Stereo in the Presence of Specular Reflection

Dinkar N. Bhat and Shree K. Nayar[1]

CUCS-030-94

bhat@cs.columbia.edu
nayar@cs.columbia.edu

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

December, 1994

[1]This work is supported in part by ARPA contract DACA-76-92-C-007 and in part by the David and Lucile Packard Fellowship.

Abstract

The problem of accurate depth estimation using stereo in the presence of specular reflection is addressed. Specular reflection is viewpoint dependent and can cause large intensity differences at corresponding points. Hence, mismatches can result causing significant depth errors. Current stereo algorithms largely ignore specular reflection which is a fundamental reflection phenomenon from surfaces, both smooth and rough. We analyzed the physics of specular reflection and the geometry of stereopsis which led us to an interesting relationship between stereo vergence, surface roughness, and the likelihood of a correct match. Given the lower bound on surface roughness, an optimal binocular stereo configuration can be determined which maximizes precision in depth estimation despite specular reflection. However, surface roughness is difficult to estimate in unstructured environments. Therefore, multiple view configurations independent of surface roughness are determined such that at each scene point visible to all sensors, at least one stereo pair provides a correct depth estimate. We have developed a simple algorithm to reconstruct depth from the multiple view images. Experiments with real surfaces confirm the viability of our approach. A key feature of this approach is that we do not seek to eliminate or avoid specular reflection, but rather minimize its effect on stereo matching.

Categories: Stereo, Physics-based vision

1 Introduction

Stereo is a direct method of obtaining three-dimensional information of the visual world which makes it attractive for a variety of applications like autonomous navigation, cartography and aerial surveying. These applications are commonly used in unstructured and unknown environments. For example, when an autonomous vehicle navigates, it encounters various objects with different shapes and reflectance properties. A robust stereo system must be able to obtain accurate depth estimates in any such environment containing objects with varying surface characteristics.

Depth recovery using stereo [Barnard and Fischler-1982] involves capturing images from different viewpoints, matching corresponding points in the images (known as the correspondence problem), and obtaining scene depth at corresponding points using triangulation. Inherently, the correspondence problem is under-constrained. While stereo geometry can constrain the range of possible regions of correspondence in the two images, it is not sufficient to make the problem tractable. Therefore, constraints have to be imposed by making assumptions regarding the scene. One major assumption made is that intensities of corresponding points in the images are identical. However, this is true only when the associated scene is Lambertian. Corresponding point intensities are not identical in the presence of *specular reflection*. The reason is that specular intensity at any scene point is dependent on the viewing direction. This effect is more clearly manifest on smoother surfaces where *highlights* – bright regions due to specular reflection – shift on the surface with a even with slight changes in viewpoint. Figure 1 shows a stereo pair of a rendered cup, and depth obtained along two scanlines; one including a highlight and the other away from it. Depth was computed by assuming that the surface of the cup is Lambertian. It can be seen that depth computed is erroneous at points where corresponding intensities are vastly different. Our paper deals with accurate depth estimation in the presence of specular reflection which is a fundamental form of reflection from surfaces, both smooth and rough.

Many algorithms have been developed to establish point correspondence. Since unambiguous intensity matching of individual points is impossible, search-based strategies have been used that match image regions (*area-based*), [Panton-1978], [Okutomi and Kanade-1993], [Hannah-1989], or primitives like edges and lines (*feature-based*). Area-based schemes rely on optimal statistical correlation between corresponding regions. These methods can compute wrong matches in the presence of specular reflection because corresponding regions could be poorly correlated in terms of intensity. Similarly in feature-based methods, highlights could be mistaken for real features and matched. This would result in incorrect depth computation because the motion of highlights does not correspond to that of any scene point or feature. Other methods have been developed for image matching which avoid direct intensity-based correspondence. Recently, Wolff and Angelopoulou[Wolff and Angelopoulou-1994] developed a stereo system which matches photometric ratios between images. However, this scheme requires two illumination conditions which makes it unattractive for passive stereo. Furthermore, it does not extend to

