeds.

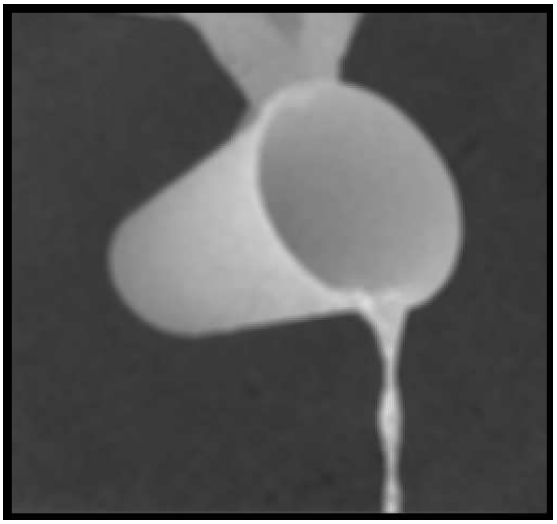
In this paper, we present a range sensor based on focus analysis that produces a 512x480 depth map at 30 Hz (video frame-rate). The sensor uses inexpensive off-the-shelf imaging and processing hardware and is shown to have an accuracy of approximately 0.3%. Focus analysis has a major advantage over stereo and structure from motion, two or more images of a scene are taken under different optical settings but from the same viewpoint, as initially demonstrated in [Pentland-1987][Subbarao-1988][Pentland *et al.*-1989]. This circumvents the need for correspondence or feature tracking. The algorithm presented here uses only two scene images. These images correspond to different levels of focus and local frequency analysis implemented typically via linear operators yields depth estimates. However, differences between the two images tend to be very subtle and we believe that previous solutions to depth from defocus have met with limited success as they are based on rough approximations to the optical and sensing mechanisms involved in focus analysis. In contrast, our approach is based on a careful physical modeling of all the optical, sensing, and computational elements at work; the optical transfer function, defocus, image sensing and sampling, and focus measure operators.

Depth from defocus shares one inherent weakness with stereo and motion, it requires that the scene have high frequency textures. A textureless surface appears the same focused or defocused and resulting images do not contain information necessary for depth computation. This has prompted us to develop a focus range sensor that uses active illumination. The key idea is to force a texture on the scene and then analyze the relative defocus of the texture in two images. Illumination projection has been suggested in the past [Girod and Scherock-1989][Pentland *et al.*-1994] for both depth from defocus and depth from pattern size distortion under perspective projection. However, these projected patterns were selected in an ad-hoc fashion and do not guarantee desired precision in computed depth. A critical problem therefore is determining an illumination pattern that would maximize the accuracy and robustness of depth from defocus. In this paper, a solution to this problem is arrived at through a detailed Fourier analysis of the entire depth from defocus system. First, theoretical models developed for each of the optical and computational elements of the system are expressed in spatial and Fourier domains. The derivation of the illumination pattern (or filter) is then posed as an optimization problem in Fourier domain. The optimal pattern is one that maximizes sensitivity of the focus measure to depth variations while minimizing the size of the focus operator to achieve high spatial resolution in computed depth.

A prototype real-time focus range sensor has been developed. It uses two CCD image detectors that view the scene through the same optical elements. The derived illumination pattern is fabricated using micro-lithography and incorporated into the sensor. The illumination pattern is projected onto the scene via the same optical path used to image the scene. This results in several advantages. It enables precise registration of the illumination pattern with the sampling grid of the image sensors. Light rays projected out through the imaging optics are subjected to similar geometric distortions as rays reflected back to the sensors. Therefore, despite ever-present lens distortions, the illumination pattern and the sensing grid of the detector are well registered. The coaxial illumination and imaging also results in a shadowless image; all surface regions that are visible to the sensor are also illuminated. Furthermore, since both images are acquired from the same viewing direction, the missing part or occlusion problem in stereo is avoided. Figure 1 shows two brightness images and the computed depth map of a cup with milk flowing out of it. Structures of such dynamic scenes can only be recovered by a high-speed sensor. Numerous experiments have been conducted to evaluate the accuracy and real-time capability of the sensor. In addition to a quantitative error analysis, real-time depth map sequences of moving objects are presented. The performance of the sensor demonstrates its relevance not only to problems such as recognition and inspection but also automatic CAD model generation for vision and graphics and remote visualization and real-time tracking of three-dimensional objects. These applications are presently being explored.



(a)



(b)

Figure 1: (a) Two images of a scene taken using different focus settings. (b) A depth map of the scene computed in 33 msec by the focus range sensor.

2 Depth from Defocus

Fundamental to depth from defocus is the relationship between focused and defocused images [Born and Wolf-1965]. Figure 2 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. For a thin lens, the relationship between the object distance d, focal length of the lens f, and the image distance d_i is given by the Gaussian lens law:

$$\frac{1}{d} + \frac{1}{d_i} = \frac{1}{f}. \tag{1}$$

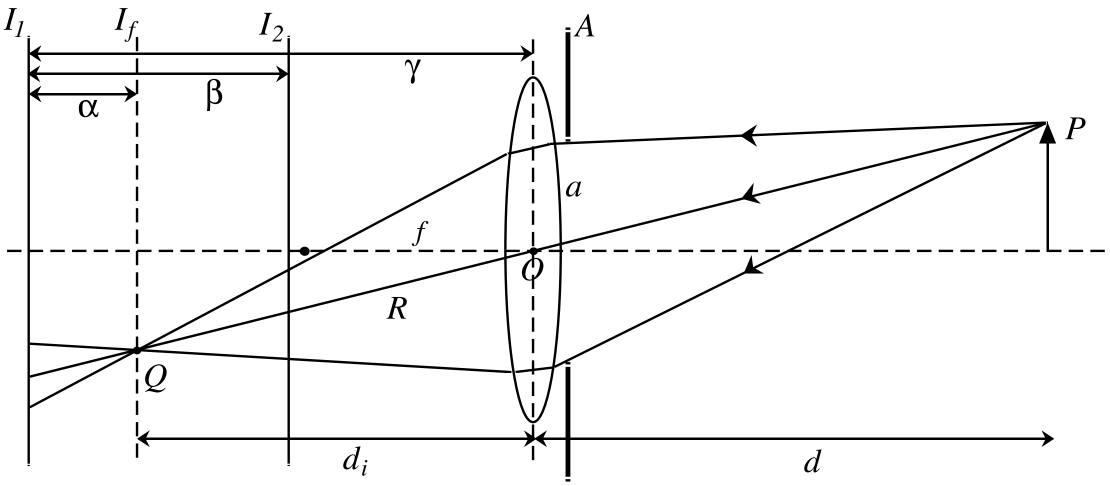


Figure 2: Image formation and depth from defocus.

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 scenes defocused to different degrees can produce identical images. A solution to depth is achieved by using two images I_1 and I_2 separated by a known physical distance β [Pentland-1987] [Subbarao and Surya-1992]. The problem is reduced to analyzing the relative blurring of each scene point in the two images and computing the distance α of its focused image. Then, using $d_i = \gamma - \alpha$, the lens law (1) yields depth d of the scene point. Simple as this procedure may appear, several technical problems emerge when implementing an algorithm of practical value.

- **Determining Relative Defocus:** In frequency domain, blurring can be viewed as low-pass filtering of scene texture. Relative blurring can thus in principle be estimated by frequency analysis. This problem is non-trivial since local scene texture includes frequencies with unknown magnitudes and phases. Since the effect of blurring is frequency dependent, it is not meaningful to investigate the net blurring of the entire collection of frequencies that constitute scene texture. This observation has forced investigators to use narrow-band filters that isolate more or less single frequencies and estimate their relative attenuation due to defocus in two or more images. Given that the dominant frequencies of the scene are unknown and possibly spatially varying, one is forced to use a large bank of tuned filter such as Gabor filters [Gokstorp-1994] [Xiong and Shafer-1994b] or hypergeometric filters [Xiong and Shafer-1994a]. Three problems surface with this approach. (a) While it is rigorous, the necessity to use scores (at times over 100) filters makes it impractical for any real-time application without the use of expensive customized hardware. (b) The filters are typically chosen by assuming the images to be continuous. Filter design for discrete images requires the analysis be carried further to avoid undesirable artifacts in computed depth. (c) Irrespective of the reliability of a filter in extracting focus measures, its output can be put to good use only if all optical and sensing elements of the depth from defocus system are accurately modeled. For instance, previous work has relied heavily on the Gaussian blur function, an approximation that may suffice for depth from focus[1] but severely limits the accuracy of depth from defocus.

- **Textureless Surfaces:** Depth from defocus shares a severe weakness with stereo and structure from motion. If the imaged surface is textureless (a white sheet of paper, for instance) defocus and focus produce identical images and any number of filters would prove ineffective in estimating relative blurring. A similar situation would arise in stereo or motion; correspondence and feature tracking would be ill-posed. Particularly in structured environments this problem can be obviated by projecting an illumination pattern on the scene of interest, i.e. forcing scene texture [Girod and Scherock-1989] [Pentland-1987]. However, careful attention must be given to the pattern that is used, else the problem is at best reduced to applying depth from defocus to a scene with unknown texture. In our work we are interested in both textured and textureless scenes and hence adopt illumination projection. In contrast to previous work, however, we seek an optimal pattern that would ensure that all scene points have the same dominant texture, one that maximizes the spatial resolution and accuracy of computed depth. Derivation of the optimal projected pattern is posed as an optimization in Fourier domain.

- **Varying Magnification:** Finally, the relation between magnification and focus is well worth mentioning. In the imaging system shown in Figure 2, the effective image location

[1] All work in focus based depth computation can be broadly classified into depth from focus and depth from defocus. The former relies on a large number of images taken by varying α in Figure 2 in small increments (or through search) and uses a focus operator to detect the image of maximum focus for each scene point (see [Krotkov-1987, Darrell and Wohn-1988, Nayar and Nakagawa-1994, Nair and Stewart-1991, Krishnan and Ahuja-1993, Asada *et al.*-1993, Xiong and Shafer-1994b, Noguchi and Nayar-1994]). In contrast, depth from defocus typically uses two images and estimates relative blurring to get depth (see [Pentland-1987, Subbarao-1988, Grossman-1987, Pentland *et al.*-1989, Bove, Jr.-1993, Ens and Lawrence-1991, Xiong and Shafer-1994b, Gokstorp-1994]).

of point P moves along ray R as the sensor plane is displaced. This causes a shift in image coordinates of P that in turn depends on the unknown scene coordinates of P. This variation in image magnification with defocus manifests as a correspondence-like problem in depth from defocus as the right set of points in images I_1 and I_2 are needed to estimate blurring. This problem has been underemphasized in previous work with the exception of [Willson and Shafer-1994] where a precise focus-magnification calibration of motorized zoom-lenses is suggested and [Darrell and Wohn-1988] where a registration-like correction in image domain is proposed. The calibration approach while effective is cumbersome and not viable for many off-the-shelf lenses. We present a simple but very effective solution based on first principles of optics.

