Real-time computation of depth from defocus

Masahiro Watanabe*, Shree K.Nayar† and Minori Noguchi*

*Hitachi Ltd., Production Engineering Research Lab.
292 Yoshida-cho, Totsuka, Yokohama 244, Japan

†Columbia University, Department of Computer Science
New York, NY 10027

ABSTRACT

A new range sensing method based on depth from defocus is described. It uses illumination pattern projection to give texture to the object surface. Then the image of the scene is split into two images with different focus settings and sensed simultaneously. The contrast map of the two images are computed and compared pixel by pixel to produce a dense depth map. The illumination pattern and the focus operator to extract the contrast map are designed to achieve finest spatial resolution of the computed depth map and to maximize response of the focus operator. As the algorithm uses only local operations such as convolution and lookup table, the depth map can be computed rapidly on a data-flow image processing hardware. As this projects an illumination pattern and detects the two images with different focus setting from exactly the same direction, it does not share the problem of shadowing and occlusion with triangulation based method and stereo. It's speed and accuracy are demonstrated using a prototype system. The prototype generates 512×480 range maps at 30 frame/sec with a depth resolution of 0.3% relative to the object distance. The proposed sensor is composed of off-the-shelf components and outperforms commercial range sensors through its ability to produce complete three-dimensional shape information at video rate.

Keywords: *range sensor, focus, depth from defocus, telecentric optics, pipeline processor, real-time sensor*

1 INTRODUCTION

For applications such as object recognition, automatic CAD model generation and remote visualization, a range sensor which produces fast and dense depth maps is necessary. In the past, many techniques for range sensing have been proposed,[11] which can be categorized into passive techniques which do not use an active illumination or active techniques which use an active illumination.

Passive techniques such as stereo and shape from motion are based on correspondence matching between two or more images. From the disparity or the motion vector extracted from this correspondence matching, one can get the range data of the scene. The correspondence matching is computationally expensive. These techniques also suffer from the occlusion problem, i.e. one cannot get depth data for areas in the scene which are visible to only one of the camera. Another passive technique is depth from focus/defocus. Depth from focus uses a sequence of images taken by incrementing the focus setting in small steps. For each pixel, the focus setting that maximizes image contrast is determined. This in turn can be used to compute the depth of the corresponding scene point.[8,11,14,15,4,25] In contrast, depth from defocus uses only two images with different optical settings.[18,5,2,22,26] Though depth from focus/defocus does not have the problem of occlusion, it is also computationally expensive to get a reliable depth map.[26] This is because the texture of the object has a variety of spectrum distribution, and one must analyse carefully to get a reliable focus estimate. Another draw back that is common to the passive techniques is that one cannot compute depth for scene areas without texture.

Popular active techniques are based on the principles of structured light and time of flight. Popular structured light methods include light striping method,[10] moiré interferometry[6] and Fourier-transform profilometry.[23] They are based on triangulation and determines the depth from the deformation of the image of the projected pattern. They provide reasonable accuracy. For light striping method, one must project many sets of light stripe pattern to encode the stripes in order to discriminate the stripes. This makes sensing time long, which implies that the scene must be static during the sensing. New hope for light stripe range finding has been instilled by advances in VLSI. Based on the notion of cell parallelism,[12] 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 current VLSI density limits the total number of cells, and hence spatial depth resolution, to 28×32.[7] Future advances in VLSI are expected to yield high-resolution depth maps at high speeds. For moiré interferometry and Fourier-transform profilometry, one needs only one image but the scene should not have a steep depth gap, as it must keep track of the fringe order. In addition, the depth it gives is just a relative depth, not absolute depth. Another type of active method, time of flight, uses a modulated laser beam and measures the time for the light to come back from the object surface.[11] Although this method is suitable at getting a rough depth map for relatively far scenes, it takes a long time to get a dense depth map as it scans the scene point by point.

A sensor which uses focus error information and active illumination has been proposed by Rioux et al.[20] and Pentland et al.[19] They project a matrix of dots[20] or light-stripe pattern.[19] Using the phenomenon that the diameter or the width of the defocused dot or strip gets larger when it is defocused, this dimension is measured from the image and is converted into a depth value. These sensors are able to detect the depth map of a dynamic scene in real-time. However, as they use a coarse matrix of dots or coarse light-stripe pattern, resulting spatial resolution and depth accuracy are insufficient for most real-world applications.

Our approach uses co-axial projection of a fine illumination pattern onto the scene and detects two images with two CCD sensors that have different focus settings. A focus operator, i.e. narrow-band-pass convolution filter is newly designed to provide estimates of the defocus of the projected illumination pattern. The operator is derived by careful modeling of the illumination, blurring and image sensing and is tuned to respond to the fundamental frequency of the projected illumination pattern. The focus operator is applied to the two images to obtain two different focus measures at each image point. The relative defocus of each image point maps to a unique depth estimate. The computation of depth is a local operation, which enables us to realize a frame-rate range sensor. Since the illumination pattern and the tuned contrast operator were designed to maximize depth accuracy and resolution, the sensor produces depth maps of high quality. The co-axial illumination and imaging also results in a shadowless image; all surface regions that are visible to the sensor are also illuminated. A prototype real-time focus range sensor has been developed. 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. In the previous paper,[17] the authors have discussed this sensor mainly on the basic concept including illumination pattern design and the usage of constant magnification optics. In this paper, we focus on the contrast operator and real-time depth computation. The performance of the sensor is demonstrated through several experiments conducted on complex scenes. Quantitative results on the accuracy, repeatability, and linearity of the sensor are included.