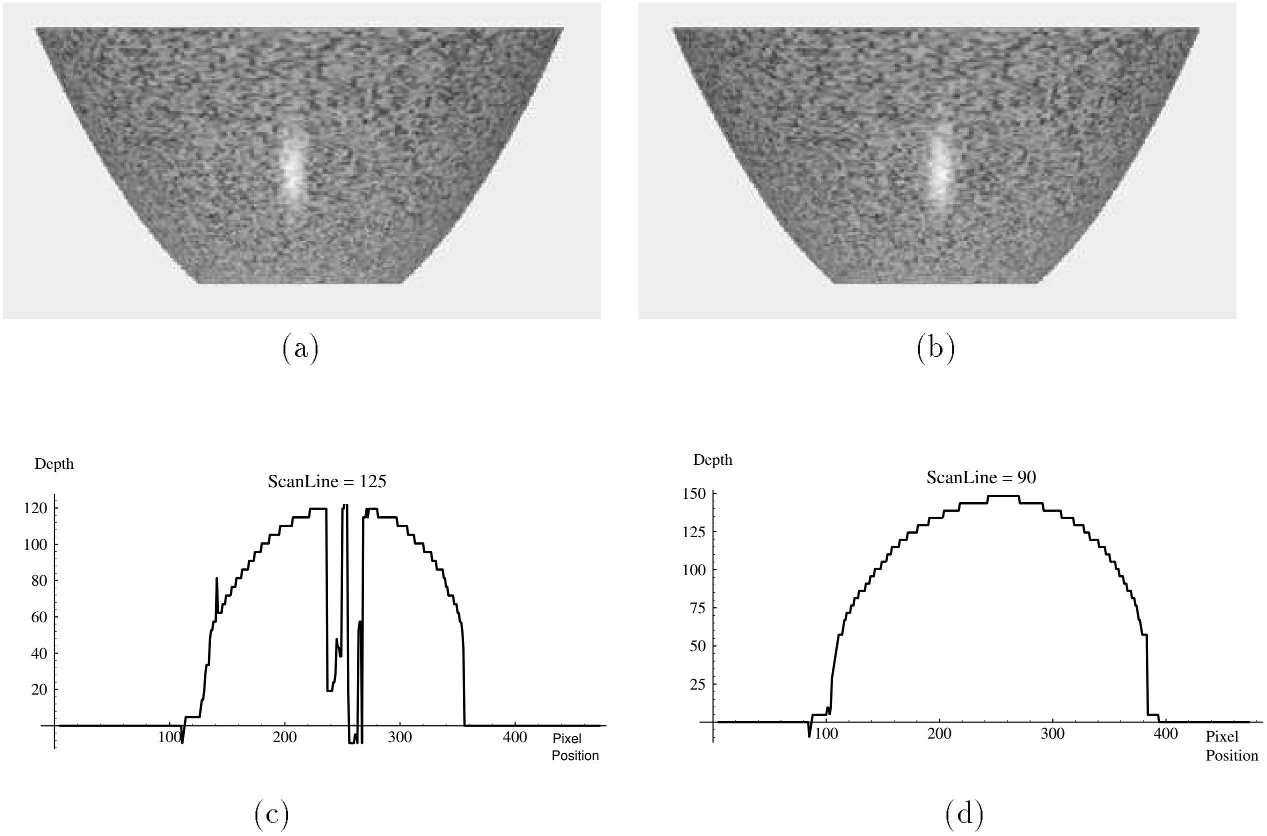


Figure 1: Rendered stereo pair and depth computed along two scanlines. (a) Left image; (b) Right image; (c) Depth along a scanline including the highlight; (d) Depth along a scanline away from the highlight. The depth was computed by assuming that the surface of the object is Lambertian. Notice that computed depth has large errors within the highlight region in (c).

specular surfaces because the ratio loses its property of being viewer independent. Smith [Smith-1986] proposed an elegant alternative to point correspondence. His approach attempts to find surface depth profiles that are consistent with image intensity profiles. Unfortunately, this does not work with complex intensity profiles like those observed in the case of specular reflection, as the solution involves ill-conditioned numerical systems.

Two view matching, both feature-based and area-based, is ambiguous. Multiple view systems have been designed to alleviate this problem. The redundant views provide strong constraints since correspondence must be established with respect to all images. In particular, trinocular stereo systems - systems using three views - have been designed to a measure of success [Pietikainen and Harwood-1986], [Ito and Ishii-1986], [Yachida *et al.*-1986]. The general strategy employed in trinocular stereo for correspondence is to locate possible matches using one stereo pair and verify each potential match using the third view. Only one matching triplet of points (or features) is expected to be consistent with respect to the three views. However, due to specular reflection, none of the matches in the first pair may correspond to the true match. Furthermore, even if the right match is located using the first stereo pair, it may not be possible to verify it in the third view as the corresponding point intensity can be greatly different. Dhond and Aggarwal [Dhond and Aggarwal-1991] analyzed the cost and benefit of adding an additional view for matching and concluded that the increased reliability of three view systems outweighs the added computational cost involved. However, they assumed diffuse surfaces and hence mismatches due to specular reflection were not considered in estimating reliability. A more complete analysis should deal with specular reflection in both binocular and trinocular stereo systems. Okutomi and Kanade [Okutomi and Kanade-1993] developed a convincing and practical multiple view system. It uses stereo pairs taken with different baselines to compute precise depth estimates without suffering from ambiguities caused by factors like repetitive texture. The assumption is that the right match at any scene point is always found in each stereo pair along with other possible other incorrect matches. By combining image pairs, only the correct match is accentuated while the incorrect ones weaken. The above assumption is not correct, however, when specular reflection is present because the right match may not be found in one or many image pairs. Thus, all the above techniques must incorporate ways to detect and handle mismatches due to specular reflection.

To overcome the problem of depth errors due to strong highlights, Brelstaff and Blake [Brelstaff and Blake-1988] suggested removing them from images before matching. Removal of highlights is difficult for real scenes in unstructured environments and is an active area of research [Nayar *et al.*-1993]. Ching *et al* [Ching *et al.*-1993] developed a correlation based technique to detect and avoid specular reflection when the camera is active. But the method is heuristic and uses an empirical adaptive window algorithm. On a different note, information from highlights has been used to recover local surface shape. Blake [Blake-1985] related the movement of a highlight to the Hessian of the surface which describes local surface geometry. The above techniques assume ideal specular reflection which as we will show is only an extreme case as surface roughness tends to zero.

It is clear that current stereo algorithms are seriously deficient in dealing with specular reflection. In this paper, we address the problem of precise depth estimation in the presence of specular reflection. Roughness of surfaces is considered since it strongly influences the degree of specular reflection. First, we seek an optimal *binocular stereo* configuration such that intensity differences at corresponding points is limited, while depth resolution is maximized. The optimal configuration is determined independent of surface normal and source direction as they are measurable only in some structured environments. The configuration parameters are shown to be a function of surface roughness. Surface roughness appears as an independent parameter because it cannot be constrained without prior knowledge of the scene. Therefore, given a scene where the lower bound on surface roughness can be estimated, possible in indoor structured environments, the two cameras can be positioned to minimize mismatches without losing precision in depth computation.

Next, we seek to avoid estimation of surface roughness. The reason is that measurement of surface roughness in outdoor scenes is not practical. To overcome this problem, we derive *multiple view* configurations based on the binocular stereo framework. The important characteristic of these configurations is that for each scene point in the common field of view of the sensors, *at least* one binocular pair provides the correct depth estimate. We have developed a novel correspondence algorithm (currently implemented for a trinocular system) to extract correct depth estimates of scene points from different binocular pairs so as to yield an accurate and dense depth map of the scene. The algorithm is simple, can be easily implemented and is easily parallelizable. We have shown its effectiveness on different scenes.

Our approach is premier in considering specular reflection from rough surfaces in the context of stereo. All previous methods have implicitly [Ching *et al.*-1993] or explicitly [Blake-1985] assumed ideal specular reflection. We do not attempt to avoid or detect the immediate artifacts of specular reflection like strong highlights but rather perform accurate matching in their presence. Thus preprocessing of images, like removal of highlights, is avoided. Our approach is not limited by any specific reflectance model or to any image correspondence scheme. Therefore, it is general and can be incorporated into existing stereo algorithms. Our new multiple view correspondence algorithm uses computationally simple window-based operators which makes it attractive for real-time stereo. Experiments with real surfaces demonstrate the viability of our approach.

The paper is organized as follows. First, we describe basic reflection mechanisms and the effect of specular reflection on stereo. The binocular stereo framework is developed and experiments are presented to illustrate its applicability. Next, the multiple view approach is elucidated. We describe the configuration derived and present experiments on real objects. We conclude with a brief discussion on the proposed framework.

2 Surface Reflection Mechanisms

The aim of this section is to qualitatively describe the effect of specular reflection on stereo. Surfaces exhibit two fundamental forms of reflectance - *body* and *surface*. The popular terminology for these components are *diffuse* and *specular reflectance*, respectively.

2.1 Body Reflection

Body reflection occurs due to internal scattering of light. Incident rays enter the surface, undergo multiple reflections and refractions due to surface inhomogeneities, and a fraction of them emerge out of the surface. Because of its non-directional nature, this form of reflection is also called diffuse reflection. We prefer the term diffuse reflection, because it is more widely used in computer vision literature.