3 Constant Magnification Defocus

We begin with the last of the problems raised in the above discussion; the variation of image magnification with defocus. We approach the problem from an optical perspective rather a computational one. Consider the image formation model shown in Figure 3. The only modification made with respect to the model in Figure 2 is the use of the external aperture A'. The aperture is placed at the *front-focal plane*, i.e. a focal length in front of the *principal point* O of the lens. This simple addition solves the prevalent problem of magnification variation with distance α of the sensor plane from the lens. Simple geometrical analysis reveals that a ray of light R' from any scene point that passes through the center O' of aperture A' emerges parallel to the optical axis on the image side of the lens [Kingslake-1983]. Furthermore, this parallel ray is the axis of a cone that includes all light rays radiated by the scene point, passed through by A' and intercepted by the lens. As a result, despite blurring, the effective image coordinates of point P in both images I_1 and I_2 are the same, namely the coordinate of its focused image Q on I_f.

This invariance of magnification to defocus holds true for any depth from defocus configuration (all values of α and β). It can also be shown that the constant-magnification property is unaffected by the aperture radius a' used. Furthermore, the lens law of (1) remains valid. This modification is realizable not only in single lens systems but any compound lens system (see [Kingslake-1983]). Given an off-the-shelf lens, such an aperture is easily appended to the casing of the lens. The resulting optical system is called a *telecentric lens*. While the nominal and effective *F-numbers* of the classical optics in Figure 2 are f/a and d_i/a, respectively, they are both equal to f/a' in the telecentric case.

4 Modeling

Effective solutions to both illumination projection and depth estimation require careful modeling and analysis of all physical phenomena involved in depth from defocus. There are five different elements, or components, that play a critical role. We briefly describe them before proceeding to model them.

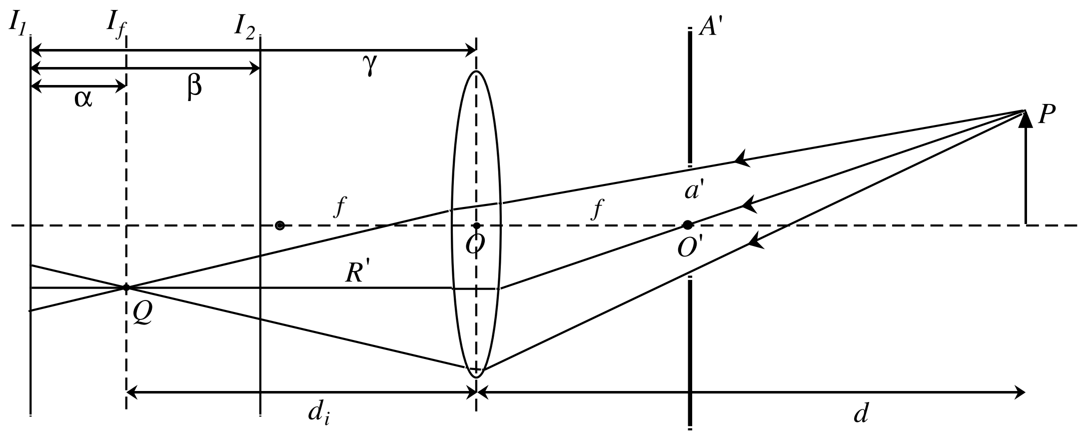


Figure 3: A constant-magnification imaging system for depth from defocus is achieved by simply placing an aperture at the front-focal plane of the optics.

1. **Illumination Pattern:** The exact pattern used to illuminate the scene determines its final texture. The spatial and frequency characteristics of this texture determine the behavior of the focus measure and hence the accuracy of depth estimation. It is the parameters of this component that we set out to optimize so as to achieve maximum depth accuracy.

2. **Optical Transfer Function:** The finite size of the lens aperture imposes restrictions on the range of spatial frequencies that are detectable by the imaging system. These restrictions play a critical role in the optimization of the illumination pattern. Upon initial inspection, the optical transfer function (OTF) seems to severely constrain the range of useful illumination patterns. However, as we shall see, the OTF's limited range also enables us to avoid serious problems such as image aliasing.

3. **Defocusing:** The depth of a surface point is directly related to its defocus (or lack of it) on the image plane. It is this phenomenon that enables us to recover depth from defocus. It is well-known that defocus is essentially a low-pass filter. However, a realistic model for this phenomenon is imperative for focus analysis. Our objective is to determine depth from two images by estimating the plane of best focus for each scene point.

4. **Image Sensing:** The two images used for shape recovery are of course discrete. The relationship between the continuous image formed on the sensor plane and the discrete image used in computations is determined by the shape and spatial arrangement of sensing elements (pixels) on the image detector. As we shall see, the final illumination pattern will include elements that are comparable to the size of each pixel on the sensor. Therefore, an accurate model for image sensing is essential for illumination optimization.

5. **Focus Operator:** The relative degree of defocus in two images is estimated by using a focus operator. Such an operator is typically a high-pass filter and is applied to discrete images. Interestingly, the optimal illumination pattern is also dependent on the parameters of the focus operator used.

All the above components together determine the the relation between the depth of a scene point and its two focus measures. Therefore, the optimal illumination is viewed as one that maximizes the sensitivity and robustness of the focus measure function. To achieve this each component is modeled in spatial as well as Fourier domains. Since we have used the telecentric lens (Figure 3) in our implementation, it's parameters are used in developing each model. However, all of the following expressions can be made valid for the classical lens system (Figure 2) by simply replacing the factor $\frac{f}{a'}$ by $\frac{d_i}{a}$.

4.1 Illumination Pattern

Before the parameters of the illumination pattern can be determined, an illumination model must be defined. Such a model must be flexible in that it must subsume a large enough variety of possible illumination patterns. In defining the model, it is meaningful to take the characteristics of the other components into consideration. As we will describe shortly, the image sensor used has rectangular pixels arranged on a rectangular spatial grid. With this in mind, we define the following illumination model. The basic building block of the model is a rectangular illuminated patch, or cell, with uniform intensity:

$$i_c(x,y) = i_c(x,y;b_x,b_y) = {}^2\Pi(\frac{1}{b_x}x, \frac{1}{b_y}y) \tag{2}$$

where, ${}^2\Pi()$ is the two-dimensional *Rectangular* function [Bracewell-1965]. The *unknown* parameters of this illumination cell are b_x and b_y, the length and width of the cell.

This cell is assumed to be repeated on a two-dimensional grid to obtain a periodic pattern. This periodicity is essential since our goal is to achieve spatial invariance in depth accuracy, i.e. all image regions, irrespective of their distance from each other, must possess the same textural characteristics. The periodic grid is defined as:

$$i_g(x,y) = i_g(x,y;t_x,t_y) = {}^2\mathrm{III}(\tfrac{1}{2}(\tfrac{1}{t_x}x + \tfrac{1}{t_y}y), \tfrac{1}{2}(\tfrac{1}{t_x}x - \tfrac{1}{t_y}y)) \tag{3}$$

where, ${}^2\mathrm{III}()$ is the 2-dimensional *Shah* function [Bracewell-1965], and $2t_x$ and $2t_y$ determine the periods of the grid in the x and y directions. Note that this grid is not rectangular but has vertical and horizontal symmetry on the x-y plane. The final illumination pattern $i(x,y)$ is obtained by convolving the cell $i_c(x,y)$ with the grid $i_g(x,y)$:

$$i(x,y) = i(x,y;b_x,b_y,t_x,t_y) = i_c(x,y) * i_g(x,y) \tag{4}$$

The exact pattern is therefore determined by four parameters, namely, b_x, b_y, t_x and t_y. The above illumination grid is not as restrictive as it may appear upon initial inspection. For instance, the parameters b_x, b_y, $2t_x$ and $2t_y$ can each be stretched to obtain repeated illumination and non-illumination stripes in the horizontal and vertical directions, respectively. Alternatively, they can also be adjusted to obtain a checkerboard illumination pattern with large or small illuminated patches. The exact values for b_x, b_y, t_x and t_y will be evaluated by the optimization procedure described later. In practice, the illumination pattern determined by the optimization is used to fabricate a filter with the same pattern.

The optimization procedure requires the analysis of each component of the system in spatial domain as well as frequency domain (u, v). The Fourier transforms of the illumination cell, grid, and pattern are denoted as $I_c(u, v)$, $I_g(u, v)$, and $I(u, v)$, respectively, and found to be:

$$I_c(u,v) = I_c(u,v;b_x,b_y) = b_x \frac{\sin(\pi b_x u)}{\pi b_x u} \cdot b_y \frac{\sin(\pi b_y v)}{\pi b_y v} \tag{5}$$

$$I_g(u,v) = I_g(u,v;t_x,t_y) = {}^2\mathrm{III}((t_x u + t_y v), (t_x u - t_y v)) \tag{6}$$

$$I(u,v) = I(u,v;b_x,b_y,t_x,t_y) = I_c(u,v) \cdot I_g(u,v) \tag{7}$$

4.2 Optical Transfer Function

Adjacent points on the illuminated surface reflect light waves that interfere with each other to produce diffraction effects. The angle of diffraction increases with the spatial frequency of surface texture. Since the lens aperture of the imaging system (Figure 3) is of finite radius a', it does not capture the higher order diffractions radiated by the surface (see [William and Becklund-1989] for details). This effect places a limit on the optical resolution of the imaging system, which is characterized by the optical transfer function (OTF):

$$\begin{aligned} O(u,v) &= O(u,v;a',f) \\ &= \begin{cases} (\frac{a'}{f})^2(\gamma - \sin\gamma)\ , & \sqrt{u^2+v^2} \le \frac{2a'}{\lambda f} \\ 0\ , & \sqrt{u^2+v^2} > \frac{2a'}{\lambda f} \end{cases} \\ &\quad \text{where } \gamma = 2\cos^{-1}(\frac{\lambda f}{a'} \frac{\sqrt{u^2+v^2}}{2})\ . \end{aligned} \tag{8}$$

where, (u, v) is the spatial frequency of the two-dimensional surface texture as seen from the image side of the lens, f is the focal length of the lens, and λ is the wavelength of incident light. It is clear from the above expression that only spatial frequencies below the limit $\frac{2a'}{\lambda f}$ will be imaged by the optical system (Figure 4). This in turn places restrictions on the frequency of the illumination pattern. Further, the above frequency limit can be used to "cut off" any desired number of higher harmonics produced by the illumination pattern. In short, the OTF is a curse and a blessing; it limits the detectable range of frequencies and at the same time can be used to minimize the detrimental effects of aliasing and high-order harmonics.