2 DEPTH FROM DEFOCUS

2.1 Basic concept

Fundamental to depth from defocus is the relationship between focused and defocused images.[1] 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 lens law:

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

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.[18,21] 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. These include (a) accurate estimation of relative defocus in the two image, (b) recovery of textured and textureless surfaces, and (c) achieving constant magnification that is invariant to the degree of defocus.

2.2 Telecentric optics

We begin with the last of the problems mentioned above. In the imaging system shown in Figure 2, the effective image location of point P moves along the *principal ray* R as the sensor plane is displaced. This causes a shift in image coordinates of the image of P. This variation in image magnification with defocus manifests as 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. We approach the problem from an optical perspective rather than 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 (see book.[13]) As a result, despite blurring, the effective image coordinates of point P in both images I_1 and I_2 are the same as the coordinate of its focused image Q on I_f. The detailed discussion of this is found in the technical report.[24]

2.3 Defocus function and depth estimation

The defocus function is described in detail in previous work.[1,9] 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 with a radius of $\alpha a'/f$ on the sensor plane.† This patch, also called the *pillbox*, is the defocus function;

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

where a' is the radius of the telecentric lens aperture, and $\Pi(r)$ is a rectangular function which takes a value 1 for $|r| < \frac{1}{2}$, 0 otherwise. In Fourier domain, the above defocus function is given by:

$$H(u,v) = H(u,v;\alpha,a',f) = \frac{f}{\pi a'\alpha\sqrt{u^2+v^2}} J_1\left(\frac{2\pi a'\alpha}{f}\sqrt{u^2+v^2}\right) \tag{3}$$

where J_1 is the first-order Bessel function. 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. Figure 4 visualizes the above discussion; (a) is the image $i(x,y)$ at the focused plane, I_f, and its Fourier spectrum $I(u,v)$. When the sensor plane is displaced to I_1, the defocused image is the convolution of the focused image $i(x,y)$ with the pill-box $h_1(x,y)$ as in (b). In the Fourier domain, it is the product of Fourier spectrum of the focused image $I(u,v)$ and the Fourier transform of the pill-box $H_1(u,v)$. (c) is the equivalent set to (b) when the sensor is placed at I_2, i.e. at distance $\beta - \alpha$ from the focused plane I_f. As the image is defocused more, the low-pass response of the transfer function $H_2(u,v)$ is more notable.

*Since we have used the telecentric lens (Figure 3) in our implementation, it's parameters are used in the model. However, the following expressions can be made valid for the conventional lens model (Figure 2) by simply replacing the factor $\frac{f}{a'}$ by $\frac{d_i}{a}$. In addition, the nominal F-number of the lens equals $\frac{f}{2a'}$.

†This geometric model is valid as far as the lens is not exactly focused and the aberration is small compared to the radius $\alpha a'/f$.[1]

2.4 Active illumination

If one can get the amplitudes g_1 and g_2 of the spectrum of the two defocused images at a predefined frequency as in Figure 4, one can get the depth estimate from g_1 and g_2. This is done by applying a convolution operator to the images. But this is not trivial since the image texture includes all kinds of frequency. Uncertainty relation[3] tells us that, when we try to conduct a frequency analysis for a small area, the frequency resolution reduces proportionally to the inverse of the area size. To get a dense depth map, one must get the g_1 and g_2 for a very small area around each pixel. But this means the operator output is actually an averaged spectrum over a wide band of frequency. As the response of the defocus function H depends not only on defocus α but also on the texture frequency, this band width of the operator causes error in depth value. If the texture has only one frequency, the problem is solved. This is the reason why we have introduced active illumination. The projection filter pattern has been designed to achieve finest spatial resolution of the computed depth map and to maximize response of the focus operator (see papers.[17,16]) The resulting pattern is a checkerboard pattern with a horizontal period of t_x and a vertical period of t_y such that;

$$t_x \;=\; 4p_x, \quad t_y \;=\; 4p_y, \tag{4}$$

where p_x and p_y are the CCD pixel pitch in horizontal and vertical direction, respectively. The horizontal and vertical spacing between neighboring elements of the discrete Laplacian kernel (q_x, q_y) that corresponds to the optimal pattern obeys;

$$q_x \;=\; 2p_x, \quad q_y \;=\; 2p_y, \tag{5}$$

This means the 3×3 Laplacian operator kernel has zeros between each element, and it is actually a 5×5 kernel.‡ Figure 5 shows the effect of pattern projection. (a) is a image of a scene under normal lighting and its spectrum. (b) is the image of the same scene under the coaxial pattern projection and its spectrum. The spectrum in (b) shows that projected pattern creates strong peaks in the spectrum at positions $(\pm 1/t_x, \pm 1/t_y)$.