Diffuse reflection is often assumed to be Lambertian. A Lambertian surface appears equally bright from all viewing directions. However, Oren and Nayar [Oren and Nayar-1994] showed that this assumption is incorrect for surfaces with macroscopic roughness. They developed a comprehensive reflectance model by observing that a surface modelled as a collection of Lambertian facets is not Lambertian when viewed at low magnification, i.e when a pixel includes many such facets. In fact, the Lambertian case was shown to be a limiting instance of their model. However, even for rough surfaces, the change in diffuse component with viewing direction is much less pronounced than the change in specular component.

2.2 Surface reflection

Surface reflection occurs at the boundary between materials. It can be broken down to two components - coherent and incoherent[2]. Coherent reflection obeys the mirror law of reflection, and accordingly incident light is reflected only in the specular direction. This component is also termed as the specular spike [Nayar *et al.*-1991]. When surface variations are very small in comparison to the wavelength of light, coherent reflection dominates. As the surface gets rougher, the coherent component weakens rapidly and the incoherent component begins to dominate. The incoherent component spreads in directions other than and including the specular direction, the distribution depending on the roughness of the surface. This component is also termed as the specular lobe [Nayar *et al.*-1991]. We prefer the term specular reflection over surface reflection, because the underlying reflection mechanism of the surface reflection components is mirror-like or *specular*. In this paper, we do not deal with surfaces smooth in comparison to the wavelength of the incident light as they rarely occur in real scenes. Therefore, specular reflection refers to the specular lobe only. It is described by the Torrance-Sparrow model [Torrance and Sparrow-1967] which is based on geometrical optics [Nayar *et al.*-1991]. The following is a brief description of this model.

A surface is viewed as a collection of planar microfacets, each behaving like a perfect mirror. A rough surface can be modelled using a probability distribution for the slopes of the microfacets. The slope distribution model uses a parameter σ which represents surface roughness. A smoother surface is characterized by a lower value for σ. Using this surface model, the specular intensity I_s at any point was shown in [Torrance and Sparrow-1967] to be:

$$\begin{aligned} I_s &= \frac{K_s F\, G}{\hat{n}\cdot\hat{v}} exp(-\frac{1}{2\sigma^2}(cos^{-1}(\hat{h}\cdot\hat{n}))^2) \\ G &= min(1, \frac{2(\hat{n}\cdot\hat{h})(\hat{n}\cdot\hat{v})}{\hat{v}\cdot\hat{h}}, \frac{2(\hat{n}\cdot\hat{h})(\hat{n}\cdot\hat{s})}{\hat{v}\cdot\hat{h}}) \\ \hat{h} &= \frac{\hat{v}+\hat{s}}{||\hat{v}+\hat{s}||} \end{aligned} \tag{1}$$

where $\hat{v}$, $\hat{s}$ and $\hat{n}$ are unit vectors pointing along the viewing, source and normal directions, respectively. In the above equation, G is the Geometrical Attenuation Factor which

[2]Coherence refers to the predictability of phase of reflected light wave with respect to the incident light wave. If the phase of the reflected wave is well-defined, then it is called *coherent* and if it is random it is called *incoherent*.

accounts for masking and shadowing effects, F is the Fresnel's coefficient and $\hat{h}$ is the bisector of $\hat{v}$ and $\hat{s}$. K_s is a constant which accounts for the gain of the sensor measuring the intensity, the source strength, normalization for the Gaussian term in the specular intensity expression, and the reflectivity of the surface.

The plots in Figure 2 illustrate the variation in specular intensity with changing viewing and source vectors, and surface roughness. In all plots, the source, normal and viewing vectors are constrained to lie in one plane. Any of the vectors can change in the plane, but I_s depends only on two independent variables: θ_s is the angle between source and normal vectors, and θ_r is the angle between viewing and normal vectors. From the plots it can be seen that: (a) When the surface is smooth, the distribution of I_s is concentrated around a small region around the specular direction; (b) As the surface becomes rougher, the peak value of I_s decreases and the distribution of I_s widens. We will discuss the implications of these observations to stereo.

2.3 Implications for Stereo

The image intensity I_t for any point in the scene is given by:

$$I_t = I_d + I_s \tag{2}$$

where I_d and I_s are the diffuse and specular intensities components, respectively. Given a pair of stereo images, the intensities at corresponding points are different in general because each component varies due to the change in viewing direction[3]. The variation in I_d can be described using the model in [Oren and Nayar-1994] while I_s varies according to (1). Thus, I_t^1 and I_t^2, the total intensities in the two images are given by:

$$I_t^1 = I_d^1 + I_s^1,\ I_t^2 = I_d^2 + I_s^2 \tag{3}$$

Since the change in I_d is much smaller than the change in I_s, it follows from (3) that the overall intensity difference I_{diff} can be assumed to be equal to the difference in specular intensities:

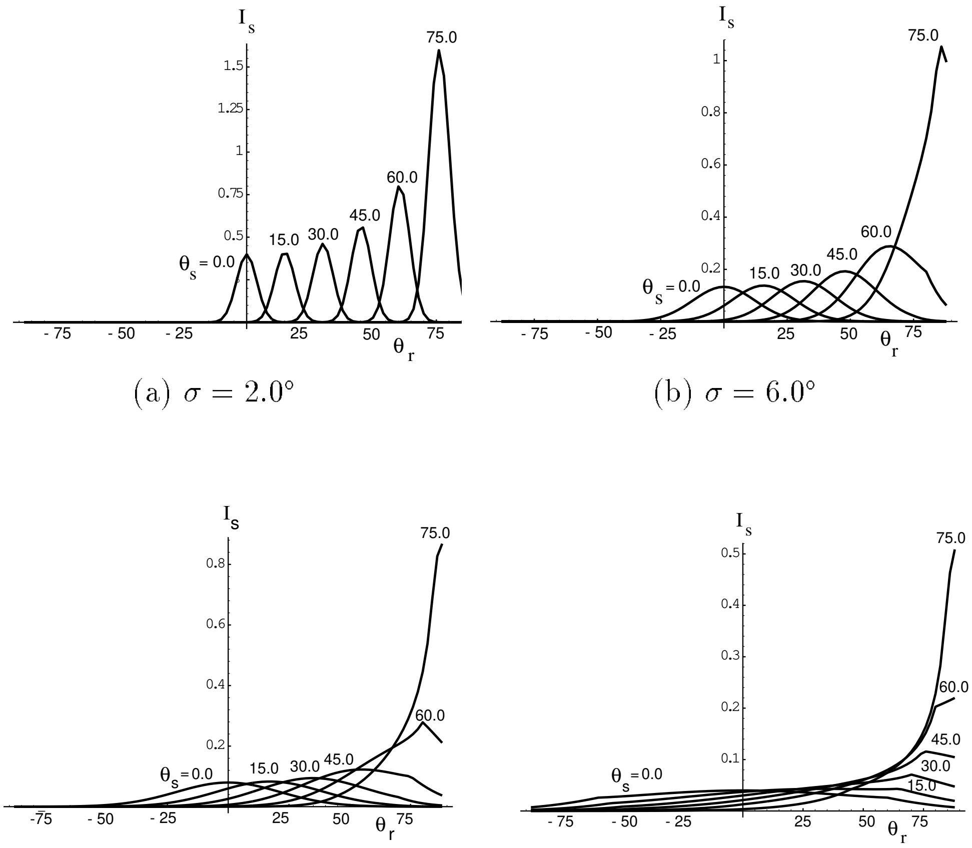
$$I_{diff} = |I_s^1 - I_s^2| \tag{4}$$

I_{diff} varies over the scene as the surface normal and roughness are generally not constant. Further, I_{diff} could be large if the viewing directions are chosen arbitrarily, resulting in wrong matches while computing stereo correspondence.

For any given surface roughness σ, I_{diff} is small if the viewing directions are close[4]. This naturally leads to the question as to how far apart can the viewing vectors be placed beyond which I_{diff} becomes larger than a specified quantity. In other words, what is the limit for the angle between the viewing vectors at which I_{diff} exceeds a threshold? This

[3]We assume the scene is illuminated by a light source whose direction is fixed but unknown. We also assume that the gain of the cameras are identical while obtaining the images. The response of the sensors is assumed to be linear with respect to radiance.

[4]The trivial case is when the viewing vectors coincide and both stereo images are identical.



(a) $\sigma = 2.0°$ (b) $\sigma = 6.0°$

(c) $\sigma = 10.0°$ (d) $\sigma = 20.0°$

Figure 2: Specular intensity plotted against viewing angle θ_r for different roughness values. For each roughness value, intensity is plotted for incidence angles θ_s. All angles are measured in degrees.

upper limit determines the range of acceptable viewing directions. In the case of smoother surfaces, a change in viewing direction can cause a comparatively large change in I_s, i.e the peak value of I_{diff} is larger for a smoother surface. Consequently, the upper limit is bound to be smaller for the smoother surface. It should be ascertained independent of surface normal and source direction since they are unknown and indeterminable except in some structured environments.

3 Vergence Definitions

In this section, we discuss how the specular intensity difference at scene points can be affected, by changing parameters related to stereo geometry. Parameters of interest include camera vergence and baseline (Figure 4). Camera vergence [Diner and Fender-1993] is the angle β between the intersecting optical axes (O_1V and O_2V) of the cameras, and baseline is the distance between the two viewpoints (O_1 and O_2).

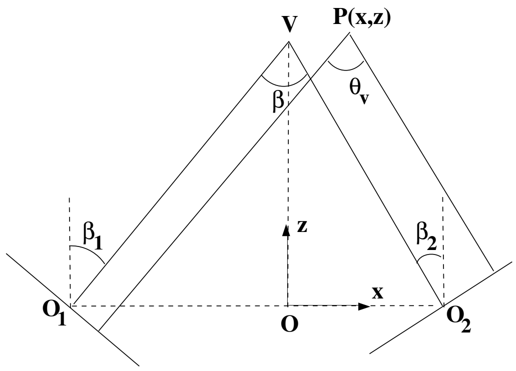


Figure 3: Point vergence and camera vergence under orthographic projection.

When points are projected *orthographically* to the two stereo image planes, as shown in Figure 3, corresponding rays are parallel to their respective optical axes. The implications are that: (a) The viewing vectors at all points are *equal* with respect to either viewpoint; (b) The viewing directions at all points can be simultaneously varied, by changing camera vergence alone. The camera vergence in turn can be controlled[5] by varying the individual camera angles (the angle between the z-axis and camera optical axis). In effect, we are changing the angle between the projected rays from any point in the scene θ_v termed as *point vergence.* Point vergence interests us because it is a controllable parameter, independent of surface normal, and affects the specular intensity difference at scene points. The relation between point vergence and camera vergence for orthographic projection is:

$$\begin{aligned} \theta_v &= \beta \\ \beta &= \beta_1 + \beta_2 \end{aligned} \tag{5}$$

where β_1 and β_2 are the individual camera angles.

In the case of *perspective* projection (Figure 4), viewing direction varies at each point in the scene, with respect to either viewpoint, i.e point vergence varies across the scene. In order to define a single controllable parameter to affect specular intensity differences, the point vergence is averaged over a *workspace* W[6]. We call this parameter as *field vergence* which is denoted by $\bar{\theta}_v$. In effect, we approximate every point vergence

[5]It is assumed that the cameras are yoked about the x and z axes.

[6]Since we are dealing with a single epipolar plane, the workspace is an area. In the general case, it would be a volume in three dimensional world.

in the workspace W by the field vergence. If dA represents an infinitesimal area in W, then $\bar{\theta_v}$ is given by:

$$\bar{\theta_v} = \frac{\int_W \theta_v dA}{\int_W dA} \tag{6}$$

There are two ways of choosing the workspace. The first is to define the workspace

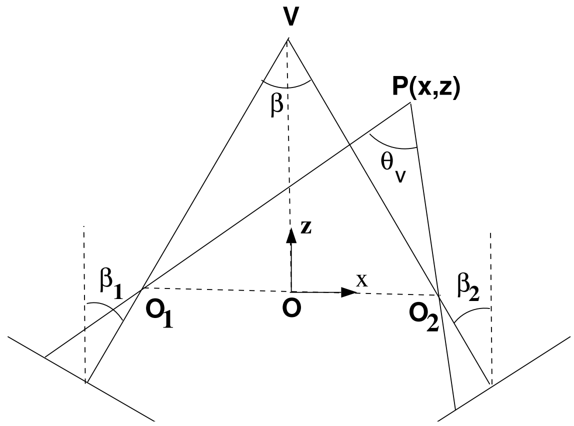


Figure 4: Variation of point vergence under perspective projection in a single epipolar plane ($\beta \neq \theta_v$).

in world coordinates. A rectangular workspace could be chosen, for simplicity. The second is to specify limits of the projected workspace in both images. For example, the projected workspace could be defined as $(-\delta, \delta)$ with respect to the image centers where δ is specified in pixels. The actual workspace can be obtained by backprojection from the images. The shape of the workspace hence obtained would be quadrilateral. In this paper, we adopt the first scheme because it is direct. Also, $\bar{\theta_v}$ could be related to a single camera parameter namely baseline (see Appendix B for details), whereas in the second it would be a function of baseline, camera vergence and the focal lengths of the cameras. Figure 5 shows the variation of $\bar{\theta_v}$ with baseline for a sample rectangular workspace defined by $[(x_{min} : 0, x_{max} : 2),(z_{min} : 2, z_{max} : 5)]$. It can be seen that $\bar{\theta_v}$ increases monotonically with baseline.

