Efficient Depth Recovery through Inverse Optics

Muralidhara Subbarao
SUNY at Stony Brook

In the case of machine vision . . . one ought to understand image formation if one wishes to recover information about the world from images.
– B. K. P. Horn(1986)[7].

Abstract

The image of a scene formed by an optical system such as a lens contains both *photometric* and *geometric* information about the scene. 'Inverse Optics' is the problem of recovering this information from a set of images sensed by the camera. Previous solutions to this problem– the depth-from-focusing methods– required a large number (in principle, infinitely many) of images to be recorded and processed. Hence the methods were slow and computationally intensive. Recent work in this area suggests solutions that require only a few images and therefore are fast and computationally efficient. Here we present a coherent view of recent developments. Theoretical principles, practical issues, and unsolved problems are discussed. Preliminary experimental results are presented.

1 Introduction

1.1 Lens based inverse optics is a well-posed problem

One of the early goals of a visual system is to recover the three-dimensional geometry of scenes. In machine vision, most of the research for recovering the scene geometry is based on a a pin-hole camera model (e.g.: [1, 15, 7]). While the image of a pin-hole camera provides *photometric information* (i.e. scene radiance along different directions of view), it completely lacks *geometric information* (i.e. the distance of visible surfaces along different viewing directions). Therefore, analyses based on a pin-hole camera model have to use heuristic assumptions about the scenes to recover geometric information. For example, in the shape-from-shading process, assumptions are made about the reflectance and shape of visible surfaces (e.g.: the Lambertian reflectance model and "smoothness" of surface structure).

Practical camera systems, and also the human eye, are not pinhole cameras but consist of convex lenses. In contrast to a pin-hole camera, the image formed by a lens contains both photometric and geometric information. For an aberration-free convex lens, (i) the radiance at a point in the scene is proportional to the irradiance at its *focused image* [7], and (ii) the position of the point in the scene and the position of its focused image are related by the *lens formula*

$$\frac{1}{f} = \frac{1}{u} + \frac{1}{v} \tag{1}$$

where f is the focal length, u is the distance of the object from the lens plane, and v is the distance of the focused image from the lens plane (see Figure 1). Given the irradiance and the position of the focused image of a point, its radiance and position in the scene are uniquely determined. In fact the positions of a point-object and its image are *interchangeable,* i.e. the image of the image is the object itself. Now, if we think of an object surface in front of the lens to be comprised of a set of points, then the focused images of these points define another surface behind the lens (see Figure 1). We can think of this surface and the image irradiance on it as the *focused image.* There is a *one to one correspondence* between this focused image and the object surface. The geometry (i.e. the shape) and the radiance distribution of the object surface is uniquely determined by the focused image.

In this paper we are concerned with the principles and computational methods for recovering the geometry and the radiance of an object from its *sensed image.* (Note that a sensed image is in general

Figure 1: Image formed by a convex lens.

quite different from the focused image of an object.) This recovery involves *inverting* the image formation process in a lens based camera. We will term this inversion process as *inverse optics.* We shall see that, under suitable conditions, inverse optics is a *well-posed problem,* though, perhaps, *ill-conditioned.*

1.2 Depth-from-focusing

In the depth-from-focusing method (e.g. [6, 8, 16, 9]), the lens formula (1) is used for finding the distance of objects whose images are in focus. Many approaches exist for focusing an object. These approaches are primarily found in the autofocusing literature for cameras and microscopes (e.g.: [10, 24]).

In the depth-from-focusing method, an object is focused onto the image detector by continuously varying one or both of the following camera parameters: (i) distance between the lens and image detector, and (ii) the focal length. For each setting of these camera parameters, one image is recorded and processed to compute a "sharpness" measure. The image which gives a global maximum for the sharpness measure is taken to be the focused image. A simple measure of sharpness of an image $g(x, y)$ is its energy $\int\int g^2(x, y)dx\ dy$. It can be shown that the global maximum of this measure corresponds to the focused image (e.g. [18]). Measures based on the derivatives of images can also be used. These measures have been found to be more reliable because, usually, the attenuation caused by blurring increases with increasing spatial frequency while the amplification caused by derivatives increase with increasing spatial frequency.

The depth-from-focusing method is inherently slow and computationally intensive because it involves recording a large number of images, computing the sharpness measure for each image, and then finding the global maximum of the sharpness measure.

In this paper we are mainly concerned with methods that do not require an object to be focused in order to find its distance or its radiance. The approach to be presented here requires recording and processing only a few images and therefore is much faster than the depth-from-focusing method. It is also faster than active methods such as laser ranging and does not suffer from the *correspondence problem* of stereo ranging method. The computations involved are simple and has the potential for hardware implementation.

1.3 Organization

This paper is primarily based on our work reported in [18, 19, 20, 21, 23]. The ideas and results there have been refined in many respects and reorganized in a major way. Many details there have been left out. The net result (we are afraid!) is that the approach here comes out as quite simple and straightforward. The most closely related work to this paper is Pentland's [12, 13, 14].