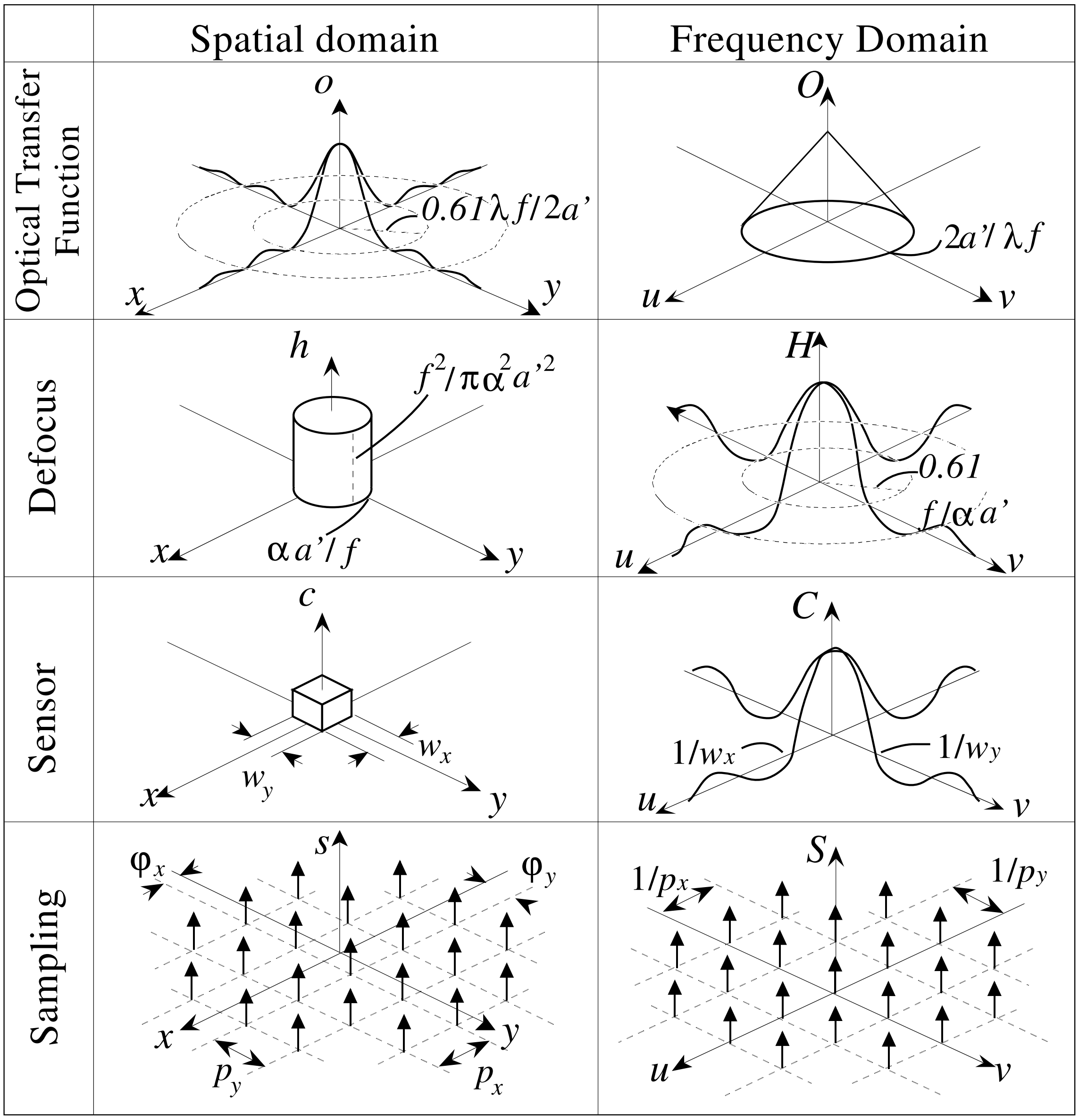


Figure 4: Spatial and frequency models for the optical and sensing elements of depth from defocus.

4.3 Defocusing

The defocus function is described in detail in previous work (see [Born and Wolf-1965] [Horn-1986], for example). As in Figure 3, let α be the distance between the focused image of a surface point and its defocused image formed on the sensor plane. The light energy radiated by the surface point and collected by the imaging optics is uniformly distributed over a circular patch on the sensor plane. This patch, also called the *pillbox*, is the defocus function (Figure 4):

$$h(x,y) = h(x,y;\alpha,a',f) = \frac{f^2}{2\pi a'^2\alpha^2}\Pi(\frac{d}{2a\alpha}\sqrt{x^2+y^2}) \tag{9}$$

where, once again, a' is the radius of the telecentric lens aperture. In Fourier domain, the above defocus function is given by:

$$H(u,v) = H(u,v;\alpha,a',f) = \frac{f}{2\pi a'\alpha\sqrt{u^2+v^2}}J_1(\frac{2\pi a'\alpha}{f}\sqrt{u^2+v^2}) \tag{10}$$

where J_1 is the first-order Bessel function [Born and Wolf-1965]. As is evident from the above expression, defocus serves as a low-pass filter. The bandwidth of the filter increases as α decreases, i.e. as the sensor plane gets closer to the plane of focus. In the extreme case of $\alpha = 0$, $H(u,v)$ passes all frequencies without attenuation producing a perfectly focused image. Note that in a defocused image, all frequencies are attenuated at the same time. In the case of passive depth from focus or defocus, this poses a serious problem; different frequencies in an unknown scene are bound to have different (and unknown) magnitudes and phases. It is difficult therefore to estimate the degree of defocus of an image region without the use of a large set of narrow-band focus operators that analyze each frequency in isolation. This again indicates that it would be desirable to have an illumination pattern that has a single dominant frequency, enabling robust estimation of defocus and hence depth.

4.4 Image Sensing

We assume the image sensor to be a typical CCD TV camera. Such a sensor can be modeled as a rectangular array of rectangular sensing elements (pixels). The quantum efficiency [Horn-1968] of each pixel is assumed to be uniform over the area of the pixel. Let $m(x,y)$ be the continuous image formed on the sensor plane. The finite pixel area has the effect of averaging the continuous image $m(x,y)$. In spatial domain, the averaging function is the rectangular cell:

$$s_c(x,y) = s_c(x,y;w_x,w_y) =^2 \Pi(\frac{1}{w_x}x, \frac{1}{w_y}y) \tag{11}$$

where, w_x and w_y are the length and width of the pixel, respectively. The discrete image is obtained by sampling the convolution of $m(x,y)$ with $s_c(x,y)$. This sampling function is a rectangular grid:

$$s_g(x,y) = s_g(x,y;p_x,p_y,\varphi_x,\varphi_y) = \frac{1}{p_x p_y}{}^2\mathrm{III}(\tfrac{1}{p_x}(x-\varphi_x), \tfrac{1}{p_y}(y-\varphi_y)) \tag{12}$$

where, p_x and p_y are spacings between discrete samples in the two spatial dimensions, and (φ_x, φ_y) is phase shift of the grid. The final discrete image is therefore:

$$m_d(x, y) = (s_c(x, y) * m(x, y)) \cdot s_g(x, y) \tag{13}$$

The parameters w_x, w_y, p_x, and p_y are all determined by the particular image sensor used. These parameters are therefore known and their values are substituted after the optimization is done. On the other hand, the phase shift (φ_x, φ_y) of the sampling function is with respect to the illumination pattern and will also be viewed as illumination parameters during optimization. To recognize the importance of these phase parameters one can visualize the variations in a discrete image that arise from simply translating a high-frequency illumination pattern with respect to the sensing grid.

In Fourier domain, the above averaging and sampling functions are:

$$S_c(u, v) = S_c(u, v; w_x, w_y) = w_x \frac{\sin(\pi w_x u)}{\pi w_x u} \cdot w_y \frac{\sin(\pi w_y v)}{\pi w_y v} \tag{14}$$

$$\begin{aligned} S_g(u, v) &= S_g(u, v; p_x, p_y, \varphi_x, \varphi_y) \\ &= {}^2\mathrm{III}(p_x u, p_y v) \cdot e^{-i2\pi(\varphi_x u + \varphi_y v)} \end{aligned} \tag{15}$$

The final discrete image is:

$$M_d(u, v) = (S_c(u, v) \cdot M(u, v)) * S_g(u, v) \tag{16}$$

4.5 Focus Operator

Since defocusing has the effect of suppressing high-frequency components in the focused image, it is desirable that the focus operator respond to high frequencies in the image. For the purpuse of illumination optimization, we use the Laplacian. However, the derived pattern will remain optimal for a large class of symmetric focus operators. In spatial domain, the discrete Laplacian is:

$$\begin{aligned} l(x, y) &= l(x, y; q_x, q_y) \\ &= 4\delta(x) \cdot \delta(y) - [\delta(x) \cdot \delta(y - q_y) + \delta(x) \cdot \delta(y + q_y) \\ &\quad + \delta(x - q_x) \cdot \delta(y) + \delta(x + q_x) \cdot \delta(y)] \end{aligned} \tag{17}$$

Here, q_x and q_y are the spacings between neighboring elements of the discrete Laplacian kernel. In the optimization, these spacings will be related to the illumination parameters. The Fourier transform of the discrete Laplacian is:

$$\begin{aligned} L(u, v) &= L(u, v; q_x, q_y) \\ &= 2(1 - \cos(2\pi q_x u)) * \delta(u) + 2(1 - \cos(2\pi q_y v)) * \delta(v) \\ &= 4 - 2\cos(2\pi q_x u) - 2\cos(2\pi q_y v) \end{aligned} \tag{18}$$

The required discrete nature of the focus operator comes with a price. It tends to broaden the bandwidth of the operator. Once the pattern has been determined, the above filter will be tuned to maximize sensitivity to the fundamental illumination frequency while minimizing the effects of spurious frequencies caused either by the scene's inherent texture or image noise.

4.6 Focus Measure

The focus measure is simply the output of the focus operator. It is related to defocus α (and hence depth d) via all of the components modeled above. Note that the illumination pattern $(i_c * i_g)$ is projected through optics that is similar to that used for image formation. Consequently, the pattern is also subjected to the limits imposed by the optical transfer function o and the defocus function h. Therefore, the texture projected on the scene is:

$$i(x, y; b_x, b_y, t_x, t_y) * o(x, y; a', f) * h'(x, y; \alpha', a', f) \tag{19}$$

where, α' represents defocus of the illumination itself that depends on the depth of the illuminated point. However, the illumination pattern once incident on a surface patch plays the role of surface texture and hence defocus α' of illumination does not have any significant effect on depth estimation. The projected texture is reflected by the scene and projected by the optics back onto the image plane to produce the discrete image:

$$\begin{aligned} &\{i(x, y; b_x, b_y, t_x, t_y) * o(x, y; a', f)^{*2} * h'(x, y; \alpha', a', f) * h(x, y; \alpha, a', f) \\ &\quad * s_c(x, y; w_x, w_y)\} \cdot s_g(x, y; p_x, p_y, \varphi_x, \varphi_y) \end{aligned} \tag{20}$$

where, $o^{*2} = o * o$. The final focus measure function $g(x, y)$ is the result of applying the discrete Laplacian to the above discrete image:

$$\begin{aligned} g(x, y) &= \{(i(x, y; b_x, b_y, t_x, t_y) * o(x, y; a', f)^{*2} \\ &\quad * h'(x, y; \alpha', a', f) * h'(x, y; \alpha, a', f) * s_c(x, y; w_x, w_y)) \\ &\quad \cdot s_g(x, y; p_x, p_y, \varphi_x, \varphi_y)\} * l(x, y; q_x, q_y) \\ &= \{(i * o^{*2} * h^{*2} * s_c) \cdot s_g\} * l \end{aligned} \tag{21}$$

Since the distance between adjacent weights of the Laplacian kernel must be integer multiples of the period of the image sampling function s_g, the above expression can be rearranged as:

$$\begin{aligned} g(x, y) &= (i * o^{*2} * h' * h * s_c * l) \cdot s_g \\ &= g_0 \cdot s_g \end{aligned} \tag{22}$$

where, $g_0 = i * o^{*2} * h' * h * s_c * l$. The same can be expressed in Fourier domain as:

$$\begin{aligned} G(u, v) &= (I \cdot O^2 \cdot H' \cdot H \cdot S_c \cdot L) * S_g \\ &= G_0 * S_g \end{aligned} \tag{23}$$

The above expression gives us the final output of the focus operator for any value of the defocus parameter α. It will be used in the following section to determine the optimal illumination pattern.