In this paper, we will refer to point vergence and field vergence jointly as *vergence*, unless explicitly stated. Vergence is an important design parameter because it is related to *depth resolution*. Depth accuracy and hence resolution, are limited by image quantization [7] amongst other factors. In Appendix A, we show that the depth resolution attainable at any point is directly proportional to vergence, assuming quantization is the primary

[7]Two types of quantization, spatial and intensity, occur in imaging. In this paper, quantization implies the former where one pixel in the image corresponds to more than one point in the world.

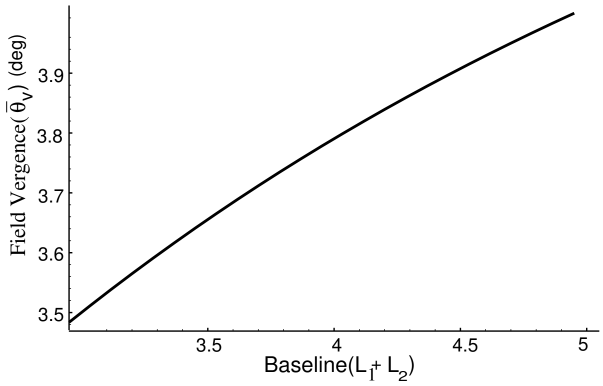


Figure 5: Variation of the field vergence with baseline for the given stereo workspace.

source of depth errors. To understand this phenomenon, consider a region in space (called a *lozenge*) [Diner and Fender-1993]) projected onto one pixel on each camera, as shown in Figure 6. Any point in this lozenge will have the same discrete image locations. The lozenge size which depends on vergence and pixel size effectively determines the depth resolution attainable. Thus, our objective is to attain maximum vergence while minimizing intensity differences at corresponding points.

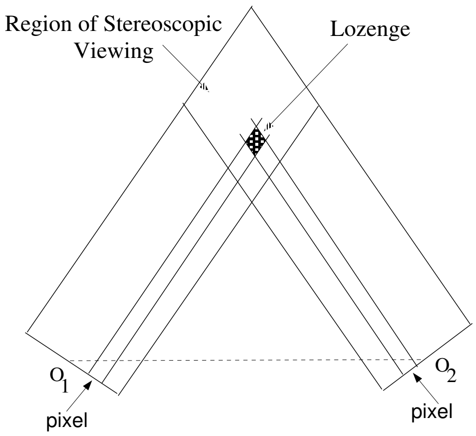


Figure 6: Region of stereoscopic viewing and a lozenge shaped area viewed by one pixel on each camera, under orthographic projection.

4 Problem Formulation – Binocular Stereo

Determining the maximum acceptable vergence in the presence of specular reflection can be formulated as a constrained optimization problem. In this section we describe the mathematical formulation of the problem, by defining the objective function and suitable scene related constraints. Every scene point on a surface can be mapped to a point O, the origin of our left-handed coordinate system (Figure 7). It's corresponding normal orientation is described by a unit vector $\hat{n}$ pointing away from O. If we assume that the point source is distant and image projection is orthographic, then the source and viewing vectors can also be located at O, and pointing away from it. The same representation holds in the case of perspective projection too since we approximated point vergences in a workspace by a single value, namely the field vergence (section 3).

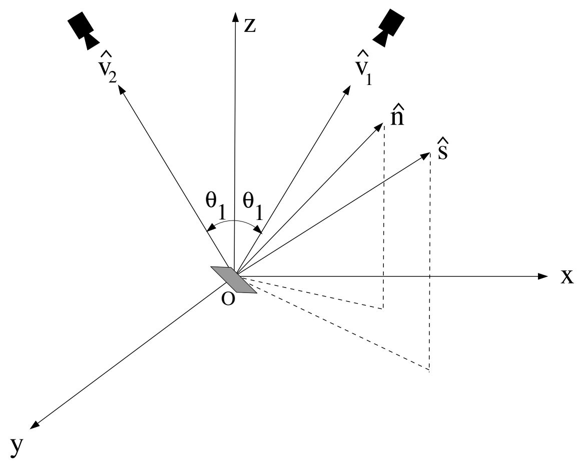


Figure 7: Coordinate system used for the stereo problem. Each scene point is mapped to the origin (O) of the coordinate system.

The aim is to attain maximum vergence. Hence a suitable objective function f_{obj} would be:

$$f_{obj} = \hat{v}_1 \cdot \hat{v}_2 \tag{7}$$

To limit the specular intensity difference I_{diff} at every point in the scene, the following constraint ($c1$) is imposed:

$$I_{diff} < T \tag{8}$$

where T is a threshold. Appendix C explains the significance of minimizing I_{diff} from a statistical perspective. The cameras are restricted to lie in the positive x-z plane, and tilt symmetrically about the z-axis. These constraints ($c2$) can be expressed as:

$$\hat{v}_1 \cdot \hat{j} \;\; = \;\; \hat{v}_2 \cdot \hat{j} = 0$$

$$\begin{aligned}\hat{v}_1 \cdot \hat{k} &= \hat{v}_2 \cdot \hat{k} \\ \hat{v}_i \cdot \hat{k} &> 0,\ i : 1,2\end{aligned} \tag{9}$$

where $\hat{i}$, $\hat{j}$ and $\hat{k}$ are unit vectors along the x,y and z axes, respectively.

To obtain a physically meaningful solution, only those scene points whose normal orientations are such that they can be illuminated and are visible to both the cameras should be considered. In other words, incidence and viewing angles greater than grazing value must be avoided. Therefore, the following constraints ($c3$) are imposed:

$$\begin{aligned}\hat{v}_1 \cdot \hat{n} &> 0 \\ \hat{v}_2 \cdot \hat{n} &> 0 \\ \hat{s} \cdot \hat{n} &> 0\end{aligned} \tag{10}$$

Finally, the unit vector constraints ($c4$) are added:

$$||\hat{v}_1|| = ||\hat{v}_2|| = ||\hat{s}|| = ||\hat{n}|| = 1 \tag{11}$$

The optimization problem can now be stated as:

$$\begin{aligned}Maximize:&\ f_{obj} \\ subject\ to\ constraints:&\ (c1, c2, c3, c4)\end{aligned} \tag{12}$$

The variables are $\hat{v}_1$, $\hat{v}_2$, $\hat{s}$ and $\hat{n}$[8]. To solve the problem, the associated Lagrangian and the Kuhn-Tucker's conditions for optimality [Kaplan-1959] must be defined. By eliminating $\hat{s}$ and $\hat{n}$ from the resulting equations, the optimal viewing directions $\hat{v}_1^{opt}$ and $\hat{v}_2^{opt}$ and hence the optimal vergence θ_v^{opt} can be obtained, either numerically or analytically.