We begin our discussion with methods for simple objects such as points, lines, etc. and proceed in steps to increasingly complex ones. First we focus on the basic principles of the approach. Many details of the approach are delayed until much later to avoid distraction.

Figure 2: Camera geometry.

2 Points, Lines, and Edges

The class of objects we deal in this section may be too simple to be of much use in practice, but the principles presented here lay the foundation for the following sections. We also present experimental results which verify the mathematical model developed here.

Consider a convex lens camera as shown in Figure 2. Let P be a point object in front of the lens and p' be its focused image. The relation between the positions of P and p' is given by the lens formula (1). If P is not in focus then it gives rise to a circular image called *blur circle* on the image detector. From simple plane geometry (see Figure 2) and the lens formula (1) we can show that the diameter d of the blur circle is

$$d \;=\; D\, s \left(\frac{1}{f} - \frac{1}{u} - \frac{1}{s}\right) \tag{2}$$

where D is the diameter of the lens aperture and s is the distance from the lens to the image detector. Note that d can be either positive or negative depending on whether $s \geq v$ or $s < v$. In the former case the image detector is *behind* the focused image p' and in the latter case it is *in front* of it. According to geometric optics [7, 3], the intensity[1] within the blur circle is approximately constant. If b is the brightness of P when focused, then its blurred image is

$$h_1(x,y) \;=\; \begin{cases} \frac{4b}{\pi d^2} & \text{if } x^2+y^2 \le \frac{d^2}{4} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Note that

$$\int\!\!\int h_1(x,y)\; dx\; dy \;=\; b. \tag{4}$$

If P is an isolated point object with no other light source in its vicinity, then its distance can be determined from the diameter of its blur circle. Knowing the camera parameters s, f, D and the blur circle diameter d, the distance u is obtained from equation (2). The diameter d itself can be either measured directly on the image detector or estimated from the brightness within the blur circle. First b is computed from equation (4). This provides photometric information about P (i.e. b is proportional to the brightness of P). If b_0 is the brightness within the blur circle then d is obtained from equation (3) as

$$d \;=\; \pm 2\sqrt{b/\pi b_0}. \tag{5}$$

However the sign of d remains ambiguous. This ambiguity can be resolved in two ways: (i) by setting $s = f$, or (ii) by changing s by a small value and observing the change in the magnitude of d. In the first case the sign of d is always negative (since $u > 0$ in equation (2)). In the second case, from Figure 2 we see that, the magnitude of d increases for a small increase in s if the image detector is behind the focused point p' (i.e. $s \ge v$ and consequently d is positive) and the magnitude of d decreases otherwise. Therefore d has the same sign as the derivative $d|d|/ds$. Therefore, under ideal conditions, finding the distance of a point object is straightforward.

In practice, the image of a point object is not a crisp circular patch of constant brightness as suggested by geometric optics. Instead, due to diffraction, lens aberrations, and noise, it will be a

[1] We shall use the terms ‘brightness’, ‘intensity’, ‘radiance’, ‘irradiance’, ‘grey-level’, etc. interchangeably, relying on context to convey the intended meaning.

roughly circular blob with the brightness falling off gradually at the border rather than sharply. Because of these non-idealities, a two-dimensional Gaussian is often suggested as a model for the image of a point object [7, 17] (also see [22] for the derivation of a model from diffraction theory). For a point object of unit brightness (i.e. $b = 1$), the intensity distribution of its image under the Gaussian model would be

$$h_2(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{6}$$

where σ is the spread parameter. We shall not restrict ourselves to the Gaussian model at this point. Indeed our experiments indicate that Gaussian is not a good model for our camera. However we shall use this model later to illustrate certain concepts.

Let $h(x, y)$ be the intensity distribution of the image of the point object P. Let the brightness of the focused image of P be unity (i.e. $b = 1$) so that

$$\int\int_I h(x, y)\, dx\, dy = 1 \tag{7}$$

where I is a region on the image detector containing the entire image of P. $h(x, y)$ as defined here is indeed the *point spread function* of the camera. We shall assume that $h(x, y)$ is *circularly symmetric*. This is largely true of cameras with circular apertures. In order to determine the distance u, we need to relate $h(x, y)$ to the camera parameters s, f, D, and the distance u. For this purpose, we characterize h with a single parameter σ defined to be the *standard deviation* of the distribution $h(x, y)$, i.e.

$$\sigma^2 = \int\int_I [(x - \bar{x})^2 + (y - \bar{y})^2]\, h(x, y)\, dx\, dy \tag{8}$$

where $(\bar{x}, \bar{y})$ is the *center of mass* defined by

$$\bar{x} = \int\int_I x\, h(x, y)\, dx\, dy \quad \text{and} \quad \bar{y} = \int\int_I y\, h(x, y)\, dx\, dy\,. \tag{9}$$

σ is the *square root of the second central moment* of $h(x, y)$. It can also be thought of as the *radius of gyration* of h about its *center of mass*. We shall call σ the *spread parameter* of the point spread function $h(x,y)$. Sometimes we also refer to σ as the *blur parameter* as it is an indicator of the degree of blur.