2.5 Depth from two images

Now let us introduce following normalized ratio;

$$q(x,y) \;=\; \frac{g_1(x,y) \;-\; g_2(x,y)}{g_1(x,y) \;+\; g_2(x,y)} \;=\; \frac{H(\frac{1}{t_x},\frac{1}{t_y};\alpha) \;-\; H(\frac{1}{t_x},\frac{1}{t_y};\alpha-\beta)}{H(\frac{1}{t_x},\frac{1}{t_y};\alpha) \;+\; H(\frac{1}{t_x},\frac{1}{t_y};\alpha-\beta)} \tag{6}$$

Here, g_1, g_2 and q are functions of image coordinate (x, y). As shown in Figure 6, q is a monotonic function of α such that $-p \le q \le p$ and $p \le 1$. This monotonic response is obtained as far as β and a' are chosen so that the analyzed frequency $(1/t_x, 1/t_y)$ is within the main lobe of the defocus function H;

$$\sqrt{\frac{1}{{t_x}^2} + \frac{1}{{t_y}^2}} \;<\; 0.61\frac{f}{\beta a'} \tag{7}$$

In practice, the above relation can be precomputed 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.

3 TUNED FOCUS OPERATOR

3.1 Design of the kernel

For the purpose of illumination optimization, we used the Laplacian operator as is described in the previous papers.[17,16] The resulting illumination pattern has only a single dominant absolute frequency, $(1/t_x, 1/t_y)$. Given

‡ In the papers,[17,16] we have shown another checkerboard pattern that is not used for the implementation, where $(t_x, t_y) = 2(p_x, p_y)$ and $(q_x, q_y) = (p_x, p_y)$. However, this pattern requires perfect registration between illumination pattern and sensor pixel. It is because the focus measure also depends on the phase between pattern and pixel. In this case, as the peak frequency is at the Nyquist frequency, the phase error cannot be compensated using *quadrature operation* which will be described in section 3.2.

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 3×3 discrete Laplacian (see Figure 7(a) and (b)). We see that though the Laplacian does have peaks at $(\pm 1/t_x, \pm 1/t_y)$, it has a fairly broad bandwidth allowing other spurious frequencies to contribute to the focus measure. 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. (i) To maximize spatial resolution in computed depth we force the operator kernel to be 3×3 or 4×4.[§] This is also a requirement from the convolution hardware of the pipeline processor we use, which can execute up to 8×8 convolution. (ii) Since the fundamental frequency of the illumination pattern has a symmetric quadrapole arrangement, the focus operator must be either reflection-symmetric or anti-reflection-symmetric about vertical and horizontal axis. (iii) 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. If we use 3×3 operator, condition (ii) forces the operator to have the structure shown in Figure 7(c), and condition (iii) becomes;

$$a + 4b + 4c = 0 \tag{8}$$

It is also imperative that the response of the operator;

$$L(u,v) = a + 2\,b\,(\cos 2\pi q_x u + \cos 2\pi q_y v) + 4\,c\,\cos 2\pi q_x u \cos 2\pi q_y v. \tag{9}$$

is not zero at the fundamental frequency, i.e. $L(\frac{1}{t_x}, \frac{1}{t_y}) \neq 0$. This reduces to:

$$a - 4b + 4c \neq 0 \tag{10}$$

Expressions (8) and (10) imply that $b \neq 0$. Without loss of generality, we set $b = -1$. Hence, (8) gives $a = 4(1-c)$. Therefore, the tuned operator is determined by a single unknown parameter, c, as shown in Figure 7(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 second-order moment of the power $|| L(u,v) ||^2$ with respect to $(1/t_x, 1/t_y)$:

$$\begin{aligned} M &= \frac{1}{|| L(\frac{1}{t_x}, \frac{1}{t_y}) ||^2} \int_{u=0}^{\frac{2}{t_x}} \int_{v=0}^{\frac{2}{t_y}} [(u-\frac{1}{t_x})^2 + (v-\frac{1}{t_y})^2] \; || L(u-\frac{1}{t_x}, v-\frac{1}{t_y}) ||^2 \, dv\, du \qquad (11) \\ &= \frac{t_x^2 + t_y^2}{768\,\pi^2\, t_x^3\, t_y^3} (20\pi^2 c^2 + 6c^2 + 48c - 32\pi^2 c + 20\pi^2 - 93) \end{aligned}$$

The above measure is minimized when $\frac{\partial M}{\partial c} = 0$, i.e. when $c = 0.658$ as shown in Figure 7(e). The resulting tuned focus operator has the response shown in Figure 7(f). It has substantially sharper peaks than the discrete Laplacian. We have also solved an optimization problem for 4×4 kernel case. This time it becomes a two-parameter minimization problem after considering the symmetric property. The resultant kernel and its spectrum response is shown in Figure 7 (f) and (g), respectively. 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.

3.2 Quadrature operation

As was discussed above, the focus operator passes the spectrum at the frequency $(\pm 1/t_x, \pm 1/t_y)$ and stops DC spectrum component. Let's denote the spectrum of the checkerboard illuminated image (Figure 5 (d)) as;

$$G_0(u,v) = g\{\delta(u-\frac{1}{t_x}, v-\frac{1}{t_y}) + \delta(u+\frac{1}{t_x}, v-\frac{1}{t_y}) + \delta(u-\frac{1}{t_x}, v+\frac{1}{t_y}) + \delta(u+\frac{1}{t_x}, v+\frac{1}{t_y})\}. \tag{12}$$

If the operator gain at the illumination frequencies $(\pm 1/t_x, \pm 1/t_y)$ is c, operator output is the inverse Fourier transform of $c\,G_0(u,v)$;

$$c\,g_0(x,y) = 4\,c\,g\,\cos 2\pi \frac{1}{t_x} x \cdot \cos 2\pi \frac{1}{t_y} y. \tag{13}$$

[§]Here, 3×3 or 4×4 counts the pixels with non-zero kernel values. Actual kernel has zero kernel element in-between, resulting in a kernel with a size of 5×5 or 7×7.