5 Optimization

In our implementation, the illumination pattern is projected on the scene using a high power light source and a telecentric lens identical to the one used to image the scene. This allows us to assume that the projected illumination is the primary cause for surface texture and is stronger than the natural texture of the surface. As a result our results are applicable not only to textureless surfaces but also textured ones. The illumination optimization problem is formulated as follows: Establish closed-form relationships between the illumination parameters (b_x, b_y, t_x, t_y), sensor parameters $(w_x, w_y, p_x, p_y, \varphi_x, \varphi_y)$, and discrete Laplacian parameters (q_x, q_y) so as to maximize the sensitivity, robustness, and spatial resolution of the focus measure $g(x, y)$. High sensitivity implies that a small variation in the degree of focus results in a large variation in $g(x, y)$. This would ensure high depth estimation accuracy in the presence of image noise, i.e. high signal-to-noise ratio. By robustness we mean that all pixels with the same degree of defocus produce the same focus measure independent of their location on the image plane. This ensures that depth estimation accuracy is invariant to location on the image plane. Lastly, high spatial resolution is achieved by minimizing the size of the focus operator. This ensures that rapid depth variations (surface discontinuities) can be detected with high accuracy.

In order to minimize smoothing effects and maximize spatial resolution of computed depth, the support (or span) of the discrete Laplacian must be as small as possible. This in turn requires the frequency of the illumination pattern be as high as possible. However, the optical transfer function described in section 4.2 imposes limits on the highest frequency that can be imaged by the optical system. This maximum allowable frequency is $\frac{2a'}{\lambda f}$, determined by the numerical aperture of the telecentric lens. With this in mind, let us examine the Fourier transform of the illumination pattern. Since the pattern is periodic, its Fourier transform must be discrete. It may have a zero-frequency component, but this can be safely ignored since the Laplacian operator, being a sum of second-order derivatives, will eventually remove any zero-frequency component in the final image. Our objective then is to maximize the fundamental spatial frequency $(1/t_x, 1/t_y)$ of the illumination pattern. In order to maximize this frequency while maintaining high detectability, we must have $\sqrt{(1/t_x)^2 + (1/t_y)^2}$ close to the optical limit $\frac{2a'}{\lambda f}$. This in turn pushes all higher frequencies in the illumination pattern outside the optical limit. What we are left with is a surface texture whose image has only the quadrapole fundamental frequencies $(\pm 1/t_x, \pm 1/t_y)$. As a result, these are the only frequencies we need consider in our analysis of the focus measure function $G(u, v)$.

Before we consider the final measure $G(u, v)$, we examine $G_0(u, v)$ the focus measure prior to image sampling. For the reasons given above, the two-dimensional $G_0(u, v)$ is reduced to four discrete spikes at $(1/t_x, 1/t_y)$, $(1/t_x, -1/t_y)$, $(-1/t_x, 1/t_y)$ and $(-1/t_x, -1/t_y)$. Since all components (I, O, H, S_c and L) of G_0 are reflection symmetric about $u = 0$ and $v = 0$, we have:

$$G_0(\frac{1}{t_x}, \frac{1}{t_y}) = G_0(\frac{1}{t_x}, -\frac{1}{t_y}) = G_0(-\frac{1}{t_x}, \frac{1}{t_y}) = G_0(-\frac{1}{t_x}, -\frac{1}{t_y}) \quad (24)$$

where:

$$G_0(\frac{1}{t_x}, \frac{1}{t_y}) = I(\frac{1}{t_x}, \frac{1}{t_y}; b_x, b_y, t_x, t_y) \cdot O^2(\frac{1}{t_x}, \frac{1}{t_y}; a', f) \quad (25)$$

$$\cdot H'(\frac{1}{t_x},\frac{1}{t_y};\alpha',a',f)\cdot H(\frac{1}{t_x},\frac{1}{t_y};\alpha,a',f)$$

$$\cdot S_c(\frac{1}{t_x},\frac{1}{t_y};w_x,w_y)\cdot L(\frac{1}{t_x},\frac{1}{t_y};q_x,q_y)\cdot$$

Therefore, in frequency domain the focus measure function prior to image sampling reduces to:

$$\begin{aligned} G_0(u,v) &= G_0(\frac{1}{t_x},\frac{1}{t_y}) \\ &\cdot\{\delta(u-\frac{1}{t_x},v-\frac{1}{t_y})+\delta(u+\frac{1}{t_x},v-\frac{1}{t_y}) \\ &+\delta(u-\frac{1}{t_x},v+\frac{1}{t_y})+\delta(u+\frac{1}{t_x},v+\frac{1}{t_y})\} \end{aligned} \tag{26}$$

The function $g_0(x,y)$ in image domain, is simply the inverse Fourier transform of $G_0(u,v)$:

$$g_0(x,y) = G_0(\frac{1}{t_x},\frac{1}{t_y})\cdot\{4\cos 2\pi\frac{1}{t_x}x\cdot\cos 2\pi\frac{1}{t_y}y\} \tag{27}$$

Note that $g_0(x,y)$ is the product of cosine functions weighted by the coefficient $G_0(1/t_x,1/t_y)$. The defocus function h has the effect of reducing the coefficient $G_0(1/t_x,1/t_y)$ in the focus measure $g_0(x,y)$. Clearly, the sensitivity of the focus measure to depth (or defocus) is optimized by maximizing the coefficient $G_0(1/t_x,1/t_y)$ with respect to the unknown parameters of the system. This optimization procedure can be summarized as:

$$\frac{\partial}{\partial t_x}G_0(\frac{1}{t_x},\frac{1}{t_y})=0,\quad \frac{\partial}{\partial t_y}G_0(\frac{1}{t_x},\frac{1}{t_y})=0, \tag{28}$$

$$\frac{\partial}{\partial b_x}G_0(\frac{1}{t_x},\frac{1}{t_y})=0,\quad \frac{\partial}{\partial b_y}G_0(\frac{1}{t_x},\frac{1}{t_y})=0, \tag{29}$$

$$\frac{\partial}{\partial q_x}G_0(\frac{1}{t_x},\frac{1}{t_y})=0,\quad \frac{\partial}{\partial q_y}G_0(\frac{1}{t_x},\frac{1}{t_y})=0, \tag{30}$$

Since t_x and t_y show up in all the components in (25), the first two partial derivatives (equation (28)) are difficult to evaluate. Fortunately, the derivatives in (29) and (30) are sufficient to obtain relations between the system parameters. For details of the optimization procedure, we refer the reader to Appendix A. The following result maximizes sensitivity and spatial resolution of the focus measure $g(x,y)$:

$$b_x=\frac{1}{2}t_x,\quad b_y=\frac{1}{2}t_y \tag{31}$$

$$q_x=\frac{1}{2}t_x,\quad q_y=\frac{1}{2}t_y \tag{32}$$

Next, we examine the spatial robustness of $g(x,y)$. Imagine the imaged surface to be planar and parallel to the image sensor. Then, we would like the image sampling to produce the same absolute value of $g(x,y)$ at all discrete sampling points on the image. This entails relating the illumination and sensing parameters so as to facilitate careful sampling of the product of cosine functions in (27). Note that the final focus measure is:

$$g(x,y) = g_0 \cdot s_g = G_0(\frac{1}{t_x}, \frac{1}{t_y}) \cdot \{4 \cos 2\pi \frac{1}{t_x} x \cdot \cos 2\pi \frac{1}{t_y} y\} \\ \cdot {}^2\mathrm{III}(\tfrac{1}{p_x}(x - \varphi_x), \tfrac{1}{p_y}(y - \varphi_y)) \tag{33}$$

All samples of $g(x,y)$ have the same *absolute* value when the two cosines in the above expression are sampled at their peak values. Such a sampling is possible when:

$$p_x = \frac{1}{2} t_x, \quad p_y = \frac{1}{2} t_y \tag{34}$$

and:

$$\varphi_x = 0, \quad \varphi_y = 0 \tag{35}$$

Alternatively, the cosines can be sampled with a period of $\pi/2$ and phase shift of $\pi/4$. This yields the second solution:

$$p_x = \frac{1}{4} t_x, \quad p_y = \frac{1}{4} t_y, \tag{36}$$

$$\varphi_x = \pm \frac{1}{8} t_x, \quad \varphi_y = \pm \frac{1}{8} t_y. \tag{37}$$

The above equations give two solutions, both are checkerboard illumination patterns but differ in their fundamental frequencies, size of the illumination cell, and the phase shift with respect to the image sensor. Equations (31), (32), (34), (35) yield the filter pattern shown in Figure 5(a). In this case the filter and detector are registered with zero phase shift, and the illumination cell has the same size and shape as the sensor elements (pixels). The second solution, shown in Figure 5(b), is obtained using the sampling solutions (36) and (37), yielding a filter pattern with illumination cell two times the size of the sensor element and phase shift of half the sensor element size. Exactly how such patterns can be projected and perfectly registered with the image detector will be described in the experimental section.

6 Tuned Focus Operator

For the purpose of illumination optimization, we used the Laplacian operator. The resulting illumination pattern has only a single dominant absolute frequency, $(1/t_x, 1/t_y)$. Given this, we are in a position to further refine our focus operator so as to minimize the effects of all other frequencies caused either by the physical texture of the scene or image noise. To this end, let us consider the properties of the 3x3 discrete Laplacian (see Figure 6(a) and (b)). We see that

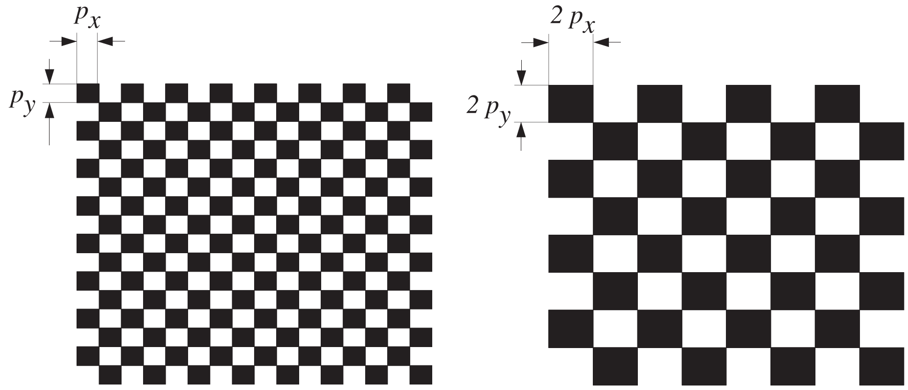


Figure 5: Optimal illumination filter patterns: (a) $t_x = 2p_x, t_y = 2p_y, \varphi_x = 0, \varphi_y = 0$; and (b) $t_x = 4p_x, t_y = 4p_y, \varphi_x = 1/8t_x, \varphi_y = 1/8t_y$. Here, (t_x, t_y) is the illumination period, (p_x, p_y) is the pixel size, and (φ_x, φ_y) is the illumination phase shift with respect to the image sensing grid.

though the Laplacian does have peaks exactly at $(1/t_x, 1/t_y)$, $(1/t_x, -1/t_y)$, $(-1/t_x, 1/t_y)$, and $(-1/t_x, -1/t_y)$, it has a fairly broad bandwidth allowing other spurious frequencies to contribute to the focus measure G in (23). Here, we seek a narrow band operator with sharp peaks at the above four coordinates in frequency space.