Now we need a relation (analogous to equation (2)) between σ, the camera parameters, and the distance u. The actual derivation of such a relation from theoretical considerations appears to be complicated and therefore will not be considered. However we hypothesize that

$$\sigma \;=\; k\,d \;\text{ for }\; k > 0 \tag{10}$$

for some constant k where d is the diameter of the blur circle given by equation (2). Our experiments, to be described shortly, support this hypothesis strongly. The proportionality constant k is characteristic of a given camera and can be determined through calibration.

Using equations (2,10) one can determine the distance u of a point object from its image $h(x,y)$ and the camera parameters k, s, f, D. The explicit expression for u is

$$u \;=\; \frac{kDsf}{kD(s-f) - f\sigma}\,. \tag{11}$$

Direct experimental verification of this method poses some practical difficulties. For example, realizing a point object of sufficient brightness, and accurate measurement of σ, are difficult due to noise, quantization, and digitization effects. Therefore we verify this through the image of a step edge.

First let us define what we mean by *focused image on the image detector* in a general case. For any point p (see Figure 2) on the image detector, consider a line through that point and the optical center. Let P be the point on a visible surface in the scene whose focused image lies on this line. Let p' be the focused image of P. Then the intensity of the *focused image on the image detector* at p is the intensity of the focused image at p'. In the rest of this paper, we abbreviate 'focused image on the image detector' to just 'focused image'.

Consider a planar object normal to the optical axis at a distance u in front of the lens. Let its focused image be $f(x,y)$ which is a step edge along the y-axis on the image detector. Let a be the image intensity to the left of the y-axis and b be the height of the step. The image can be expressed as

$$f(x, y) \;=\; a \;+\; b\, u(x) \tag{12}$$

where $u(x)$ is the standard *unit step function.* If $g(x, y)$ is the observed image, then, assuming the camera to be a *linear shift-invariant system* (cf. [15]), we have

$$g(x, y) \;=\; h(x, y) \otimes f(x, y) \tag{13}$$

where $\otimes$ represents the convolution operation. Note that, if $a = 0$ and $b = 1$ then $g(x, y)$ gives the *edge spread function* of the camera, i.e. the response of the camera to a unit step edge. The response of the camera to a line (e.g. $\delta(x)$ on the $x - y$ plane) is called the *line spread function* of the camera. Relations between the three spread functions– *point, line,* and *edge*– are well known in the image processing literature [15, 7]. We will not elaborate on this here. Using these relations it can be shown that (see [23] for detailed derivations) the line spread function $\theta(x)$ can be obtained from the observed image $g(x, y)$ from the expression

$$\theta(x) \;=\; \frac{\frac{\partial g}{\partial x}}{\int_{-\infty}^{\infty} \frac{\partial g}{\partial x}\, dx}\;. \tag{14}$$

The point spread function $h(x, y)$ can be obtained from the above line spread function using *Abel Transform* (because $h(x, y)$ is circularly symmetric; see [7]). However this involves taking derivative of the line spread function and the resulting $h(x, y)$ becomes highly unstable due to noise and discretization. Note that (at least for now) we are only interested in finding the standard deviation σ of the distribution $h(x, y)$, not $h(x, y)$ itself. If σ_l is the standard deviation of the distribution of the line spread function $\theta(x)$, then it can be shown that (see [23])

$$\sigma \;=\; \sqrt{2}\, \sigma_l\;. \tag{15}$$

Therefore σ can be estimated directly from the line spread function $\theta(x)$; it is not necessary to compute the point spread function $h(x, y)$!

The relation between σ_l and the camera parameters is obtained from equations (2,10,15) as

$$\sigma_l \;=\; \frac{kDs}{\sqrt{2}}\;\left(\frac{1}{f}-\frac{1}{u}-\frac{1}{s}\right) \tag{16}$$

This equation suggests that, for a given setting of camera parameters, the relation between σ_l and the distance u can be expressed in the form

$$\sigma_l \;=\; mu^{-1}+c \tag{17}$$

where m, c are some camera constants (which depend on the actual values of k, s, f, and D). These constants can be determined through calibration. The important point to note here is that *the spread parameter σ_l is linearly related to inverse distance.* Therefore, having determined the spread parameter from the observed image, the distance can be easily computed.

The discussion above suggests that, in addition to the distance of point objects, the distances of line objects and also step edges can be obtained from their blurred pictures.

Experiments were conducted with the following intent: (i) to verify the applicability of the mathematical model to practical camera systems, and (ii) to test the usefulness of the method in practical applications. Black and white sheets of papers were pasted on a cardboard to create a step intensity edge. Many images of this step edge were acquired with a Panasonic CCD camera (focal length 16mm, aperture diameter 11.4mm) by keeping the camera parameters fixed and varying the distance of the cardboard from the camera. Two of these pictures are shown in Figure 3. The range of distance variation was from 8 inches to about 8 feet. For each image the standard deviation σ_l was computed and plotted against the reciprocal of distance. Typical results for one set of nine pictures is shown in Figure 4. We see that the graph is linear on either side of the focused position. Although the spread parameter should be zero at the focused position, it is about one pixel due to non-idealities such as lens aberrations and discretization effects. The same experiment was carried out on three more sets of pictures with different camera parameter settings (by changing s, the lens to image detector distance). In all cases

Figure 3: Pictures of a blurred step edge

Figure 4: Plot of spread parameter *vs* inverse distance.

the same linear behavior was observed. See [23] for more details on the experiments and a discussion of the results.