In order to demonstrate a particular solution, the equation for specular intensity given by the Torrance-Sparrow model (1) is used in constraint ($c1$). Dividing both sides by K_s, the constraint can be written as:

$$I_{diff}/K_s < T/K_s \tag{13}$$

By writing the constraint in this form, it can be seen that T/K_s is an independent parameter. We call it the *relative threshold*. It is related to image correspondence which we discuss in the next section. Roughness σ is also unconstrained because surfaces in the scene are unknown. Thus, the *optimal vergence* $\theta_{\rm v}{}^{opt}$ *is a function of surface roughness* σ *and relative threshold* T/K_s.

Obtaining a closed-form solution for the optimal vergence in terms of roughness and relative threshold is not possible, given the highly nonlinear form of the constraints. Therefore, we use a numerical technique for constrained optimization called successive quadratic programming[Press *et al.*-1989]. Using this approach, the relationship between $\theta_v{}^{opt}$, σ and T/K_s is determined (see Figure 8). The implications of this relationship are:

[8] Only one of the viewing vectors is an independent variable. The other is its mirror reflection about the z-axis and hence can be eliminated.

- The optimal vergence increases with roughness. The reason is that I_{diff} weakens with increasing roughness allowing larger vergence. The surface progressively behaves in a diffuse manner, and thus the effects of specular reflection on matching diminish.

- The optimal vergence also increases with relative threshold. This is intuitive because a larger variation in I_s is permitted. Selecting the right relative threshold is a design issue. While too small a value will result in low depth resolution as discussed in section 3, a high value can cause mismatches. We will deal with this issue in subsequent sections.

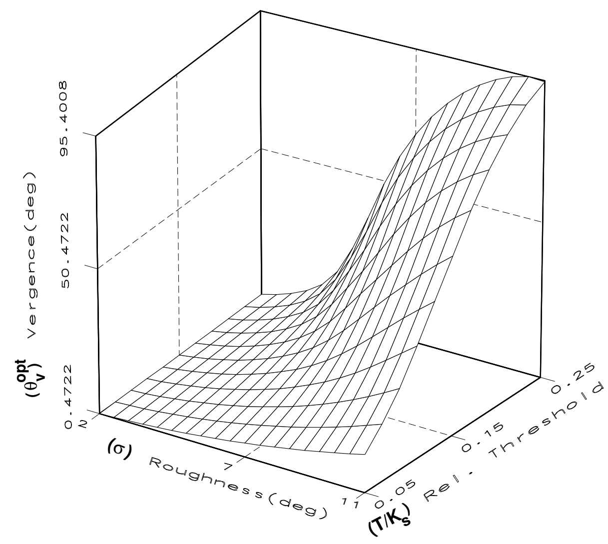


Figure 8: Graph illustrating the relationship between roughness, relative threshold and optimal vergence.

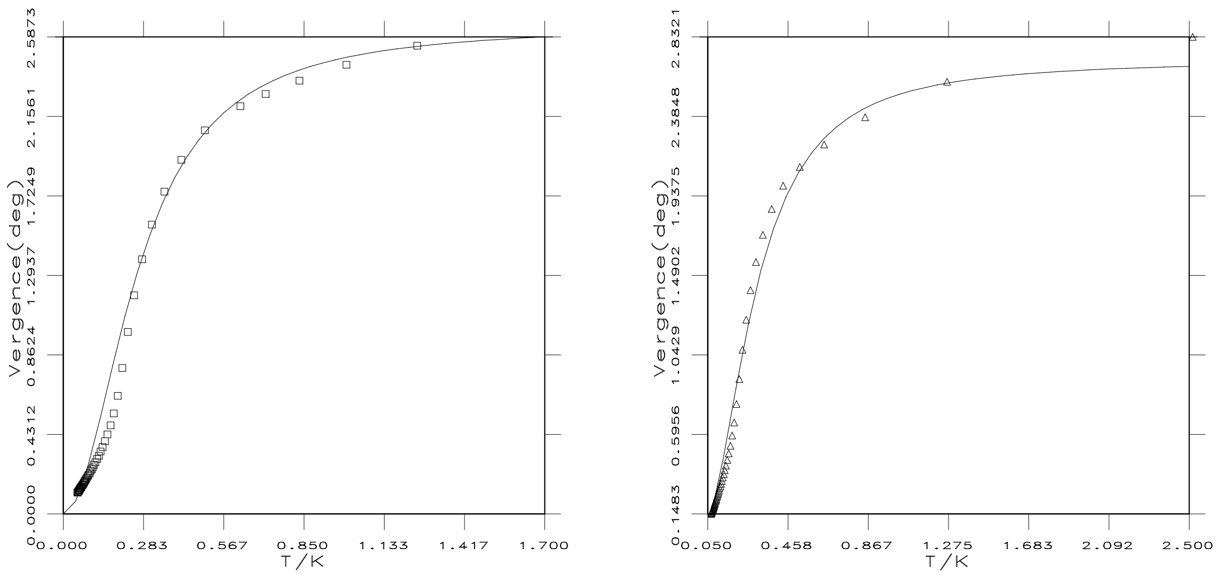
Variations to the general problem can now be considered by further constraining the variable vectors. For example, they could be constrained to lie in one plane, i.e all objects in the scene are polyhedral or cylindrical and oriented appropriately. However, the approach to determine the optimal stereo configuration remains unchanged; only additional constraints need to be imposed. It must be pointed out that very large vergences are not desirable, even though high depth resolution is attainable. The reason is that with increasing vergence the range over which depth can be determined decreases. In other words, the vertical extent of the region of stereoscopic viewing decreases (Figure 6).

5 Functional Approximation

Since we do not have a closed-form solution for optimal vergence in terms of relative threshold and roughness, we provide a reasonable functional approximation. If $\tilde{\theta_v}^{opt}$ approximates $\theta_v{}^{opt}$, then the corresponding function approximation is:

$$\tilde{\theta_v}^{opt} = \frac{a\,(T/K_s)^2\,\sigma^2}{(T/K_s)^2 + b\,\sigma^2} \tag{14}$$

where a and b are constants. This function adequately captures the nature of the graph shown in Figure 8. Figure 9 illustrates the fitting of vergence data to the approximating function (14) with two different values of σ. The constants are calculated by nonlinear fitting using the Levenberg-Marquardt algorithm [Press *et al.*-1989]. The vergence data was obtained by solving (12) numerically.



