Focus Range Sensors

Shree K. Nayar
Department of Computer Science, Columbia University New York, NY 10027, U.S.A.

Minori Noguchi, Masahiro Watanabe, Yasuo Nakagawa
Productional Engineering Research Laboratory, Hitachi Ltd. Yokohama, 244 Japan

Abstract

Structures of dynamic scenes can only be recovered using a real-time range sensor. Focus analysis offers a direct 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. A prototype focus range sensor has been developed that produces up to 512x480 depth estimates at 30 Hz with an accuracy better than 0.3%. In addition, a microscopic shape from focus sensor is described that uses the derived illumination pattern and a sequence of images to recover depth with an accuracy of 1 micron. Several experimental results are included to demonstrate the performances of both sensors. We conclude with a brief summary of our recent results on passive focus analysis.

1 Introduction

Of all problems studied in computational vision, recovery of three-dimensional scene structure has by far attracted the most attention. This has resulted in a panoply of sensors and algorithms [15] [2] 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, they are expected to prove complementary to other techniques but not serve as standalone 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. Recently, it was demonstrated that stereo could achieve real-time performance, but only with the use of significant customized hardware [16]. Further, 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 [15]. 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 [14]. All the information needed is then acquired in $log_2(N)\, T_f$, a significant improvement. An alternative approach uses color-coded stripe patterns [5]; 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 [17], 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 [12].

Here, we summarize our work on a class of range sensors that are based on focus analysis. In particular, we describe a real-time range sensor that produces high-resolution (512x480) depth maps at 30 Hz (video rate) [25] [32]. 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 by [27][29] and subsequently by others[1]. This circumvents the need for correspondence or feature tracking. The real-time sensor mentioned above 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 practical 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, in that, it requires that the scene have high frequency textures. A textureless surface appears the same focused or defocused and the resulting images do not contain information necessary for depth computation. This has prompted us to develop a 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 [9][28] for both depth from defocus and depth from pattern size distortion under perspective projection. However, these projected patterns were selected in a more or less arbitrary 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. We arrive at a solution to this problem 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.

An implementational problem that has repeatedly surfaced in previous work is the variation in image magnification that occurs when images are taken under different focus settings [33]. This manifests into a correspondence-like problem. It has forced investigators to resort to techniques varying from image registration and warping [7] to the use of precise lens calibration for correcting magnification variations [33] [7]. We have a simple but effective optical solution to this problem [31]. By appending an additional aperture to the optics, we show that the focus setting of an imaging system can be varied substantially without altering magnification.

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 [25][32]. 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

[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 displacing the sensor in small increments and uses a focus operator to detect the image of maximum focus for each scene point (see [20, 7, 22, 21, 19, 1, 34, 26]). In contrast, depth from defocus typically uses two images and estimates relative blurring to get depth (see [27, 29, 11, 28, 4, 8, 34, 10]).

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. Several 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 shown.

Finally, we describe a second system we have developed that is based on focus analysis [26]. This system is complementary to the first one, in that, it is based on depth from focus (rather than defocus) and is capable of reocovering microscopic objects with an accuracy as good as 1 micron. The starting point for this system is an algorithm described in [22], where a shape from focus method was developed for microscopic objects. The high magnification of a microscope results in images that capture brightness variations caused by the micro-structure of the surface. Most surfaces that appear smooth and non-textured to the naked eye produce highly textured images under a microscope. Examples of such surfaces are paper, plastics, ceramics, etc. Microscopic shape from focus [22] was therefore demonstrated to be an effective approach, offering solutions to a variety to challenging shape inspection problems. However, there exist surfaces that are smooth at the micro-structure level and consequently do not produce sufficient texture even under a microscope.