The experiments for obtaining the graph in Figure 4 can be considered as camera calibration. Given this graph, it is now straightforward to find the distance of a new object. The image of the object is acquired and the spread parameter σ_l is computed for it. From this computed σ_l, we just read off the distance u from the graph. However we see that there could be a two-fold ambiguity corresponding to the two line segments in the graph. This ambiguity can be resolved by the same two methods mentioned earlier (i.e. setting $s = f$, or observing the sign of the derivative $d|\sigma_l|/ds$; the latter method requires two images with slightly different values of s).

We see from the graph that the linear behaviour predicted by the mathematical model (equation (17)) holds remarkably well. This verifies the hypothesis in equation (10) and suggests that the mathematical model is applicable to practical camera systems.

Pentland [12, 13, 14] and Grossman [5] both addressed the problem of recovering depth from blurred edges. Pentland's method is restricted to the case where the point spread function of the camera can be approximated by a two-dimensional Gaussian. Also the computational algorithm of Pentland is relatively complicated in comparison with the above method. Grossman [5] showed experimentally that useful depth information can be obtained from blurred edges. However he did not provide a theoretical justification for his computational algorithm.

3 Finite Planar Object with Known Brightness Pattern

We now consider a slightly more complicated case as compared to the case of points, lines, and edge objects considered above. We consider a planar object with arbitrary, but known brightness pattern. The object is taken to be placed normal to the optical axis at a distance u and surrounded by a dark background (or background with constant brightness). The object should be small enough such that, even when it is blurred, its *entire image* is sensed by the image detector (note:

a focused image "spreads" when blurred and hence becomes larger; for example, a point object spreads into a circle when blurred).

As before, let $f(x, y)$ denote the focused image, $g(x, y)$ the observed image, and $h(x, y)$ the point spread function. Also, let $F(\omega, \nu)$, $G(\omega, \nu)$ and $H(\omega, \nu)$ be their respective Fourier transforms. The functions f, g and h are related according to equation (13). Therefore their Fourier transforms are related as follows (because convolution in the spatial domain is equivalent to multiplication in the Fourier domain):

$$G(\omega, \nu) \;=\; H(\omega, \nu)\; F(\omega, \nu)\;. \tag{18}$$

Now, if the focused image $f(x, y)$ is known, then the point spread function can be obtained through deconvolution. This operation in the Fourier domain is

$$H(\omega, \nu) \;=\; \frac{G(\omega, \nu)}{F(\omega, \nu)}\;. \tag{19}$$

$H(\omega, \nu)$ above is called the *optical transfer function* of the camera. Its inverse Fourier transform gives the point spread function $h(x, y)$. The spread parameter σ can be computed from $h(x, y)$. It may also be possible to compute σ directly from the optical transfer function without computing its inverse Fourier transform.

To illustrate this method, consider the case where the point spread function is a Gaussian as in equation (6). The corresponding optical transfer function is

$$H(\omega, \nu) \;=\; e^{-\frac{1}{2}(\omega^2+\nu^2)\sigma^2} \tag{20}$$

where ω, ν are spatial frequencies in radians per unit distance. Having obtained $H(\omega, \nu)$ from the focused and the observed images (according to equation (19)), we can solve for σ from equation (20):

$$\sigma^2 \;=\; \frac{-2}{(\omega^2+\nu^2)}\;\ln H(\omega, \nu)\;. \tag{21}$$

In principle, according to the above equation, measuring $H(\omega, \nu)$ at a single point (ω, ν) is sufficient to obtain the value of σ. However, in practice, a more robust estimate can be obtained by taking the average over some domain in the frequency space:

$$\sigma^2 \;=\; \frac{-2}{A} \int\int_R \frac{\ln H(\omega,\nu)}{\omega^2+\nu^2}\; d\omega \; d\nu \tag{22}$$

where R is a region in the (ω, ν) space not containing points where $H(\omega, \nu) \leq 0$, and A is the area of R. Having obtained σ, the distance u is determined using equation (11).

Pentland [12, 13, 14] was perhaps the first to address the problem considered in this section. Pentland suggested that a close approximation to the focused image can be obtained by setting the aperture diameter D to be nearly zero. In this case the camera effectively acts like a pin-hole camera. From equations (2,10) we see that the spread σ of the point spread function is proportional to D. Therefore, when D is reduced to pin-hole dimensions, the spread σ becomes very small. Consequently the observed image will closely resemble the focused image.

Pentland's analysis of the problem is restricted to the case of a Gaussian point spread function. Also, his computational method for finding u is slightly complicated in comparison with the method presented above.