Actual values that the discrete focus operator gives are the $c\,g_0(x,y)$ values at discrete sampling positions $(\phi_x + mp_x, \phi_y + np_y)$;

$$g_d(m,n) \;=\; cg_0(\phi_x + mp_x, \phi_y + np_y) \;=\; 4\,c\,g\,\cos\pi(\frac{m}{2} + \frac{2\phi_x}{t_x}) \cdot \cos\pi(\frac{n}{2} + \frac{2\phi_y}{t_y}) \tag{14}$$

where m and n are pixel indexing integer values and (ϕ_x, ϕ_y) is the relative shift between illumination and CCD pixel. Here the relationship of equation (4) was used. Equation (14) shows that the operator output $g_d(m,n)$ is sensitive to the registration, (ϕ_x, ϕ_y). To cope with this problem, we use the fact that $g_d(m,n)$ at the next pixel has a $\pi/2$ phase difference, then It is easily shown that;

$$g \;=\; \frac{1}{4\,c}\sqrt{g_d(m,n)^2 + g_d(m+1,n)^2 + g_d(m,n+1)^2 + g_d(m+1,n+1)^2} \tag{15}$$

Here we get a focus measure g which is insensitive to the sub-pixel order mis-registration. This *quadrature operation* loosens the requirement for the accuracy to register the CCD's and the illumination pattern, making pixel-order registration enough.

4 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 converted to telecentric with an additional aperture inside. Aperture diameter is set so that *F-number* is 6.5. 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 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 corresponds to a sensor depth of field of 257 mm (a detectable range of 305∼562 mm.) This detectable range of the sensor can be varied by changing the sensor displacement and the focus distance of the lens. *F-number* of the optics should be chosen to fulfill equation (7).

The checkerboard illumination pattern was etched on a glass plate using microlithography. The filter was then placed in the path of a 300 W Xenon arc lamp. The illumination pattern 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 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 two A/D converters, one 12-bit convolver (maximum kernel size:8×8,) one arithmetic logic unit (ALU) and one 16-bit look up table, which can be aligned on a pipeline. Data in the pipeline flow at 20MHz. The pipeline also requires $1 \sim 2$msec as an overhead. For example, for a 512×480 pixel image, the pipeline is completed in about 14.3 msec. This hardware enables simultaneous digitization of the two images, convolution of both images with the tuned focus operator, and the computation of a 256×240 depth map, all within a single frame time of 33 msec with a lag of 33 msec. Figure 9 shows the data flow. The first and the second pipeline input the near focused (i_1) and far focused (i_2) images, respectively, and execute convolution with the tuned focus operator and quadrature operation to produce the focus measure image of 512×480 resolution. Each pipeline takes 14.2 msec. The focus measure image is sub-sampled in the third pipeline to a resolution of 256×240 and input to the 16-bit look up table. The look-up table is configured to take two 8-bit inputs and map each pair of focus measures (g_1 and g_2) to a unique depth estimate d. Here, the normalized ratio of focus measures q in equation (6) is not output. Instead, the depth value d or d_i is directly output. Then the depth map goes through linear and non-linear smoothing and pixel-by-pixel linear calibration in a single pipeline as the look up table, which takes 4.8 msec. The above three pipelines produce a depth map in 33 msec in total. Image grabbing of the near and far images for the next depth computation is accomplished parallelly with the above three pipelines.

A pixel-by-pixel linear calibration is executed to compensate for the image curvature and vignetting. Image curvature causes an offset of the depth value. Vignetting changes the depth detection gain. A planer target is placed perpendicularly to the optical axis of the sensor at a far position z_1 and a near position z_2 in the ranging depth. Then the depth maps are detected and smoothed using a spline function. Let us denote the smoothed depth map when the target is at z_1 by $z_{d1}(x, y)$. Similarly denote the smoothed depth map when the target is at z_2 by $z_{d2}(x, y)$. Then the calibration gain map in Figure 9 is computed by;

$$\frac{z_2 - z_1}{z_{d2}(x, y) - z_{d1}(x, y)} . \tag{16}$$

The calibration offset map is computed by;

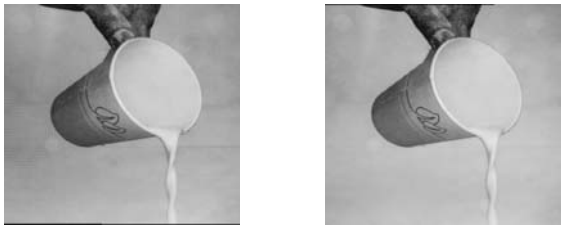
$$\frac{z_1\, z_{d2}(x, y) - z_2\, z_{d1}(x, y)}{z_{d2}(x, y) - z_{d1}(x, y)} . \tag{17}$$

The sub-sampling for the third pipeline is merely because of the time restriction. Instead, by giving up simultaneous grabbing of the near and far images, a 512×480 depth map can be computed at the same rate if the two images are taken in succession. Still, simultaneous image acquisition is clearly advantageous since it makes the sensor less sensitive to variations in both illumination and scene structure between frames. With an addition of one more MV200 to the present processing hardware, it is easy to obtain 512×480 depth maps at 30 Hz using simultaneous image grabbing. Depth maps produced by the sensor are shipped via video cable and visualized as wire-frame plots with 80×60 meshes at a speed of 18 frame/sec on a DEC Alpha workstation.