The illumination pattern we used in the real-time range sensor was initially incorporated into the microscopic shape from focus system [26]. The illumination pattern is projected onto the sample via the path of the bright field illumination of the microscope. As with the depth from defocus sensor, this enables very precise registration of the illumination pattern with the sampling grid of the image sensor; light rays projected out through the imaging optics are subjected to the same geometric distortions as rays reflected back to the image sensor. The motorized stage of the microscope is used to automatically acquire a set of images (typically 10-20) by moving the sample towards the objective lens. The depth from focus algorithm [22] is then applied to the image set to obtain a complete depth map of the sample. As examples, we show accurate and detailed depth maps of structures on silicon substrates and solder joints. These are samples of significant industrial import that have been found hard to recover using other vision techniques.

We conclude with a brief discussion on our most recent result on focus analysis that include a passive bifocal vision sensor [23] and a novel algorithm for passive depth from defocus [30] that uses a minimal operator set to recover scenes with unknown and complex textures.

2 Depth from Defocus

Fundamental to depth from defocus is the relationship between focused and defocused images [3]. Figure 1 shows the basic image formation geometry. All light rays that are radiated by object point P and pass the aperture A are refracted by the lens to converge at point Q on the image plane. 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:

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

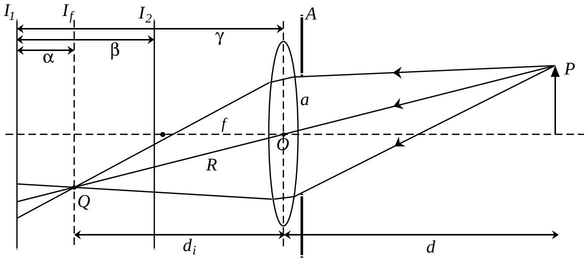


Figure 1: Image formation and depth from defocus.

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. 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= γ- α, 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. These include (a) accurate estimation of relative defocus in the two images, (b) recovery of textured **and** textureless surfaces, and (c) achieving constant magnification that is invariant to the degree of defocus.

3 Constant-Magnification Defocus

We begin with the last of the problems mentioned above. In the imaging system shown in Figure 1, the effective image location 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. We approach this problem from an optical perspective rather a computational one. Consider the image formation model shown in Figure 2. The only modification made with respect to the model in Figure 1 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 [18]. 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'. Furthermore, the lens law of (1) remains valid. This modification is realizable not only in single lens systems but any compound lens system. 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 1 are f/a and d_i/a, respectively, they are both equal to f/a' in the telecentric case.

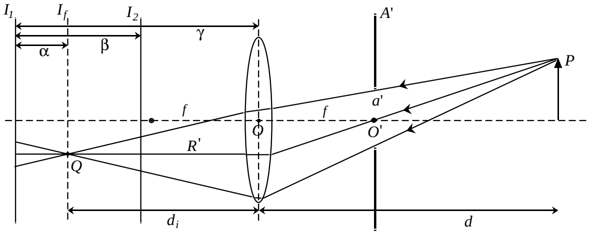


Figure 2: A constant-magnification imaging system for depth from defocus is achieved by simply placing an aperture at the front-focal plane of the optics.

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, namely, the illumination pattern, optical transfer function, defocusing, image sensing, and the focus operator. All of these together determine the the relation between the depth of a scene point and its two focus measures. Since we have used the telecentric lens (Figure 2) 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 1) by simply replacing the factor $\frac{f}{a'}$ by $\frac{d_i}{a}$. Though we use both spatial and Fourier (frequency) models of the above components, for brevity we will present Fourier models only when they are needed to make pertinent observations.

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. As we will describe shortly, the image sensor used has rectangular pixels arranged on a rectangular spatial grid. Hence, the basic building block of the model is a rectangular illuminated patch, or cell, with uniform intensity:

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

where, ${}^2\Pi()$ is the two-dimensional *Rectangular* function [6]. 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. The periodic grid is defined as:

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

where, ${}^2\mathrm{III}()$ is the 2-dimensional *Shah* function [6], and $2t_x$ and $2t_y$ determine the periods of the grid in the x and y directions. The final illumination pattern is obtained by convolving the cell $i_c(x, y)$ with the grid $i_g(x, y)$, i.e. $i(x, y) = i_c(x, y) * i_g(x, y)$. 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.