(a) $\sigma = 5.0°$; A=2.7, B=0.07 (b) $\sigma = 7.0°$; A=2.8, B=0.09

Figure 9: Graphs illustrating fitting of vergence data to the approximating function. In the two graphs, $A = a\sigma^2, B = b\sigma^2$. The solid lines indicate the approximating function and the dots indicate the vergence data values.

6 Image Correspondence

We introduced a new parameter called relative threshold T/K_s which limits the maximum deviation in intensity between corresponding points. Determining the exact value of the relative threshold when mismatches begin to occur (the breaking threshold) is difficult.

The exact value depends on the diffuse texture of the surface. It can be high if the texture variation is sufficient to find right matches even in the presence of specular reflection and *vice versa*. Unfortunately, textures of real surfaces are extremely diverse, making the estimation of the breaking threshold a hard problem. Note that the problem is inherent to stereo matching, and is not a limitation of our approach. In fact, it is only natural that the threshold appears in our formulation. The problem is mitigated when we use the multiple view approach.

To make the plot in Figure 8 usable, a correspondence operator is required which is sensitive to changes in the relative threshold and degrades gracefully. In other words when relative threshold is low, the operator must match corresponding points with high confidence, and *vice versa*. In this section, we discuss two area-based operators, the normalized correlation coefficient (NCC) and the sum of squared differences (SSD), and their sensitivity to relative threshold.

NCC (also called Karl Pearson's Coefficient) measures the degree of linearity between two random variables. It varies between 0 and 1; a value of 1 indicates perfect linearity, and 0 indicates lack of it[9]. It is invariant to scaling of the random variables. Given two matching regions (*windows*) M in the two images, containing N pixels and having intensities $I_1^{(i,j)}$ and $I_2^{(i,j)}$, $NCC = 1$ if $I_1^{(i,j)} = I_2^{(i,j)}, (i,j) \in M$, i.e if the corresponding surface is Lambertian[10]. However, due to specular reflection the intensities are not equal and hence NCC deviates from 1. The deviation is estimated by E:

$$E = \frac{1}{N} \sum_{(i,j)\in M} (\frac{I_1^{(i,j)} - I_2^{(i,j)}}{K_s})^2 \tag{15}$$

assuming that all points in each window are identically scaled in intensity by K_s. Since we limit the intensity difference at all corresponding points by T, it follows that $E < (T/K_s)^2$. Thus, it can be seen that NCC is sensitive to changes in relative threshold.

The other stereo operator of interest SSD measures the difference in intensities in windows using the Euclidean norm. Thus, it is proportional to the square of the absolute threshold, T. To make it sensitive to relative threshold, we modify it as follows:

$$SSD_m = \frac{\sum_{(i,j)\in M} (I_1^{(i,j)} - I_2^{(i,j)})^2}{max(I_1^{(i,j)}, I_2^{(i,j)})^2} \tag{16}$$

Note that the above is not strict normalization since we are using different norms in the numerator and denominator. SSD is a more popular operator than NCC because it is computationally less expensive, an important requirement for real time stereo systems. Invariance to intensity scaling has two physical implications. First, it implies that stereo

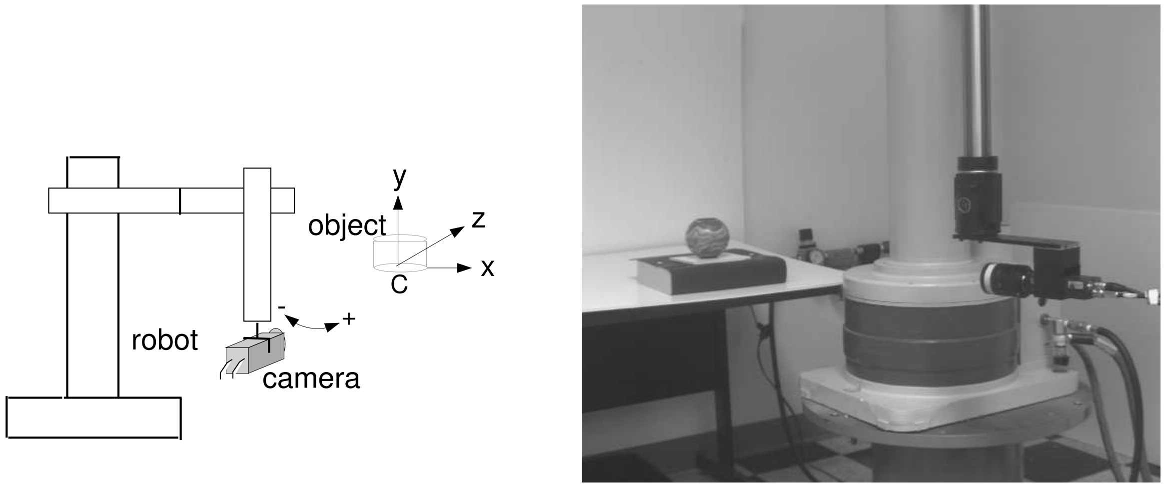
[9] $NCC = 0$ only means lack of linearity between the random variables. They may still be correlated nonlinearly.

[10] We ignore noise, and geometrical distortion in the windows, i.e windows do not necessarily correspond to the same surface area because of unequal projection.

images taken with different aperture settings can be used. Second, we do not have to determine K_s for points on a surface as we are not concerned with the absolute threshold T. Measuring K_s is in practice difficult because it depends on illumination, surface parameters and camera settings.

7 Experiments

In this section, we illustrate the effect of vergence on stereo matching using surfaces with different roughness. For these experiments, we use a 5 degree of freedom SCARA (Adept) robot (see Figure 10). The end-effector, equipped with a camera, is programmed to move in a circle. Objects are located at the center of the circle. This allows for different angles between the viewing vector and normals of points in the scene. By using a large radius of rotation in comparison to object dimensions, we approximate orthographic projection. The end-effector could also be equipped with a light source when different angles between the source and normal vectors is desired. In this case, using a large radius of rotation approximates a distant light source. We use this arrangement in the measurement of surface roughness.