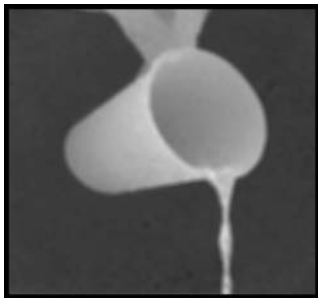
5 EXPERIMENTS

Numerous experiments have been conducted to test the performance of the sensor. Here we briefly summarize these results. Figure 10(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 10(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. Several depth maps of the plane in Figure 10(a) were computed by varying its position in the 250 mm workspace of the sensor. Relative accuracy and repeatability of the sensor were estimated for both simultaneous and successive image grabbing configurations (see Table 10(c)). The conventional definition of relative accuracy and repeatability is used, where it is a value relative to the object distance. Accuracy is measured by the fitting error to a plane. Repeatablity is measured by the standard deviation of the depth values measured at the same position at different times. Here we discuss absolute accuracy (linearity.) Figure 11(a) is the plot of detected depth for the textured planer target vs. the target depth. Detected depth is determined by a plane fitting to a 50×50 pixel area in the depth map. The fitting errors are also shown as 6σ error bars. The deviation from the linearity is 5.2mm (σ.) The slight curvature is due to the error in the optical parameter. After a calibration of the look up table using a quadratic function, the linearity is improved as is shown in Figure 11(b). The linearity is 2.5 mm. Figure 12 is the same pair of plots as Figure 11, except that the detected depth is determined by an average of 150 successive depth value at a pixel. The linearity is 1.0 mm with calibration. These results clearly demonstrate the superior performance of the sensor over previous implementations of depth from defocus.

Figure 13 shows a scene with polyhedral objects. The computed depth map in Figure 13(b) is fairly accurate despite the complex textural properties of the objects. The only filtering that is applied to the depth map is a 5×5 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 14 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 15 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



(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.

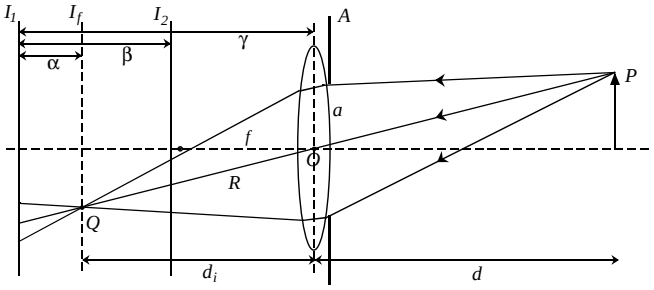


Figure 2: Image formation and depth from defocus.

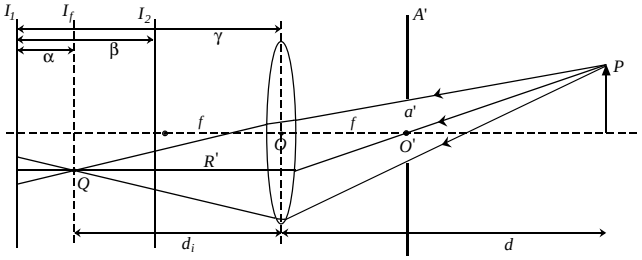
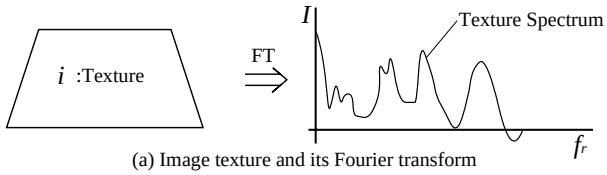
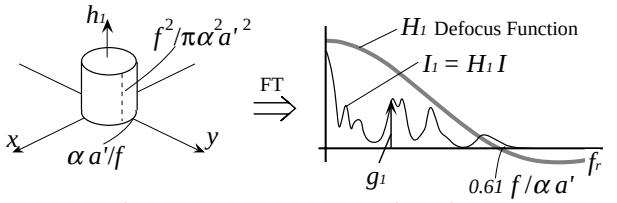


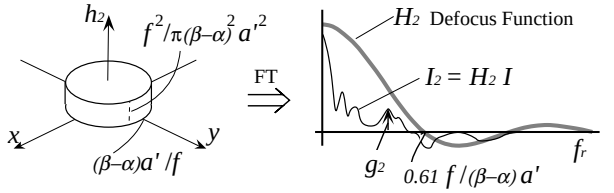
Figure 3: A constant-magnification imaging system for depth from defocus is achieved by simply placing an aperture at the front-focal plane of the optics.