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 are related as:

$$I(u, v; b_x, b_y, t_x, t_y) = I_c(u, v) \cdot I_g(u, v) \qquad (4)$$

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 2) is of finite radius a', it does not capture the higher order diffractions radiated by the surface (see [3] for details). This effect places a limit on the optical resolution of the imaging system characterized by the optical transfer function (OTF):

$$O(u, v; a', f) \qquad (5)$$

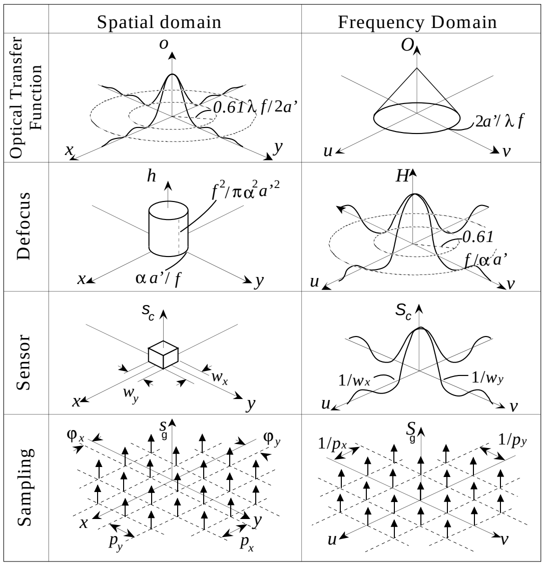


Figure 3: Spatial and frequency models for the optical and sensing elements of depth from defocus.

$$= \begin{cases} (\frac{a'}{f})^2(\gamma - \sin\gamma)\ , & \sqrt{u^2+v^2} \le \frac{2a'}{\lambda f} \\ 0\ , & \sqrt{u^2+v^2} > \frac{2a'}{\lambda f} \end{cases}$$

$$\text{where } \gamma = 2\cos^{-1}(\frac{\lambda f}{a'}\frac{\sqrt{u^2+v^2}}{2})\ .$$

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 3). This in turn places restrictions on the frequency of the illumination pattern.

4.3 Defocusing

The defocus function is described in detail in previous work (see [3][13] for example). As in Figure 2, 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 3):

$$h(x, y; \alpha, a', f) = \frac{f^2}{2\pi a'^2 \alpha^2}\Pi(\frac{d}{2a\alpha}\sqrt{x^2+y^2}) \quad (6)$$

where, a' is the radius of the telecentric lens aperture. The Fourier transform of the defocus function is:

$$\begin{aligned} & H(u,v;\alpha,a',f) & (7) \\ = \quad & \frac{f}{2\pi a'\alpha\sqrt{u^2+v^2}} J_1(\frac{2\pi a'\alpha}{f}\sqrt{u^2+v^2}) \end{aligned}$$

where J_1 is the first-order Bessel function [3]. 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. 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. Hence, 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 that can be modeled as a rectangular array of rectangular sensing elements (pixels). The quantum efficiency [13] 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:

$$s_c(x,y;w_x,w_y) ={}^2\,\Pi(\frac{1}{w_x}x,\frac{1}{w_y}y) \qquad (8)$$

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:

$$\begin{aligned} & s_g(x,y;p_x,p_y,\varphi_x,\varphi_y) & (9) \\ = \quad & \frac{1}{p_x p_y}{}^2\mathrm{III}(\tfrac{1}{p_x}(x-\varphi_x),\tfrac{1}{p_y}(y-\varphi_y)) \end{aligned}$$