Given that the operator must eventually be discrete and of finite support, there is a limit to the extent to which it can be tuned. To constrain the problem, we impose the following conditions. (a) To maximize spatial resolution in computed depth we force the operator kernel to be 3x3. (b) Since the fundamental frequency of the illumination pattern has a symmetric quadrapole arrangement, the focus operator must be rotationally symmetric. These two conditions force the operator to have the structure shown in Figure 6(c). (c) The operator must not respond to any DC component in image brightness. This last condition is satisfied if the sum of all elements of the operator equals zero:

$$a + 4b + 4c = 0 \tag{38}$$

It is also imperative that the response $L(u, v)$ of the operator to the fundamental frequency not be zero:

$$L(\frac{1}{t_x}, \frac{1}{t_y}) = a + 2\,b\,(\cos 2\pi q_x \frac{1}{t_x} + \cos 2\pi q_y \frac{1}{t_y}) + 4\,c\,\cos 2\pi q_x \frac{1}{t_x}\,\cos 2\pi q_y \frac{1}{t_y} \neq 0 \tag{39}$$

Given (32), the above reduces to:

$$a - 4b + 4c \neq 0 \tag{40}$$

Expressions (38) and (40) imply that $b \neq 0$. Without loss of generality, we set $b = -1$. Hence, (38) gives $a = 4(1 - c)$. Therefore, the tuned operator is determined by a single unknown parameter, c, as shown in Figure 6(d). The problem then is to find c such that the operator's Fourier transform has a sharp peak at $(1/t_x, 1/t_y)$. A rough measure of sharpness is given by the

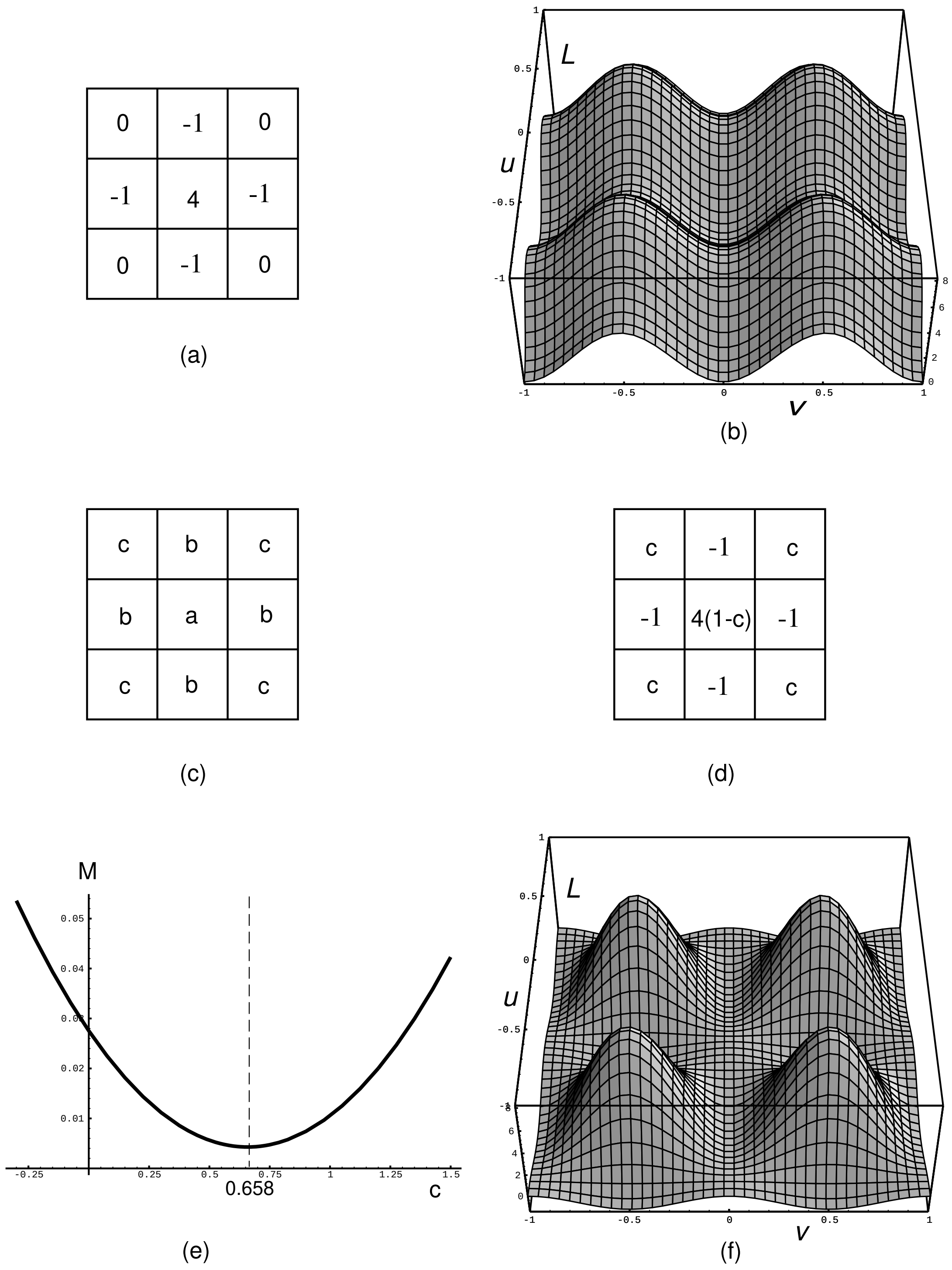


Figure 6: (a) The 3x3 Laplacian and its (b) Fourier transform. (c) The kernel structure for a 3x3 operator that is symmetric. (d) The kernel of a 3x3 operator that is insensitive to the zero-frequency component (see text).(e) The second moment M of each of the four operator peaks is minimized when $c = 0.658$. (f) Response of the tuned focus operator ($c = 0.658$) has much sharper peaks than the Laplacian.

second-order moment of the power $\| L(u,v) \|^2$ with respect to $(1/t_x, 1/t_y)$:

$$M = \frac{1}{\| L(\frac{1}{t_x}, \frac{1}{t_y}) \|^2} \int_{u=0}^{\frac{2}{t_x}} \int_{v=0}^{\frac{2}{t_y}} [(u - \frac{1}{t_x})^2 + (v - \frac{1}{t_y})^2] \; \| L(u - \frac{1}{t_x}, v - \frac{1}{t_y}) \|^2 \, dv \, du \quad (41)$$

$$= \frac{t_x^2 + t_y^2}{768\,\pi^2\, t_x^3\, t_y^3} (20\pi^2 c^2 + 6c^2 + 48c - 32\pi^2 c + 20\pi^2 - 93)$$

The above measure is minimized when $\frac{\partial M}{\partial c} = 0$, i.e. when $c = 0.658$ as shown in Figure 6(e). The resulting tuned focus operator has the response shown in Figure 6(f), it has substantially sharper peaks than the discrete Laplacian. Given that the operator is 3x3 and discrete, the sharpness of the peaks is limited. The above derivation brings to light the fundamental difference between designing tuned operators in continuous and discrete domains. In general, an operator that is deemed optimal in continuous domain is most likely sub-optimal for discrete images.

7 Depth from Two Images

Depth estimation uses two images of the scene $I_1(x,y)$ and $I_2(x,y)$ that correspond to different effective focal lengths as shown in Figure 3. Depth of each scene point is determined by estimating the displacement α of the focused plane I_f for the scene point. The tuned focus operator is applied to both images to get focus measure images $g_1(x,y)$ and $g_2(x,y,)$. From (33) we see that:

$$\frac{g_1(x,y)}{g_2(x,y)} = \frac{G_0(\frac{1}{t_x}, \frac{1}{t_y}, \alpha)}{G_0(\frac{1}{t_x}, \frac{1}{t_y}, \alpha - \beta)} \quad (42)$$

From (23) we see that the only factor in G_0 affected by parameter α is defocus function H. Therefore:

$$\frac{g_1(x,y)}{g_2(x,y)} = \frac{H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha)}{H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha - \beta)} \quad (43)$$

Note that the above measure is not bounded. This poses a problem from a computational viewpoint which is easily remedied by using the following normalization:

$$q(x,y) = \frac{g_1(x,y) - g_2(x,y)}{g_1(x,y) + g_2(x,y)} = \frac{H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha) - H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha - \beta)}{H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha) + H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha - \beta)} \quad (44)$$

As shown in Figure 7, q is a monotonic function of α such that $-p \le q \le p$, $p \le 1$. In practice, the above relation can be pre-computed and stored as a look-up table that maps q computed at each image point to a unique α. Since α represents the position of the focused image, the lens law (1) yields the depth d of the corresponding scene point. Note that the tuned focus operator designed in the previous section is a linear filter, making it feasible to compute depth maps of scenes in real-time using simple image processing hardware.

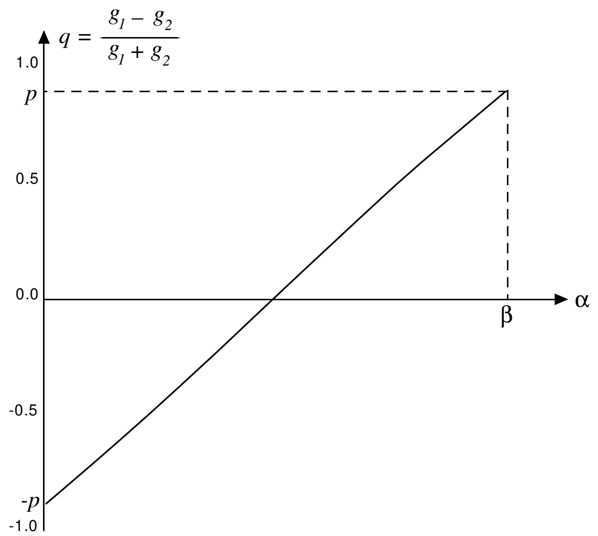


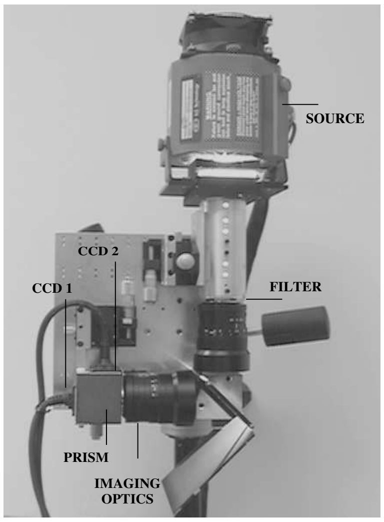
Figure 7: Relation between focus measures g_1 and g_2 and the defocus parameter α.