(a) Image texture and its Fourier transform

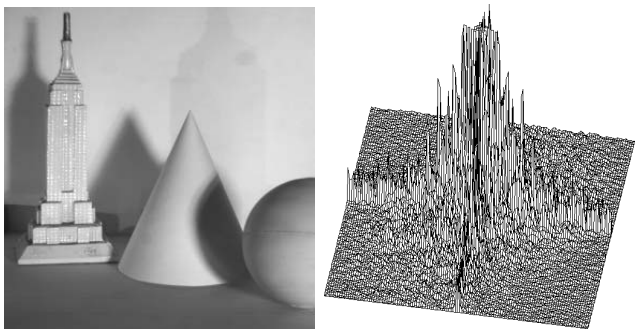


(b) Pill-box defocus model and the Fourier transform of the blured image I_1

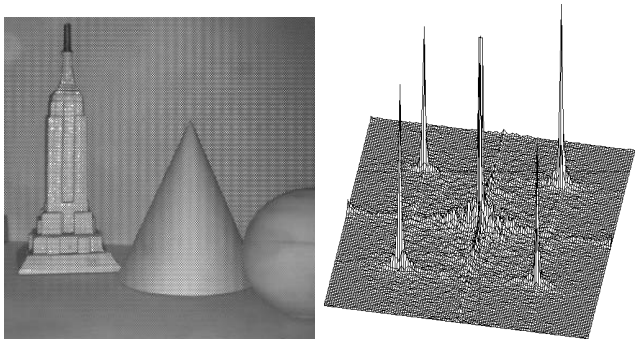


(c) Pill-box defocus model and the Fourier transform of the blured image I_2

Figure 4: Image blaring and focus measure.



(a) image under normal lighting and its power spectrum



(b) image under pattern projection and its power spectrum

Figure 5: Effect of active pattern projection.

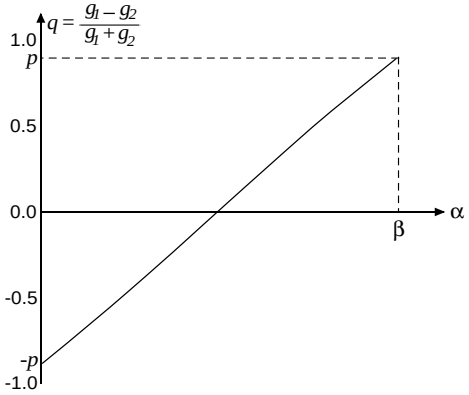


Figure 6: Relation between focus measures g_1 and g_2 and the defocus parameter α.

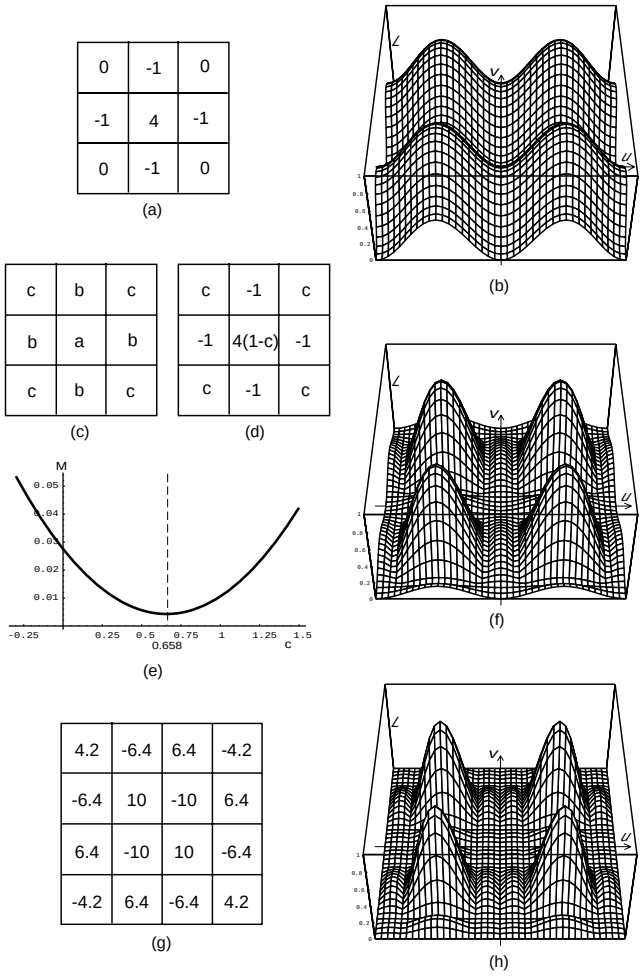


Figure 7: (a) The 3×3 Laplacian and its (b) Fourier transform. (c) The kernel structure for a 3×3 symmetric operator. (d) The kernel of a 3×3 operator that dose not pass DC component (see text). (e) The second moment M of each of the operator peaks is minimized when $c = 0.658$. (f) Response of the tuned focus operator ($c = 0.658$) has much sharper peaks than the Laplacian. (g) 4×4 optimal operator. (h) Response of (g).

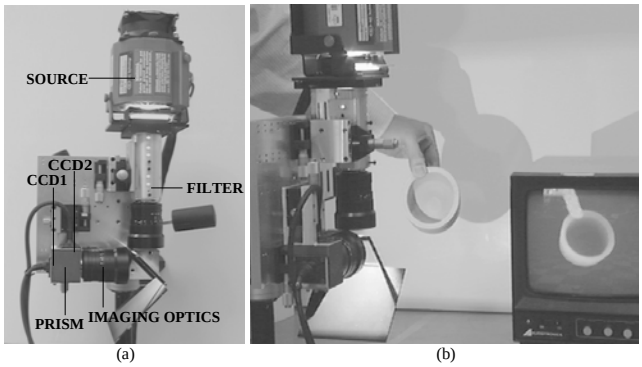


Figure 8: (a) The real-time focus range sensor and its key components. (b) The sensor can produce depth maps up to 512×480 in resolution at 30 Hz.