4 Finite Planar Object with unknown Reflectance Pattern

The method described in the previous section requires the knowledge of the focused image $f(x, y)$. Although an approximation to the focused image can be obtained by setting the aperture diameter to be very small, this poses some serious practical difficulties. First, the diffraction effects increase as the aperture diameter decreases, thus distorting the recorded image. Second, a small aperture gathers only a small amount of light and therefore the exposure period of the film will have to be increased correspondingly. The exposure period is approximately proportional to the reciprocal of the aperture diameter squared. This could slow down the method.

In this section we present a general method that does not require the knowledge of the focused image. Therefore it is faster and more practical than the previous method. The requirement of the focused image is avoided by processing two pictures acquired with different camera parameter settings.

Let $g_1(x,y)$ and $g_2(x,y)$ be the observed images for two different camera parameter settings: s_1, f_1, D_1 and s_2, f_2, D_2. g_1 and g_2 will have different spatial magnification if $s_1 \neq s_2$ (see Figure 2). In this case their magnifications will have to be made the same. In the following discussion we shall assume that all images are scaled to have unit magnification. This scaling is described by the transformation:

$$g_n(x/s, y/s) \;=\; g_o(x,y) \tag{23}$$

where $g_o(x,y)$ is the original observed image with the distance from the lens to image detector being s, and $g_n(x,y)$ is the scaled image with unit magnification. (Applying this magnification correction for digital pictures is a little tricky, but can be done through an appropriate weighted averaging (or interpolation) scheme. We leave it to the interested reader to figure it out!)

Analogous to the normalization of spatial magnification is the grey-level rescaling. The pictures g_1 and g_2 are normalized to have the same mean grey value. This step compensates for variation in mean brightness due to change in the camera parameters (e.g. a smaller aperture produces a dimmer picture, unless the exposure period is increased correspondingly). (This grey level normalization should be applied after correcting for the vignetting effect.)

For an image whose magnification has been normalized to unity, the expression for the blur circle diameter d in equation (2) also needs to be normalized by dividing the diameter by s. Therefore, the corresponding spread parameter σ of the point spread function will also be a normalized quantity. Taking this into consideration, the spread parameters σ_1 and σ_2 corresponding to g_1 and g_2 respectively are

$$\sigma_1 \;=\; k_1 D_1 \;\left(\frac{1}{f_1} - \frac{1}{u} - \frac{1}{s_1}\right) \tag{24}$$

and

$$\sigma_2 = k_2 D_2 \left(\frac{1}{f_2} - \frac{1}{u} - \frac{1}{s_2} \right). \qquad (25)$$

Eliminating u from the above two equations we get

$$\sigma_1 = \alpha \sigma_2 + \beta \qquad (26)$$

where

$$\alpha = \frac{k_1 D_1}{k_2 D_2} \quad \text{and} \quad \beta = k_1 D_1 \left(\frac{1}{f_1} - \frac{1}{f_2} + \frac{1}{s_2} - \frac{1}{s_1} \right). \qquad (27)$$

Equation (26) gives a relation between σ_1 and σ_2 in terms of the known camera parameters. This equation plays a central role in our method for depth recovery. To our knowledge, this relation has not been derived before in the literature.

Let $G_1(\omega, \nu)$ and $G_2(\omega, \nu)$ be the Fourier transforms of $g_1(x, y)$ and $g_2(x, y)$ respectively. We will denote the point spread functions corresponding to these two images by $h(x, y; \sigma_1)$ and $h(x, y; \sigma_2)$ and their Fourier transforms by $H(\omega, \nu; \sigma_1)$, $H(\omega, \nu; \sigma_2)$, where σ_1, σ_2 are the respective spread parameters. (Here the function h itself does not change with change in camera parameters.) With this notation, we can write

$$G_1(\omega, \nu) = H(\omega, \nu; \sigma_1) F(\omega, \nu) \qquad (28)$$

$$G_2(\omega, \nu) = H(\omega, \nu; \sigma_2) F(\omega, \nu). \qquad (29)$$

Dividing G_1 by G_2,

$$\frac{G_1(\omega, \nu)}{G_2(\omega, \nu)} = \frac{H(\omega, \nu; \sigma_1)}{H(\omega, \nu; \sigma_2)}. \qquad (30)$$

This is the second equation of central importance along with equation (26). While equation (26) gives a relation between σ_1 and σ_2 *in terms of the camera parameters*, this equation gives a relation *in terms of the observed images*. Equations (26) and (30) together constitute two equations in the two unknowns: σ_1 and σ_2. They are solved simultaneously to obtain σ_1 and σ_2. The depth u is then determined from either σ_1 or σ_2 using equation (11).