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:

$$m_d(x,y) = (s_c(x,y) * m(x,y)) \cdot s_g(x,y) \qquad (10)$$

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 phases (φ_x,φ_y) of the sampling function is with respect to the illumination pattern and are also viewed as parameters to be optimized. In Fourier domain, the final discrete image is:

$$M_d(u,v) = (S_c(u,v)\cdot M(u,v)) * S_g(u,v) \qquad (11)$$

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 purpose 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 3x3 discrete Laplacian is:

$$\begin{aligned} & l(x,y;q_x,q_y) \\ = \quad & 4\delta(x)\cdot\delta(y) - [\delta(x)\cdot\delta(y-q_y) + \delta(x) & (12) \\ & \cdot\delta(y+q_y) + \delta(x-q_x)\cdot\delta(y) + \delta(x+q_x)\cdot\delta(y)] \end{aligned}$$

Here, q_x and q_y are the spacings between neighboring elements of the discrete Laplacian kernel and are given by the image sensor. The Fourier transform of the Laplacian is:

$$L(u,v;q_x,q_y) = 4 - 2\cos(2\pi q_x u) - 2\cos(2\pi q_y v) \qquad (13)$$

The required discrete nature of the focus operator comes with a price. It tends to broaden the bandwidth of the operator. Once the illumination 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:

$$i(x,y;b_x,b_y,t_x,t_y)*o(x,y;a',f)*h'(x,y;\alpha',a',f) \tag{14}$$

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:

$$\begin{aligned}&\{i(x,y;b_x,b_y,t_x,t_y)*o(x,y;a',f)^{*2}\\&*h'(x,y;\alpha',a',f)*h(x,y;\alpha,a',f)\\&*s_c(x,y;w_x,w_y)\}\cdot s_g(x,y;p_x,p_y,\varphi_x,\varphi_y)\end{aligned} \tag{15}$$

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:

$$g(x,y)=\{(i*o^{*2}*h^{*2}*s_c)\cdot s_g\}*l \tag{16}$$

Since the distance between adjacent weights of the Laplacian kernel must be integer multiples of the period of the image sampling function s_g, (16) can be rearranged as:

$$g(x,y)=(i*o^{*2}*h'*h*s_c*l)\cdot s_g=g_0\cdot s_g \tag{17}$$

where, $g_0=i*o^{*2}*h'*h*s_c*l$. Alternately, in Fourier domain we have:

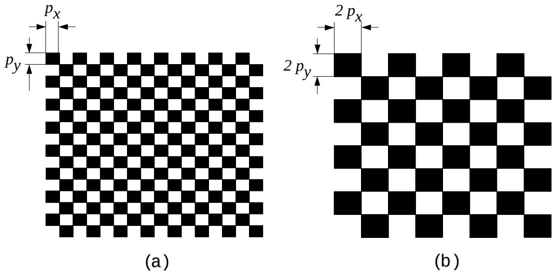
$$G(u,v)=(I\cdot O^2\cdot H'\cdot H\cdot S_c\cdot L)*S_g=G_0*S_g \tag{18}$$

The above expression gives us the final output of the focus operator for any value of the defocus parameter α.

5 Optimization

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 $g(x,y)$. 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.

The details of the optimization process are given in [24] and will be omitted in the interest of space. Here, we briefly outline the arguments we have used to arrive at the optimal pattern. 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. Our objective then is to maximize the fundamental spatial frequency $(1/t_x,1/t_y)$ of the illumination. 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 harmonics 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 $(\pm1/t_x,\pm1/t_y)$. Using this observation, the illumination pattern parameters $(b_x,b_y,t_x,t_y,)$ and the illumination phase shift (φ_x,φ_y) that maximize $\|\ G(\frac{1}{t_x},\frac{1}{t_y})\ \|$ are determined in [24]. Two optimal patterns were found and are shown in Figure 4. Exactly how such high resolution patterns can be projected and perfectly registered with the image detector will be described in the experimental section.



(a)

(b)

Figure 4: 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 sensing grid.

6 Tuned Focus Operator

For the purpose of illumination optimization, we used the Laplacian. 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.