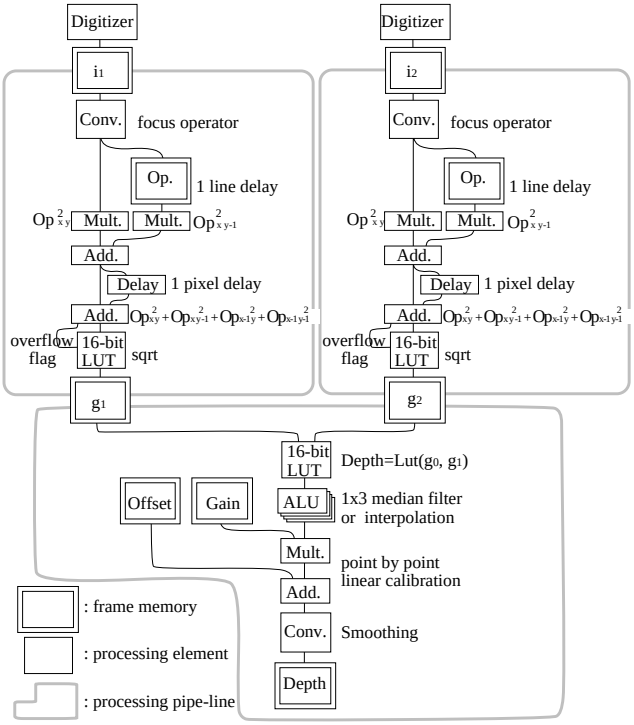
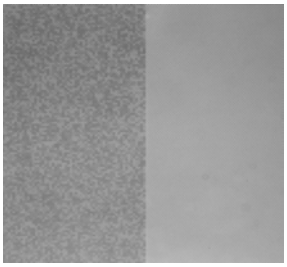
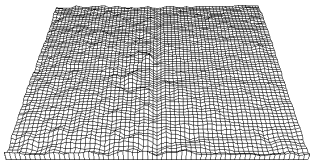


Figure 9: Dataflow for the real-time depth computation



(a)

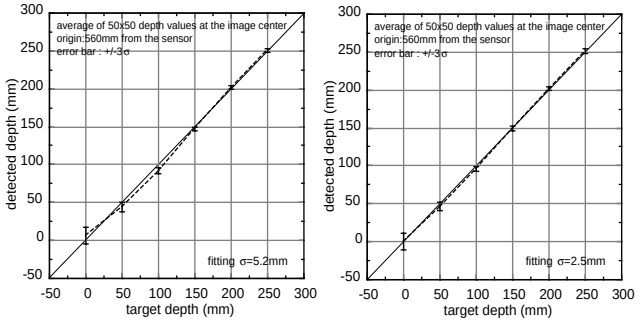


(b)

	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

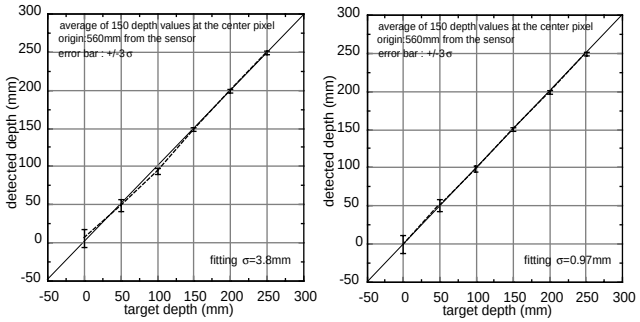
(c)

Figure 10: (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.



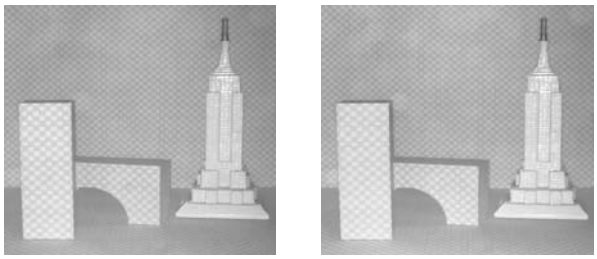
(a) without calibration (b) after calibration

Figure 11: Accuracy and linearity

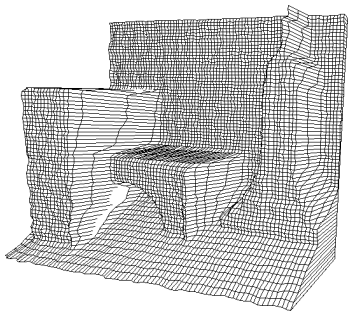


(a) without calibration (b) after calibration

Figure 12: Repeatability and linearity

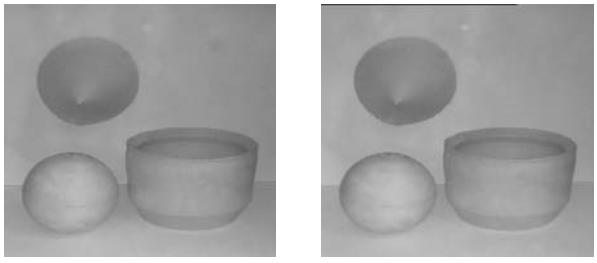


(a)