We now illustrate the above method for the case of a Gaussian point spread function. The optical transfer function for this case is as in equation (20). Therefore we get

$$\frac{G_1(\omega,\nu)}{G_2(\omega,\nu)} \;=\; e^{-\frac{1}{2}(\omega^2+\nu^2)(\sigma_1^2-\sigma_2^2)} \tag{31}$$

Taking logarithm on either side and rearranging terms, we get

$$\sigma_1^2-\sigma_2^2 \;=\; \frac{-2}{\omega^2+\nu^2}\;\ln\left(\frac{G_1(\omega,\nu)}{G_2(\omega,\nu)}\right)\;. \tag{32}$$

For some (ω,ν), the right hand side of equation (32) can be computed from the given image pair. Therefore equation (32) can be used to estimate $\sigma_1^2-\sigma_2^2$ from the observed images. As in the previous section, measuring the Fourier transform at a single point (ω,ν) is, in principle, sufficient to obtain the value of $\sigma_1^2-\sigma_2^2$, but a more robust estimate can be obtained by taking the average over some domain in the frequency space. Let the estimated average be C given by

$$C \;=\; \frac{1}{A}\int\int_R \frac{-2}{\omega^2+\nu^2}\;\ln\left(\frac{G_1(\omega,\nu)}{G_2(\omega,\nu)}\right)\;d\omega\;d\nu \tag{33}$$

where R is a region in the (ω,ν) space not containing points where $G_1(\omega,\nu)=G_2(\omega,\nu)$, and A is the area of R. Therefore, from the observed images we get the following constraint between σ_1 and σ_2:

$$\sigma_1^2-\sigma_2^2 \;=\; C\;. \tag{34}$$

Equations (26,34) together constitute two equations in two unknowns. From these equations we get

$$(\alpha^2-1)\;\sigma_2^2\;+\;2\alpha\beta\;\sigma_2\;+\;\beta^2\;=\;C. \tag{35}$$

Above we have a quadratic equation in σ_2 which is easily solved. In general there will be two solutions. However a unique solution is obtained if $D_1=D_2$. We can also derive other special cases where a unique solution is obtained (e.g.: $D_1\neq D_2$, $s_1=s_2=f_1=f_2$; in this case only the negative solution of σ is acceptable which is unique.)

Having solved for σ_2 we obtain the distance u from equation (11). Thus, the distance is determined from only two images obtained with different camera parameter settings. This should be compared to the depth-from-focusing methods [6, 10, 16, 9] which require recording and processing a large number of images. Note that *the camera parameter setting could differ in any one, any two, or all three of the parameters:* s, f, D.

This method of depth recovery requires a camera system for which one or more of the parameters s, f, D can be *changed and measured accurately.* Such a camera system is not available to us at present. We plan to acquire such a camera in the future to conduct experiments.

5 Scene containing Curved Objects

In the presence of curved objects, the distance u of visible surfaces is different along different directions of view. Let $u(x, y)$ be the distance along a line joining the optical center and the point (x, y) on the image detector. Depending on $u(x, y)$, the spread parameter σ of the point spread function also changes with position on the image detector. Let this position dependence be denoted by $\sigma(x, y)$. In this case the relation between the focused image and the observed image cannot be expressed as a convolution operation. The transformation from focused image to blurred image is still linear but *not shift-invariant.* The problem now is to recover $u(x, y)$ or the *depth-map* of the scene.

One solution to depth-map recovery is to divide an image into many smaller subimages and consider $u(x, y)$ to be approximately constant within each subimage. Then the depth corresponding to each subimage is obtained using the method in the previous section. If $u(x, y)$ is not constant within a subimage, then this scheme gives an "average" distance. This is still a useful piece of information in many applications.

Dividing an image into subimages introduces some errors due to border effects. An image region cannot be analyzed in isolation because, due to blurring (caused by the finite spread of the point-spread-function), the intensity at the border of the region is affected by the intensity immediately outside the region. We call this the *image overlap problem* because the intensity distribution produced by adjacent patches of visible surfaces in the scene overlap on the image detector plane. In indoor scenes such as the environments of industrial vision systems, the image overlap problem can be completely avoided through selective illumination of the scene. For example, the scene can be illuminated by square bright patches separated by wide dark bands with no illumination. In this case the boundaries of the subimages can be chosen to be in the middle of the dark bands. Border effects are then avoided because the image intensity is zero at and near the borders.

In situations where the illumination cannot be controlled (e.g. outdoor scenes), the image overlap problem may be reduced as follows. The image intensity is first multiplied by a suitable center weighted (e.g. a Gaussian) mask centered at the region of interest. The resulting weighted image is then used for depth recovery. Because the weights are higher at the center than at the periphery, this scheme gives a depth estimate which is approximately the depth along the center of the field of view.

6 Error Sensitivity

The effective range of our approach depends on many factors such as the values of the camera parameters, illumination condition of the scene, aberrations of the optical system, image quality (i.e. noise, spatial and grey level resolution), etc. Most existing camera systems have small apertures. This appears to be a deliberate design decision to maximize the depth-of-field. For such cameras, objects at all distances are nearly focused and consequently have low depth discrimination. Therefore a camera system designed specifically for depth recovery (having small depth-of-field) should perform significantly better than the commercially available cameras.

The depth recovery approach here requires only one or two images; however the estimate of depth can be made more robust if more images are used. If n images are available for different camera parameter settings, then $n-1$ independent estimates of depth can be made and the mean of these gives a robust estimate of the actual depth. Alternative schemes are also possible for using multiple images.