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. (c) The operator must not respond to any DC component in image brightness. The above conditions, yield a set of equations with the operator elements as variables [32]. These equations were solved to find that the operator with its symmetric structure has only one variable. This variable was optimize so as to yield a frequency response with sharpest peaks, i.e. power spectrum with the lowest second moment around the illumination frequency $(\pm 1/t_x, \pm 1/t_y)$. This tuned focus operator was found to have substantially sharper peaks than the discrete Laplacian.

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 2. 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 $g_2(x, y,)$. Since the image now has a single dominant frequency, namely $(\pm 1/t_x, \pm 1/t_y)$, a relation between the focus measures and defocus α can be derived using (18):

$$q = \frac{g_1 - g_2}{g_1 + g_2} = \frac{H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha) - H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha - \beta)}{H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha) + H(\frac{1}{t_x}, \frac{1}{t_y}; \alpha - \beta)} \tag{19}$$

As shown in Figure 5, q is a monotonic function of α such that $-p \leq q \leq p$, $p \leq 1$. In practice, the above relation can be pre-computed and stored as a look-up table that maps q 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.

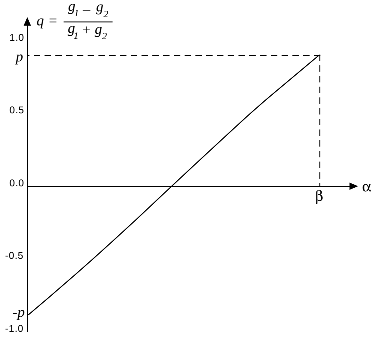


Figure 5: Relation between focus measures g_1 and g_2 and the defocus parameter α.

8 Real Time Range Sensor

Based on the above results, we have implemented [25][32] the real-time focus range sensor shown in Figure 6. 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 4(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 pix-

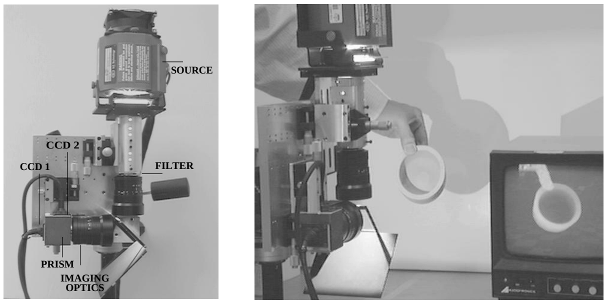


Figure 6: (a) The real-time focus range sensor and its key components. (b) The sensor can produce depth maps up to 512x480 in resolution at 30 Hz.

els of the two CCD cameras. 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.

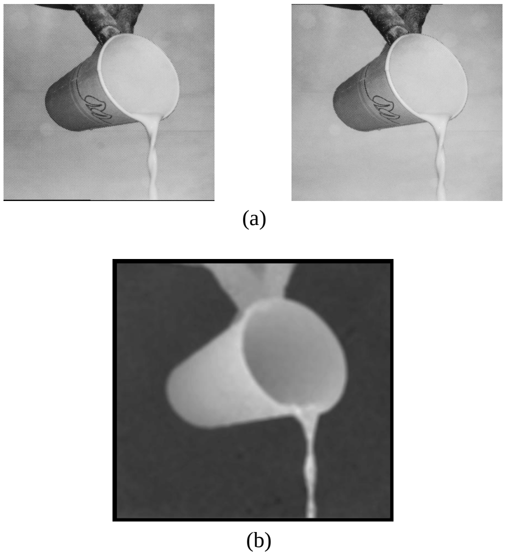
Images from the two CCD cameras are digitized and processed using MV200 Datacube image processing hardware. The present configuration includes the equivalent of two 8-bit digitizers, two A/D convertors, and one 12-bit convolver. This hardware enables simultaneous digitization of the two images, convolution of both images with the tuned focus operator, and the computation of a 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 to a unique depth estimate (see [32] for details).

Several experiments were conducted on both textured and textureless surfaces to test the performance of the sensor [25] [32]. The performance evaluation results are summarized in Table 1 and discussed in detail in [32].