8 Real Time Range Sensor

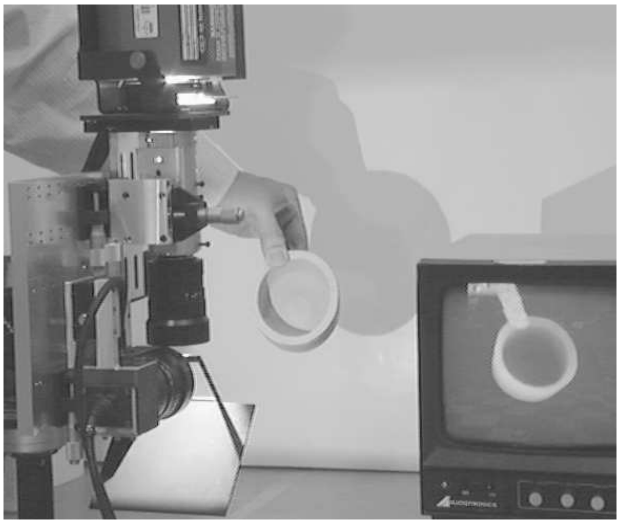
Based on the above results, we have implemented the real-time focus range sensor shown in Figure 8. The scene is imaged using a standard 12.5 mm Fujinon lens with an additional aperture added to convert it to telecentric. Light rays passing through the lens are split in two directions using a beam-splitting prism. This produces two images that are simultaneously detected using two Sony XC-77RR 8-bit CCD cameras. The positions of the two cameras are precisely fixed such that one obtains a near-focus image while the other a far-focus image. In this setup a physical displacement of 0.25mm between the effective focal lengths of the two CCD cameras translates to a sensor depth of field of approximately 30 cms. This detectable range of the sensor can be varied either by changing the sensor displacement or the focal length of the imaging optics.

The illumination pattern shown in Figure 5(b) was etched on a glass plate using microlithography, a process widely used in VLSI. The filter was then placed in the path of a 300 W Xenon arc lamp. The illumination pattern generated is projected using a telecentric lens identical to the one used for image formation. A half-mirror is used to ensure that the illumination pattern projects onto the scene via the same optical path used to acquire images. As a result, the pattern is almost perfectly registered with respect to the pixels of the two CCD cameras. Furthermore, the above arrangement ensures that every scene point that is visible to the sensor is also illuminated by it, avoiding shadows and thus undetectable regions. If objects in the scene have a strong specular reflection component, cross-polarized filters can be attached to the illumination and imaging lens to filter out specularities and produce images that mainly include the diffuse reflection component.

Images from the two CCD cameras are digitized and processed using MV200 Datacube image processing hardware. The present configuration includes the equivalent of two 8-bit digitizers, two A/D convertors, and one 12-bit convolver. This hardware enables simultaneous digitization of the two images, convolution of both images with the tuned focus operator, and the computation of a 256x240 depth map, all within a single frametime of 33 msec with a lag of 33 msec. A look-up table is used to map each pair of focus measures (g_1 and g_2) to a unique depth estimate d (see [Authors-1994] for imlpementation details). Alternatively, a 512x480 depth map can be computed at the same rate if the two images are taken in succession. Simultaneous image acquisition is clearly advantageous since it makes the sensor less sensitive to variations in both illumination and scene structure between frames. With minor additions to the present processing hardware, it is easy to obtain 512x480 depth maps at 30 Hz using simultaneous image grabbing. Depth maps produced by the sensor are visualized as wireframes at framerate on a DEC Alpha workstation.



(a)



(b)

Figure 8: (a) The real-time focus range sensor and its key components. (b) The sensor can produce depth maps up to 512x480 in resolution at 30 Hz.

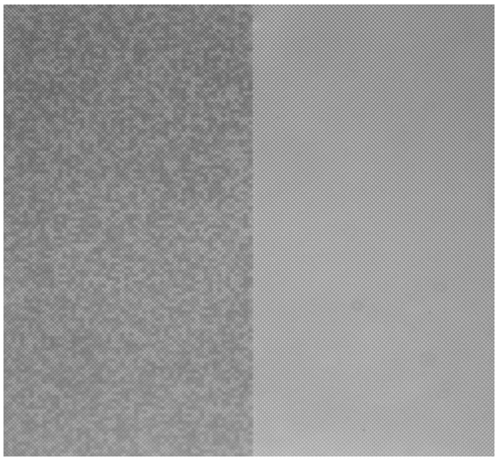
9 Experiments

Numerous experiments have been conducted to test the performance of the sensor. Here we briefly summarize these results. Figure 9(a) shows near and far focused images of a planar surface, half of the surface is textureless while the other half has strong random texture. A computed depth map of the surface is shown in Figure 9(b). As expected the textureless area is estimated almost free of errors while the textured area has small errors due to texture frequencies that lie close to the illumination frequency. It may be noted that the texture used in this experiment includes a wide spectrum of frequencies. Most scenes have weaker textures and can be expected to produce even more accurate results. Several depth maps of the plane in Figure 9(a) were computed by varying its position in the 30cm workspace of the sensor and the average accuracy and repeatability of the sensor were estimated for both simultaneous and successive image grabbing configurations (see Table 9(c)). These results clearly demonstrate the superior performance of the sensor over previous implementations of depth from defocus. This improvement results from several factors including accurate modeling of sensor optics, the use of an optimized illumination pattern, and careful imlementation of the sensor. Since depth is computed from very subtle differences between near and far focused images, high accuracy can only be achieved through careful analysis of the system components.

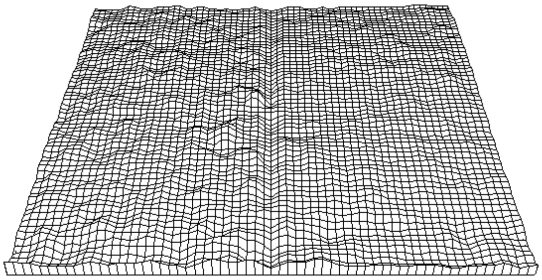
Figure 10 shows a scene with polyhedral objects. The computed depth map in Figure 10(b) is fairly accurate despite the complex textural properties of the objects. The only filtering that is applied to the depth map is a 5x5 smoothing function to reduce high frequency noise in computed depth that results from the low signal-to-noise ratio of the CCD cameras and spurious frequencies caused by surface texture. All surface discontinuities and orientation discontinuities are well preserved. The recovered shapes are precise enough for a variety of visual tasks including recognition and inspection. Similar results are shown in Figure 11 where shapes of curved objects are recovered. In the case of dynamic scenes, structure can be estimated only by using a real-time sensor. Figure 12 shows an object's depth map computed as it rotates on a motorized turntable. Such depth map sequences are valuable for automatic CAD model generation from sample objects. Computed CAD models are useful not only for visual recognition tasks but also for graphics rendering. In both cases, object models are more often than not manually designed and input to the system, a process that is not only tedious but also impractical for large numbers of complex objects. Furthermore, real-time depth computation clearly enhances the capability of any vision system as it enables recovery of a deforming shape, precise tracking of moving objects, and robust navigation in dynamic scenes.

10 Summary

We have reported theoretical results on a variety of issues related to depth estimation by focus analysis. Accurate modeling of optics and sensing were shown to be essential to precise depth estimation. Both textured and textureless surfaces are recovered by using an optimized illumination pattern that is registered with the image sensor. We also presented an elegant solution to constant magnification defocusing, a problem that has limited the precision of depth from defocus algorithms. All of these results were used to implement a real-time focus range sensor that produces high resolution depth maps at frame rate. This sensor is unique in its ability to produce fast, dense, and precise depth information at a very low cost. With time we expect the sensor to find applications ranging from visual recognition and robot control to automatic CAD model generation for visualization and virtual reality. The obvious extension to this work is the development of passive focus range finder for outdoor scenes. Such a sensor cannot afford the luxury of projected illumination. It must rely on complex scene textures for depth estimation. This problem shall be pursued in future work.



(a)

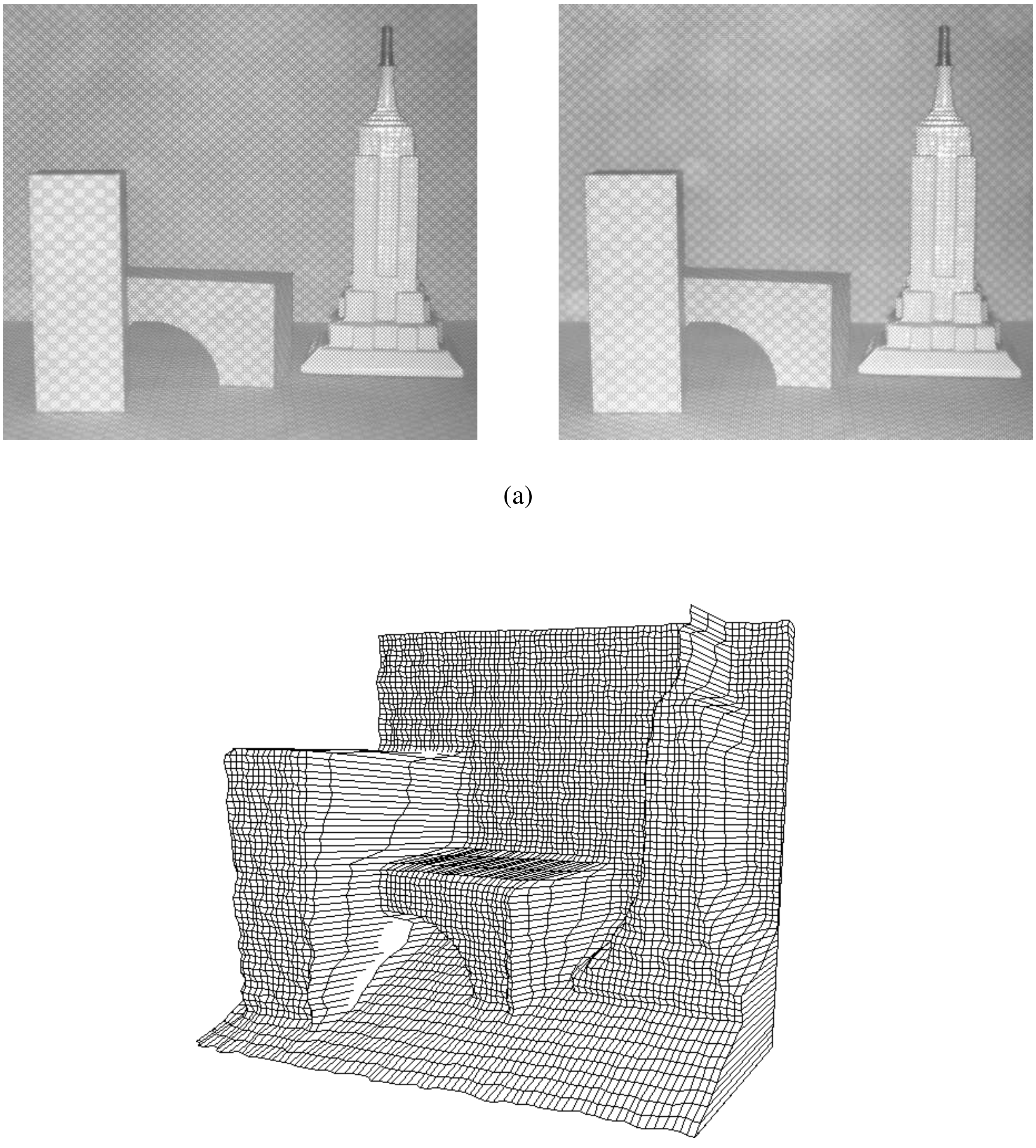


(b)