A general and complete analysis remains to be done, but one calculation under simplified assumptions leads to the following conclusions: (i) the approach is more accurate for nearby objects than for far away objects, (ii) the effective range in practical applications is about one hundred times the focal length of the camera system, and (iii) for far away objects, even if the accuracy of the quantitative estimate of depth may be unsatisfactory, we can still obtain useful qualitative information such as, for example, "object A is nearer than object B", "there are no obstacles within distance X", etc.

In some applications such as autonomous vehicle navigation where approximate depth needs to be recovered in real-time, multiple cameras could be advantageous. Each camera would have different parameters such that it is "tuned" to recover depth more accurately in a particular range than out side of this range. Cameras with smaller focal length lenses help to recover accurately the depth variations at shorter distances and those with larger focal length lenses (e.g. telescopic cameras) help to recover accurately the depth variations at longer distances. Different cameras can be made to view the scene from the same vantage point using a beam splitting device.

7 Notes

Autofocusing: The method for depth recovery presented here can be used in the autofocusing of computer controlled video cameras. Focusing by this method will be much faster than the depth-from-focusing methods used at present.

Enhancing depth-of-field: A focused image may be obtained from two observed images which are blurred. First the form of the point spread function $h(x, y; \sigma)$ is determined through calibration. Then, as in the depth recovery method, the spread σ of the point spread function is estimated. The observed image is then simply deconvolved with the corresponding point spread function. The resulting image is the required focused image. Deconvolution is done by dividing the Fourier transform of the observed image with the Fourier transform of its point spread function and then taking the inverse Fourier transform. Although, in principle, deconvolution is simple, in practice (especially in the presence of noise) it poses many serious difficulties.

Domain of analysis: We have presented our approach based on a Fourier domain analysis of the images. It is possible to do a corresponding analysis in the spatial domain. We have chosen the Fourier domain for its simplicity. In particular it should be noted that in equation (30), the power spectral densities of G_1 and G_2 could be used instead of G_1 and G_2. This usually does not complicate solving for the spread parameters. It also avoids computations on complex numbers. Power spectral density may be preferable because it can be computed very fast by optical methods [4].

Lens formula: The lens formula in equation (1) is valid exactly only for an aberration-free lens and for points near the optical axis. For an actual camera, one may consider v to be that distance of the image detector from the lens for which the image of a point at distance u is "sharpest" (i.e. the spread of the point spread function is a minimum). Therefore, for a given camera, one can experimentally determine v as a function of u, f and the direction (or angular position) of a point in the scene. Having determined this function, one can use it in place of the lens formula and derive the corresponding equations. Even if there is no satisfactory parametric representation of this function, a table-look-up method can be used. Only the computational steps become clumsy.

Avoiding the correspondence problem: While changing s or f, *the front lens of the camera receiving light directly from the object (i.e. the object piece) must never be moved.* Instead, the image detector should be moved backwards or forwards to change s, and a lens other than the object piece should be moved to change the focal length. This avoids the *correspondence problem* encountered in structure-from-motion where the camera is moved along the optical axis. Correspondence is established through simple ray tracing principles well known in geometric optics (cf. [3]). In most commercially available cameras, the front lens is moved instead of the image detector or film. This introduces the correspondence problem. This fact appears not to have been noted in previous work on depth-from-focusing. When the camera consists of multiple lenses, the effective focal length, and the positions of principal points and planes need to be computed. Methods for doing this are well established (cf. [11]).

Relevance to human vision: In the human visual system, focusing occurs by changing the focal length of the lens. In our method for depth recovery, the two images may be obtained by changing only the focal length. This suggests that humans can, in principle, perceive the depth of all objects in the field of view even if the objects are not in focus. There is evidence in support of the fact that the human eye deliberately exhibits small fluctuations in the focal length of the lens to obtain two images. The following paragraph is quoted from Weale [25, page 18]:

> "... the state of accommodation of the un stimulated eye is not stationary, but exhibits micro fluctuations with an amplitude of approximately 0.1 D (diopter: a unit of lens power given by the reciprocal of focal length expressed in meters) and a temporal frequency of 0.5 cycles/second. He (Cambell, [2]) demonstrated convincingly that these were not a manifestation of instrumental noise, since they occurred synchronously in both eyes. It follows that their origin is central."

Our approach implies that such fluctuations could be used to perceive depth in the entire scene simultaneously.

Plain objects: Objects like machine parts, wall, door, road, etc. are often "plain" or "textureless", i.e. their surfaces are smooth and have no reflectance variation. Therefore they appear as objects with constant brightness under uniform illumination. Our method fails for such objects due to the lack of spatial frequency content. However, if one has control over the illumination of the scene (as in indoor scenes), one can *introduce* "texture" by projecting an arbitrary light pattern (e.g. a random dot pattern) onto the surface of objects. Then our method becomes applicable. Note that the projected pattern need not be focused.

Magnification correction: The normalization of spatial scaling given by equation (23) appears to have been overlooked in the implementation of all depth-from-focusing methods [10, 16, 9] known to the author with the exception of Horn's [6]. We believe that applying this magnification correction will improve the reported experimental results and also alleviate some of the problems associated with local extrema and region correspondence.