	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

Table 1: Performance characteristics of the sensor.

As stated earlier, structures of dynamic scenes can only be recovered using a real-time sensor such as the one we have developed. Figure 7 illustrates the power of such a high-speed, high-resolution sensor. The figure shows two brightness images and the computed depth map of a cup with milk flowing out of it.


(a)

(b)

Figure 7: (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 8 shows a scene with polyhedral objects. The computed depth map in Figure 8(b) is fairly accurate despite the complex textural properties of the objects. All surface discontinuities and orientation discontinuities are well preserved. Figure 9 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.

9 Microscopic Shape from Focus

The optimal illumination filter shown in Figure 4(b) was also incorporated into the shape from focus system developed in [22]. A set of sample images are obtained by automatically moving the microscope stage in increments of $\Delta z = z_i - z_{i-1}$. The Laplacian focus operator is applied to each image to obtain a set of focus measure values at each image point (x, y); the number of focus measures equals the number of images taken. The dis-

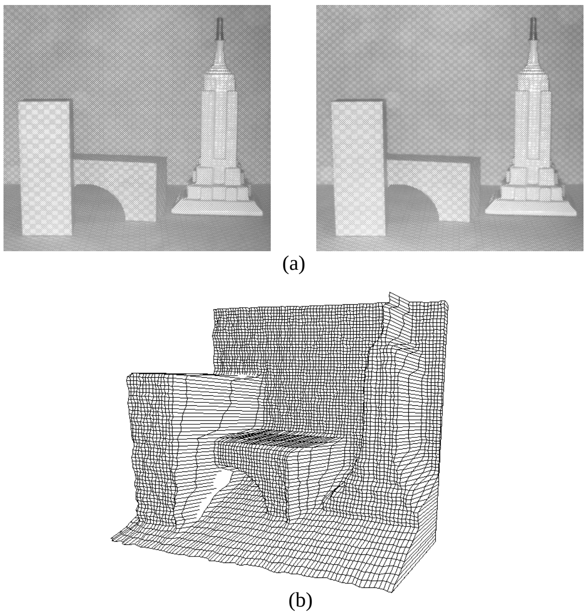


Figure 8: (a) Near and far focused images of a set of polyhedral objects. (b) Computed depth map.

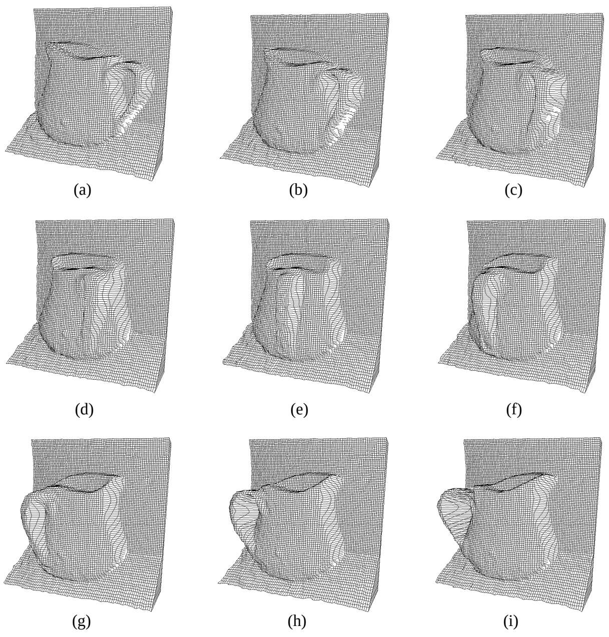


Figure 9: Depth maps generated by the sensor at 30 Hz while an object rotates on a motorized turntable.

crete stage position z_j that yields the maximum focus measure value at an image point, can be used as an approximation of the depth of the corresponding surface point. A more accurate depth estimate z is obtained by applying Gaussian interpolation [22] to the three focus measures corresponding to z_{j-1}, z_j, and z_{j+1}.