	Simulatneous Image Grab	Successive Image Grab
Depth Accuracy (rms)	0.24 %	0.34 %
Repeatability (rms)	0.23 %	0.29 %
Spatial Resolution	256 x 240	512 x 480
Speed	30 Hz	30 Hz
Delay	33 msec	33 msec

(c)

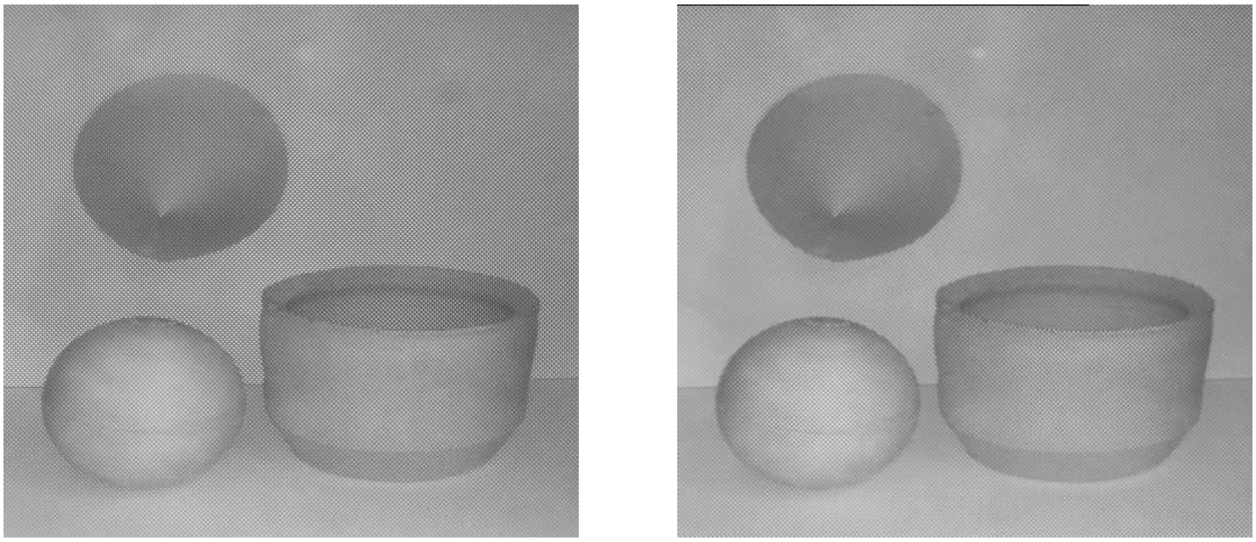
Figure 9: (a) Near focused image of a planar surface that includes highly textured and textureless areas. (b) Depth of the surface computed using the focus range sensor. (c) Performance characteristics of the sensor.



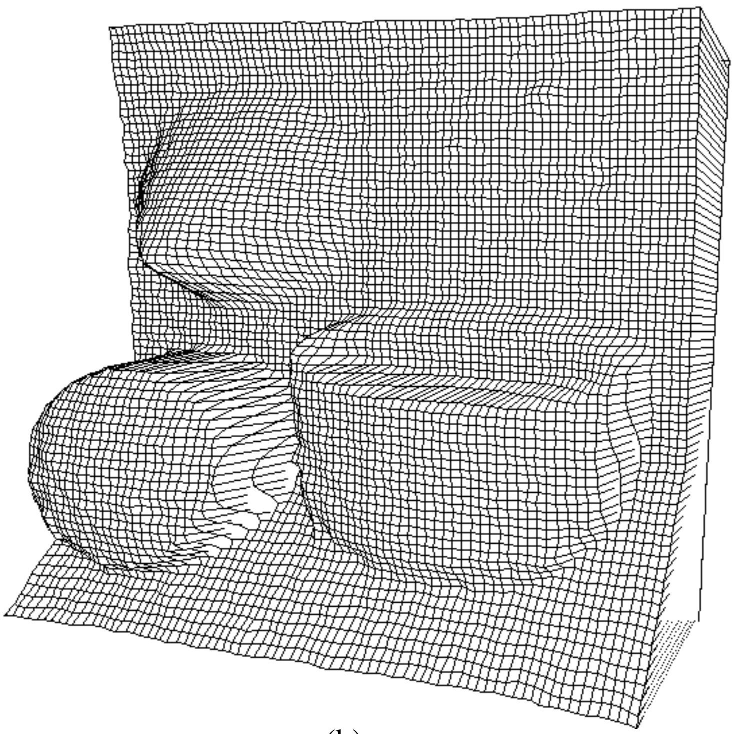
(a)

(b)

Figure 10: (a) Near and far focused images of a set of polyhedral objects. (b) Computed depth map.



(a)



(b)

Figure 11: (a) Near and far focused images of a set of curved objects. (b) Computed depth map.

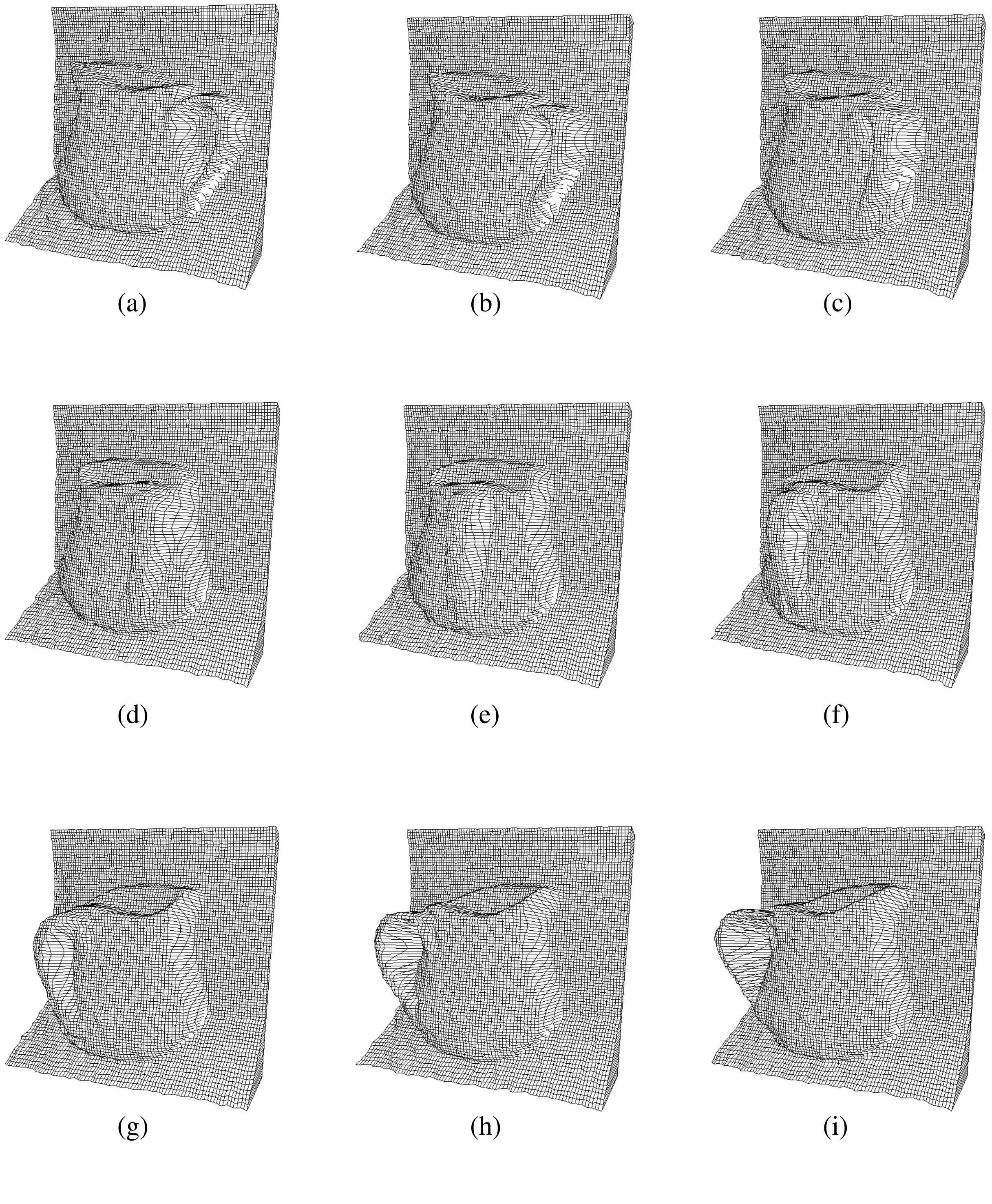


Figure 12: Depth maps generated by the sensor at 30 Hz while an object rotates on a motorized turntable.