8 Conclusions

We have presented some basic principles relevant to inverting the image formation process in a convex lens camera. This approach, termed inverse optics, suggests efficient methods for recovering scene information. Methods for recovering the distance of point, line, edge, and planar objects are presented. The difficulties associated with curved objects are discussed. Preliminary experiments suggest that the mathematical model developed here is applicable to actual camera systems and that useful depth information can be obtained through this approach.

The effective range of the current approach is limited by image quality and camera parameters. There is much scope for improvement in image quality (in terms of noise, spatial resolution, and grey-level resolution) in the future. (The advent of superconductors may advance the technology in this respect.) But, even with only the currently available technology, major improvements can be achieved through design of new camera systems meant specifically for depth recovery.

The depth information obtained from this approach, even if approximate, could be very valuable to other shape recovery methods like stereo vision. The approximate information can help to drastically prune the search space to the problem of stereo correspondence. Eventually this might indeed be the primary application of this method. A scheme for incorporating this method into a binocular system and a method for motion recovery are described in [20].

Research on the approach presented here has begun only recently and more extensive theoretical and experimental investigations are needed. The primary advantage of this approach is the absence of any inherent problem that requires heuristic solutions, such as the correspondence problem in stereo.

Acknowledgement: I am thankful to Mr. G. Natarajan for helpful discussions.

Bibliography

[1] D. H. Ballard and C. M. Brown, *Computer Vision,* Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1982, Section 2.2.2.

[2] F. W. Cambell, *Correlation of accommodation between the two eyes,* Journal of the Optical Society of America, 50, p. 738, 1960.

[3] J. D. Gaskill, *Linear Systems, Fourier Transforms, and Optics,* John Wiley & Sons, New York, 1978.

[4] J. W. Goodman, *Introduction to Fourier Optics,* McGraw-Hill, Inc., 1968.

[5] P. Grossman, *Depth from focus,* Pattern Recognition Letters 5, pp. 63–69, Jan. 1987.

[6] B. K. P. Horn, *Focusing,* Artificial Intelligence Memo No. 160, MIT, 1968.

[7] B. K. P. Horn, *Robot Vision,* McGraw-Hill Book Company, 1986.

[8] R. A. Jarvis, *A perspective on range finding techniques for computer vision,* IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5, No. 2, pp. 122–139, March 1983.

[9] E. Krotkov, *Focusing,* MS-CIS-86-22, Grasp lab 63, Dept. of Computer and Information Science, University of Pennsylvania, 1986.

[10] G. Ligthart, and F. C. A. Groen, *A comparison of different autofocus algorithms,* Proceedings of the International Conference on Pattern Recognition, 1982.

[11] K. N. Ogle, *Optics,* Chapter VII, Section 2, Charles C Thomas Publisher, Springfield, Illinois, 1968.

[12] A. P. Pentland, *Depth of scene from depth of field,* Proceedings of DARPA Image Understanding Workshop, Palo Alto, 1982.

[13] A. P. Pentland, *A new sense for depth of field,* Proceedings of International Joint Conference on Artificial Intelligence, pp. 988–994, 1985.

[14] A. P. Pentland, *A new sense for depth of field,* IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-9, No. 4, pp. 523–531.

[15] A. Rosenfeld, and A. C. Kak, *Digital Picture Processing,* Vol. I . Academic Press, 1982.

[16] J. F. Schlag, A. C. Sanderson, C. P. Neuman, and F. C. Wimberly, *Implementation of automatic focusing algorithms for a computer vision system with camera control,* CMU-RI-TR-83-14, Robotics Institute, Carnegie-Mellon University, 1983.

[17] W. F. Schreiber, *Fundamentals of Electronic Imaging Systems,* Springer-Verlag, Section 2.5.2., 1986.

[18] M. Subbarao, *Direct recovery of depth-map,* Tech. Report 87-02, Image Analysis and Graphics Laboratory, SUNY at Stony Brook, Feb. 1987. (Also appears in *Proceedings of IEEE computer society workshop on computer vision,* Miami Beach, Dec. 1987, pp. 58–65.)

[19] M. Subbarao, *Direct Recovery of Depth-map II: A New Robust Approach,* Technical report 87-03, Image Analysis and Graphics Laboratory, SUNY at Stony Brook, April 1987.

[20] M. Subbarao, *Progress in research on direct recovery of depth and motion,* Technical report 87-04, Image Analysis and Graphics Laboratory, SUNY at Stony Brook, May 1987.

[21] M. Subbarao, *Method and apparatus for determining the distances between surface patches of three-dimensional spatial scene and a camera system,* U.S. patent application no. 126407 (pending). Nov. 1987.

[22] M. Subbarao, *The optical transfer function of a diffraction-limited system for polychromatic illumination,* submitted to Applied Optics journal, June 1988.

[23] M. Subbarao, and G. Natarajan, *Depth recovery from blurred edges,* Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Ann Arbor, Michigan, pp. 498-503, June 1988.

[24] J. M. Tenenbaum, *Accommodation in Computer Vision,* Ph.D. Dissertation, Stanford University, Nov. 1970.

[25] R. A. Weale, *Focus on Vision,* Harvard University Press, Cambridge, Massachusetts, 1982.