It was shown in [22] that at least 3 focus measures are needed for depth estimation by Gaussian interpolation. In the active illumination system proposed here, the fundamental frequency of the surface texture is determined by the illumination filter used and is simply $\sqrt{(\frac{1}{t_x})^2 + (\frac{1}{t_y})^2}$. For this frequency the defocus measure has a zero-crossing at defocus value α' such that $\frac{2\pi a \alpha'}{d}\sqrt{(\frac{1}{t_x})^2 + (\frac{1}{t_y})^2} = 3.83$. This gives us an upper limit on the usable defocus range for any point on the surface, which is, $-\alpha' \leq \alpha \leq \alpha'$. Therefore, on the image side of the microscope optics, we need to obtain (for any surface point) at least 3 images within the above defocus range. This gives us the following maximum distance between consecutive focused images on the image side of the optics:

$$\Delta z' \leq \frac{2\alpha'}{4} = \frac{1}{2} \cdot 3.83 \cdot \frac{d}{2\pi a \sqrt{(\frac{1}{t_x})^2 + (\frac{1}{t_y})^2}} \qquad (20)$$

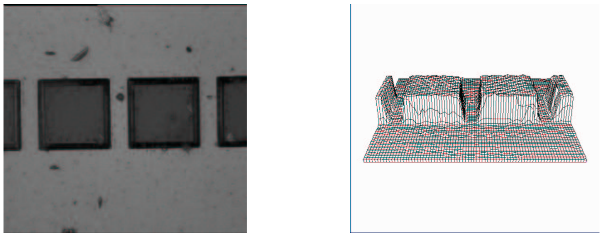
This distance on the image side is related to the maximum allowable microscope stage displacement (between consecutive images) on the sample side of the optics by the magnification M of the objective lens used:

$$\Delta z \simeq \Delta z' \frac{1}{M^2} \qquad (21)$$

In our experiments, the magnification of the objective lens is $M = 20$, the ratio $a/d = 0.025$, and illumination parameters are $t_x = 44\ \mu m$, $t_y = 52\ \mu m$. Using these values in eqs.(20) and (21) we get the maximum allowable stage displacement $\Delta z \leq 0.5\ \ \mu m$. Experiments reported in [26] illustrate that $\Delta z = 0.5\ \ \mu m$ does in fact produce the best results. Further decreasing Δz does not significantly improve the accuracy of the depth maps.

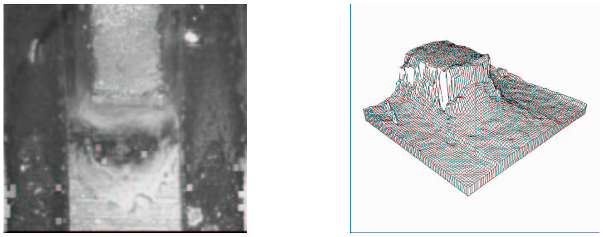
The sample in Figure 10 has rectangular structures fabricated on a smooth silicon wafer. Silicon wafer inspection is of great relevance in a variety of chip manufacturing processes. The surface of the wafer is very smooth, resulting in images (see Figure 10(a)) that are more or less textureless. This renders the original shape from focus system

[22] ineffective. A total of 16 images were taken for this sample. The derived illumination pattern produces very accurate shape information that is far superior to that produced by the bright field illumination of the microscope [26]. Similar results were obtained for the solder joint sample shown in Figure 11. For this sample, an objective lens with $M = 10$ was used and a total of 23 sample images were taken. This sample exhibits noticeable texture under a microscope. However, the texture is not consistent over the entire surface. As a result, the illumination pattern is necessary to get an accurate depth map. Solder shape inspection has remained a challenging and unresolved industrial problem. These results indicate that microscopic shape from focus may provide an effective solution to this important problem.



(a) Camera image (b) Depth Map

Figure 10: Image and depth map of rectangular structures on a smooth silicon substrate. The structures are approximately $13\mu m$ tall.