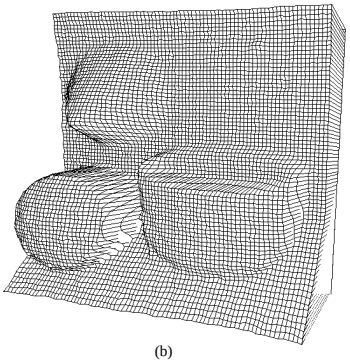


(b)

Figure 13: (a) Near and far focused images of a set of polyhedral objects. (b) Computed depth map.



(a)



(b)

Figure 14: (a) Near and far focused images of a set of curved objects. (b) Computed depth map.

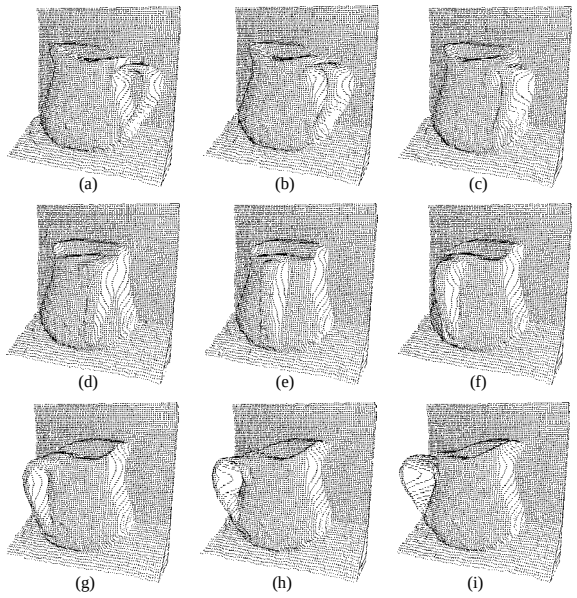
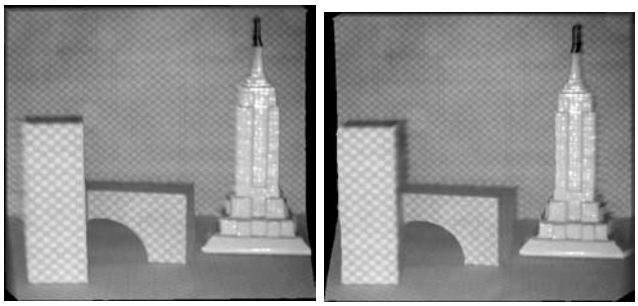


Figure 15: Depth maps generated by the sensor at 30 Hz while an object rotates on a motorized turntable.



(a) left-eye image (b) right-eye image

Figure 16: Stereo-pair of texture-mapped images are synthesized from the detected image and depth map.

large numbers of complex objects. Figure 16 is an example of graphics rendering. Since this sensor gives both depth map and image data, We can symthesize images viewed from directions that are different from the sensing direction. 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.

6 SUMMARY

In order to get a dense and accurate depth map at frame rate for both textured and textureless surface, we have incorporated co-axial active pattern projection to depth from defocus method. The projection pattern and focus operator to extract the contrast of the projected pattern has been designed through careful modeling of the optics, sensing and processing. To solve the pixel order registration problem between the image sensors, we have introduced a telecentric optics for constant magnification. To solve the sub-pixel-order registration problem between the two image sensors and the illumination pattern, we have introduced quadrature operation which is applied after the focus operator. 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.

7 ACKNOWLEDGEMENTS

This research was conducted at the Center for Research in Intelligent Systems, Department of Computer Science, Columbia University. We would like to thank Dr. Yasuo Nakagawa at Hitachi Ltd. for his encouragement to this research.

8 REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] J. J. Gibson. *The senses considered as perceptual systems*. Houghton Mifflin, Boston, 1966.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] S. K. Nayar, M. Watanabe, and M. Noguchi. "Real-time focus range sensor". Technical Report CUCS-028-94, Dept. of Computer Science, Columbia University, New York, NY, USA, November 1994.

[17] 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.

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

[19] 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.

[20] M. Rioux and F. Blais. "Compact three-dimentional camera for robotic application". *Journal of Optical Society of America A*, 3(9):1518–1521, September 1986.

[21] M. Subbarao and G. Surya. "Application of spatial-domian convolution/deconvolution transform for determining distance from image defocus". *Proc. of SPIE: Optics, Illumumination, and Image Sensing for Machine Vision VII*, 1822, November 1992.

[22] G. Surya and M. Subbarao. "Depth from defocus by changing camera aperture: A spatial domain approach". *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 61–67, June 1993.

[23] M. Takeda, H. Ina, and S. Kobayashi. "Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry". *Journal of Optical Society of America A*, pages 156–160, January 1982.

[24] M. Watanabe and S. K. Nayar. "Telecentric optics for constant-magnification imaging". Technical Report CUCS-026-95, Dept. of Computer Science, Columbia University, New York, NY, USA, September 1995.

[25] 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.

[26] 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.