A Details of Illumination Optimization

Consider expression (29) in Section 5. From (25) we see that b_x only figures in the illumination model I. Further, only the illumination cell function I_c is a function of b_x and not the illumination grid function I_g. Therefore:

$$\begin{aligned}
\frac{\partial}{\partial b_x} G_0(\frac{1}{t_x}, \frac{1}{t_y}) &= (I_g \cdot O^2 \cdot H' \cdot H \cdot S_c \cdot L) \cdot \frac{\partial}{\partial b_x} I_c \\
&= (I_g \cdot O^2 \cdot H' \cdot H \cdot S_c \cdot L) \cdot \frac{\partial}{\partial b_x} (b_x \frac{\sin(\pi b_x \frac{1}{t_x})}{\pi b_x \frac{1}{t_x}} \cdot b_y \frac{\sin(\pi b_y \frac{1}{t_y})}{\pi b_y \frac{1}{t_y}}) \\
&= (I_g \cdot O^2 \cdot H' \cdot H \cdot S_c \cdot L \cdot b_y \frac{\sin(\pi b_y \frac{1}{t_y})}{\pi b_y \frac{1}{t_y}}) \cdot \frac{\partial}{\partial b_x} (\frac{t_x}{\pi} \sin(\pi b_x \frac{1}{t_x})) \\
&= (I_g \cdot O^2 \cdot H' \cdot H \cdot S_c \cdot L \cdot b_y \frac{\sin(\pi b_y \frac{1}{t_y})}{\pi b_y \frac{1}{t_y}}) \cdot \cos(\pi b_x \frac{1}{t_x}) \;=\; 0
\end{aligned} \tag{45}$$

This yields the constraint:

$$\pi b_x \frac{1}{t_x} = \frac{\pi}{2} \cdot (2 \cdot m + 1) \qquad (m = integer) \tag{46}$$

Note that the illumination cell size b_x cannot exceed the illumination pitch t_x, else the cells overlap causing the resulting illumination to lie outside the range of illuminations that can be implemented in practice. When overlap between cells occurs, the model defined in Section 4.1 predicts that the illumination intensity in an overlap region exceeds the intensity within non-overlapping regions of the cell. Such a situation, though easy to express in theory, is difficult to realize in practice. Therefore, we have an additional constraint, namely, the illumination cell must be smaller or equal to illumination pitch ($0 < b_x < t_x$). Given (46), this gives a unique solution:

$$b_x = \frac{t_x}{2} \tag{47}$$

Using exactly the same procedure, we get:

$$b_y = \frac{t_y}{2} \tag{48}$$

Next, we evaluate the partial derivatives in (30). Again, from (25) it is seen that the parameter q_x occurs only in the focus operator L. Hence:

$$\begin{aligned}
\frac{\partial}{\partial q_x} G_0(\frac{1}{t_x}, \frac{1}{t_y}) &= \frac{\partial}{\partial q_x} (I_g \cdot I_c \cdot O^2 \cdot H' \cdot H \cdot S_c \cdot L) \\
&= (I_g \cdot I_c \cdot O^2 \cdot H' \cdot H \cdot S_c) \cdot \frac{\partial}{\partial q_x} L
\end{aligned} \tag{49}$$

$$= (I_g \cdot I_c \cdot O^2 \cdot H' \cdot H \cdot S_c) \cdot \frac{\partial}{\partial q_x}(4 - 2\cos(2\pi q_x \frac{1}{t_x}) - 2\cos(2\pi q_y \frac{1}{t_y}))$$

$$= (I_g \cdot I_c \cdot O^2 \cdot H' \cdot H \cdot S_c) \cdot (4\pi \frac{1}{t_x}) \sin(2\pi q_x \frac{1}{t_x}) = 0$$

This gives the constraint:

$$2\pi q_x \frac{1}{t_x} = \pi \cdot m \qquad (m = integer) \tag{50}$$

Note that our objective is only to maximize the absolute value of G_0, this could happen when G_0 is either maximum or minimum (since $g_0(x, y)$ is the product of cosines and can have maximum absolute value when it is positive or negative). Also, our goal is to achieve highest depth resolution. This requires that the focus operator L has minimum possible support ensuring that depth is computed using the smallest neighborhood. Therefore, (50) leads to the solution:

$$q_x = \frac{t_x}{2} \tag{51}$$

Similarly:

$$q_y = \frac{t_y}{2} \tag{52}$$

In order to verify if the above solutions do maximize the absolute value of G_0, the following second partial derivatives of G_0 with respect to each of the above parameters are used:

$$\begin{aligned} \frac{\partial^2}{\partial {b_x}^2} G_0(\frac{1}{t_x}, \frac{1}{t_y}) &= -(I_g \cdot O^2 \cdot H' \cdot H \cdot S_c) \\ &\cdot (4 - 2\cos(2\pi q_x \frac{1}{t_x}) - 2\cos(2\pi q_y \frac{1}{t_y})) \\ &\cdot \frac{t_y}{t_x} \sin(\pi b_x \frac{1}{t_x}) \sin(\pi b_y \frac{1}{t_y}) \end{aligned} \tag{53}$$

$$\begin{aligned} \frac{\partial^2}{\partial {b_y}^2} G_0(\frac{1}{t_x}, \frac{1}{t_y}) &= -(I_g \cdot O^2 \cdot H' \cdot H \cdot S_c) \\ &\cdot (4 - 2\cos(2\pi q_x \frac{1}{t_x}) - 2\cos(2\pi q_y \frac{1}{t_y})) \\ &\cdot \frac{t_x}{t_y} \sin(\pi b_x \frac{1}{t_x}) \sin(\pi b_y \frac{1}{t_y}) \end{aligned} \tag{54}$$

$$\begin{aligned} \frac{\partial^2}{\partial {q_x}^2} G_0(\frac{1}{t_x}, \frac{1}{t_y}) &= (I_g \cdot O^2 \cdot H' \cdot H \cdot S_c) \cdot (8\pi^2 \frac{1}{{t_x}^2}) \cos(2\pi q_x \frac{1}{t_x}) \\ &\cdot (b_x \frac{\sin(\pi b_x \frac{1}{t_x})}{\pi b_x \frac{1}{t_x}} \cdot b_y \frac{\sin(\pi b_y \frac{1}{t_y})}{\pi b_y \frac{1}{t_y}}) \end{aligned} \tag{55}$$

$$\frac{\partial^2}{\partial {q_y}^2} G_0(\frac{1}{t_x}, \frac{1}{t_y}) = (I_g \cdot O^2 \cdot H' \cdot H \cdot S_c) \cdot (8\pi^2 \frac{1}{{t_y}^2}) \cos(2\pi q_y \frac{1}{t_y}) \cdot (b_x \frac{\sin(\pi b_x \frac{1}{t_x})}{\pi b_x \frac{1}{t_x}} \cdot b_y \frac{\sin(\pi b_y \frac{1}{t_y})}{\pi b_y \frac{1}{t_y}}) \quad (56)$$

$$\frac{\partial^2}{\partial b_x \partial b_y} G_0(\frac{1}{t_x}, \frac{1}{t_y}) = \frac{\partial^2}{\partial b_x \partial q_x} G_0(\frac{1}{t_x}, \frac{1}{t_y}) = \frac{\partial^2}{\partial b_x \partial q_y} G_0(\frac{1}{t_x}, \frac{1}{t_y}) = \frac{\partial^2}{\partial b_y \partial q_x} G_0(\frac{1}{t_x}, \frac{1}{t_y}) = \frac{\partial^2}{\partial b_y \partial q_y} G_0(\frac{1}{t_x}, \frac{1}{t_y}) = \frac{\partial^2}{\partial q_x \partial q_y} G_0(\frac{1}{t_x}, \frac{1}{t_y}) = 0 \quad (57)$$

When the parameters have the values given by (47), (48), (51) and (52), the second derivatives in (53), (54), (55) and (56) all have the same sign (+ or -), verifying that the solutions of (47), (48), (51) and (52), maximize $\mid G_0 \mid$.

To summarize, the sensitivity of the focus measure function $g(x, y)$ is maximized when:

$$b_x = \frac{1}{2} t_x, \quad b_y = \frac{1}{2} t_y \quad (58)$$

$$q_x = \frac{1}{2} t_x, \quad q_y = \frac{1}{2} t_y \quad (59)$$

References

[Asada *et al.*, 1993] N. Asada, H. Fujiwara and T. Matsuyama. Edge and depth from focus. *Proc. of Asian Conf. on Computer Vision*, pages 83–86, November 1993.

[Authors, 1994] Authors. Real-time implementation of depth from defocus. *Technical Report, in preparation*, November 1994.

[Besl, 1988] P. J. Besl. Range imaging sensors. Technical Report GMR-6090, General Motors Research Laboratories, March 1988.

[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.

[Boyer and Kak, 1987] K. L. Boyer and A. C. Kak. Color-encoded structured light for rapid active sensing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9(1):14–28, January 1987.

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

[Darrell and Wohn, 1988] T. Darrell and K. Wohn. Pyramid based depth from focus. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 504–509, June 1988.

[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.

[Girod and Scherock, 1989] B. Girod and S. Scherock. Depth from focus of structured light. *Proc. of SPIE: Optics, Illumumination, and Image Sensing for Machine Vision IV*, 1194, November 1989.

[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.

[Grossman, 1987] P. Grossman. Depth from focus. *Pattern Recognition*, 9(1):63–69, 1987.

[Gruss *et al.*, 1993] A. Gruss, S. Tada and T. Kanade. A vlsi smart sensor for fast range imaging. *Proc. of ARPA Image Understanding Workshop*, pages 977–986, April 1993.

[Horn, 1968] B. K. P. Horn. Focusing. Technical Report Memo 160, AI Lab., Massachusetts Institute of Technology, Cambridge, MA, USA, 1968.

[Horn, 1986] B. K. P. Horn. *Robot Vision.* The MIT Press, 1986.

[Inokuchi *et al.*, 1984] S. Inokuchi, K. Sato and F. Matsuda. Range imaging system for 3-d object recognition. *Proc. of 7th Intl. Conf. on Pattern Recognition*, pages 806–808, July 1984.

[Jarvis, 1983] R. A. Jarvis. A perspective on range finding techniques for computer vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 5(2):122–139, March 1983.

[Kanade *et al.*, 1991] T. Kanade, A. Gruss and L. R. Carley. A very fast vlsi rangefinder. *Proc. of Intl. Conf. on Robotics and Automation*, pages 1322–1329, April 1991.

[Kingslake, 1983] R. Kingslake. *Optical System Design.* Academic Press, 1983.

[Krishnan and Ahuja, 1993] A. Krishnan and N. Ahuja. Range estimation from focus using a non-frontal imaging camera. *Proc. of AAAI Conf.*, pages 830–835, July 1993.

[Krotkov, 1987] E. Krotkov. Focusing. *International Journal of Computer Vision*, 1:223–237, 1987.

[Nair and Stewart, 1991] H. N. Nair and C. V. Stewart. Robust focus ranging. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 309–314, June 1991.

[Nayar and Nakagawa, 1994] S. K. Nayar and Y. Nakagawa. Shape from focus: An effective approach for rough surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(8):824–831, August 1994.

[Noguchi and Nayar, 1994] M. Noguchi and S. K. Nayar. Microscopic shape from focus using active illumination. *Proc. on Intl. Conf. on Patt. Recog.*, October 1994.

[Pentland *et al.*, 1989] A. Pentland, T. Darrell, M. Turk, and W. Huang. A simple, real-time range camera. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 256–261, June 1989.

[Pentland *et al.*, 1994] A. Pentland, S. Scherock, T. Darrell, and B. Girod. Simple range cameras based on focal error. *Journal of Optical Society of America A*, 11(11):2925–2935, November 1994.

[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, 1992] M. Subbarao and G. Surya. Application of spatial-domian convolution/deconvolution transform for determining distance from image defocus. *Proc. of OE/BOSTON '92 SPIE Conference*, 1822, November 1992.

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

[William and Becklund, 1989] C. S. William and O. A. Becklund. *Introduction to the Optical Transfer Function.* A. Wiley Interscience Publication, New York, 1989.

[Willson and Shafer, 1994] R. G. Willson and S. A. Shafer. Modeling and calibration of automated zoom lenses. Technical Report CMU-RI-TR-94-03, The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, January 1994.

[Xiong and Shafer, 1994a] Y. Xiong and S. A. Shafer. Moment and hypergeometric filters for high precision computation of focus, stereo and optical flow. Technical Report CMU-RI-TR-94-28, The Robotics Institute, Carnegie Mellon University, Pittsburg, PA, USA, September 1994.

[Xiong and Shafer, 1994b] Y. Xiong and S. A. Shafer. Variable window gabor filters and their use in focus and correspondence. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 668–671, June 1994.