(a) Camera image (b) Depth Map

Figure 11: Image and depth map of a solder joint on a circuit board. The solder joint is approximately $150\mu m$ high and $100\mu m$ wide.

10 Summary

We have summarized our results on a variety of issues related to depth estimation by focus analysis. Accurate modeling of optics and sensing were shown to be essential for 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 optical 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 second system we described targets a different class of objects, namely, microscopic structures. Using the derived illumination pattern, we have demonstrated a fully automated microscopic shape from focus system that can recover depth with an accuracy within 1 micron. This system has well-defined applications in the industrial arena, where a depth sensor for samples such as silicon wafers and solder joints is much sought after.

The obvious extension to this work is the development of passive focus range finders for both indoor as well as outdoor scenes. We have already implemented a passive bifocal vision sensor and are in the process of evaluating its capabilities [23]. Such a sensor cannot afford the luxury of projected illumination. It must rely on complex scene textures for depth estimation. In this regard, we have recently developed an efficient depth from defocus algorithm [30] that uses a minimal set of operators to recover structures of scenes with unknwown textures.

References

[1] N. Asada, H. Fujiwara, and T. Matsuyama. Edge and depth from focus. *Proc. of Asian Conf. on Comp. Vis.*, pages 83–86, Nov. 1993.

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

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

[4] V. M. Bove, Jr. Entropy-based depth from focus. *Jrn. of Opt. Soc. of Am. A*, 10:561–566, Apr. 1993.

[5] 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.

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

[7] T. Darrell and K. Wohn. Pyramid based depth from focus. *Proc. of IEEE Conf. on Comp. Vis. and Patt. Rec.*, pages 504–509, June 1988.

[8] J. Ens and P. Lawrence. A matrix based method for determining depth from focus. *Proc. of IEEE Conf. on Comp. Vis. and Patt. Rec.*, pages 600–609, June 1991.

[9] B. Girod and S. Scherock. Depth from focus of structured light. *Proc. of SPIE: Optics, Illum., and Image Sng for Mach. Vis. IV*, 1194, Nov. 1989.

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

[11] P. Grossman. Depth from focus. *Patt. Recog.*, 9(1):63–69, 1987.

[12] 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.

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

[14] 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.

[15] 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.

[16] T. Kanade. Development of a Video-Rate stereo machine. *Proc. of ARPA Image Understanding Workshop*, pages 549–557, Nov. 1994.

[17] 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.

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

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

[20] E. Krotkov. Focusing. *Intl. Jrnl. of Comp. Vis.*, 1:223–237, 1987.

[21] H. N. Nair and C. V. Stewart. Robust focus ranging. *Proc. of IEEE Conf. on Comp. Vis. and Patt. Rec.*, pages 309–314, June 1991.

[22] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 16(8):824–831, Aug. 1994.

[23] S. K. Nayar and M. Watanabe. Passive bifocal vision sensor. Technical Report (in preparation), Dept. of Computer Science, Columbia University, New York, NY, USA, Dec 1995.

[24] 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.

[25] 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.

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

[27] A. Pentland. A new sense for depth of field. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 9(4):523–531, July 1987.

[28] A. Pentland, S. Scherock, T. Darrell, and B. Girod. Simple range cameras based on focal error. *Jrnl. of Opt. Soc. of Am. A*, 11(11):2925–2935, Nov. 1994.

[29] M. Subbarao. Parallel depth recovery by changing camera parameters. *Proc. of Intl. Conf. on Comp. Vis.*, pages 149–155, Dec. 1988.

[30] M. Watanabe and S. K. Nayar. Rational filters for passive depth from defocus. Technical Report CUCS-035-95, Dept. of Computer Science, Columbia University, New York, NY, USA, September 1995.

[31] 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.

[32] M. Watanabe, S. K. Nayar, and M. Noguchi. Real-time implementation of depth from defocus. *Proceedings of SPIE Conference*, October 1995.

[33] 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, Jan. 1994.

[34] 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.