(a) (b)

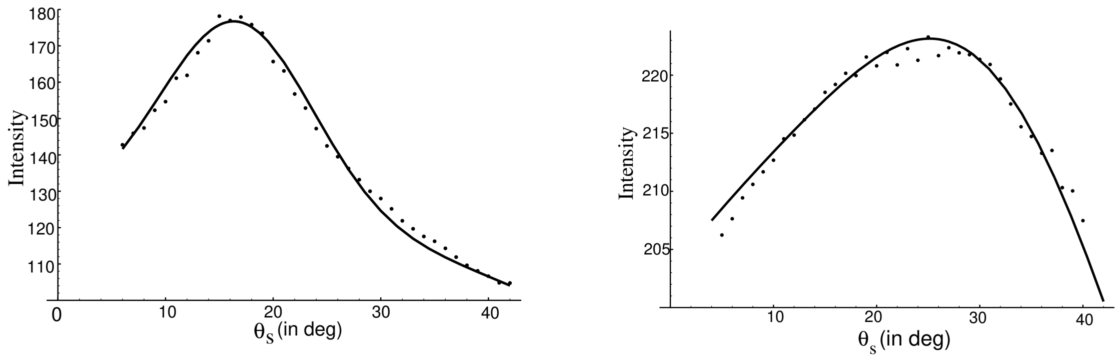
Figure 10: The experimental setup. (a) Diagram of a robot with a camera fixed to its end-effector. The coordinate system with origin C is also shown; (b) A photograph of the setting.

7.1 Roughness Measurement

We describe the method used to measure surface roughness. This is performed to subsequently illustrate the relationship between vergence and roughness, and is not required in general. We use two cylindrical objects wrapped with different surfaces (see Figure 11 and 12); a gift wrapper and a roughened xerox quality paper. The surfaces are uniformly rough, i.e all points on the surface have nearly the same σ. For each object, a set of intensity measurements are obtained at a point on the surface by varying the source direction and fixing the viewing direction. The viewing vector is along the positive z axis. The source, viewing and normal vectors are constrained to lie in a single plane (the x-z plane). This is possible because the object is cylindrical and hence all normal vectors corresponding to points on its surface are in one plane. Equation (17) expresses the total image intensity I_t (2) as the sum of a Lambertian and a specular component (derived from (1)):

$$I_t = K_d cos(\theta_s - \theta_r) + \frac{K_s}{cos\theta_r} \exp(-\frac{1}{2\sigma^2}(\frac{\theta_s - 2\theta_r}{2})^2) \tag{17}$$

where the unknown variables are θ_r, K_d, K_s and σ. K_d represents the diffuse coefficient and θ_r corresponds to the surface normal orientation. The Geometrical Attenuation Factor does not appear because θ_s and θ_r are much smaller than grazing value. The total intensity function is fitted to the measured intensity values. With non-linear fitting [Press *et al.*-1989], the respective roughness values obtained are: a) 3.5°, b) 6.3°. It can be seen that the gift wrapper surface is smoother. The other values are: a) $K_d = 125, \theta_r = 8.3°, K_s = 53$, b) $K_d = 210, \theta_r = 15.6°, K_s = 17$.



(a) (b)

Figure 11: Measurement of surface roughness for two surfaces: a gift wrapper and a roughened xerox quality sheet. (a-b) Plotting intensity measurements to the approximating function. The dots in the graphs represent intensity measurements and the solid line is the total intensity function. The roughness values obtained are 3.5° and 6.3°, respectively. Observe that the range of measured intensity values is smaller for the rougher surface.

7.2 Stereo Matching

The effect of vergence on matching is now demonstrated using the two surfaces. The source direction is fixed at approximately 30°, and the camera angle is varied to obtain a range of vergence values. In order to use approximately the same relative threshold, similar random patterns on the surfaces were marked. Images obtained at equal angles about the z-axis are matched using SSD_m along scanlines containing texture[11].

For each surface, depth obtained along a scanline at different vergence values is shown in Figure 12. It can be seen that for each surface large depth errors are computed at larger vergence: 8.0° and 11.0° respectively, although a higher vergence is acceptable for the rougher surface. The wider distribution of I_{diff} causes mismatches to be more distributed on the rougher surface. In the case of the smoother surface, mismatches are confined to the highlight region.

8 Motivation for a Multiple View Configuration

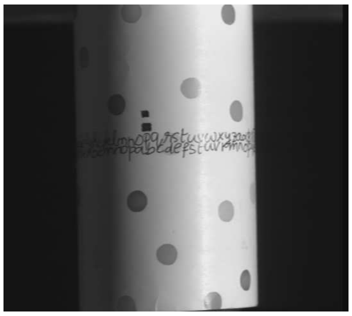
In the previous sections, we showed that choosing an appropriate binocular stereo configuration to eliminate mismatches caused by specular reflection is a viable approach for designing a working system if the roughness of objects can be estimated. Unfortunately, measurement of surface roughness is not practical in most applications. To overcome this problem, an adaptive vergence scheme could be adopted. The scheme would have to successively reduce vergence by estimating the number of mismatches at the current vergence value. In the case of a smooth cylinder, disparity variation across a scanline could serve as an estimate. However, most other surfaces would require a much more complex algorithm. This motivates us to seek for an alternate scheme.

We noticed that depth estimates at scene points are accurate when computed using certain vergences and inaccurate with others. This motivates us to seek a multiple view system from which correct depth estimates of different scene points can be combined so as to reconstruct an accurate depth map of the *entire* scene. The configuration parameters must be *independent* of roughness. In the following sections, we will see that such a multiple view configuration can be arrived at.

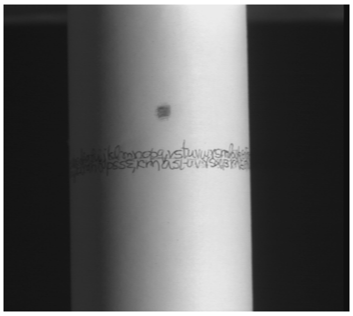
9 Problem Formulation – Multiple View Stereo

In this section, we analyze multiple view stereo configurations as shown in figure 13. The coordinate system and corresponding mapping of scene points is the same as discussed in section 4. We are interested in symmetrical planar configurations, i.e configurations in which all sensors lie in one plane and the angle between the optical axes of all adjacent sensors are equal.

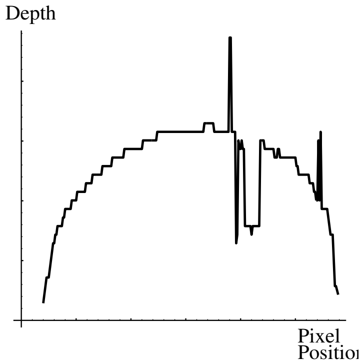
[11] We have imposed the scanline epipolarity constraint by ensuring that the robot moves in the x-z plane only.



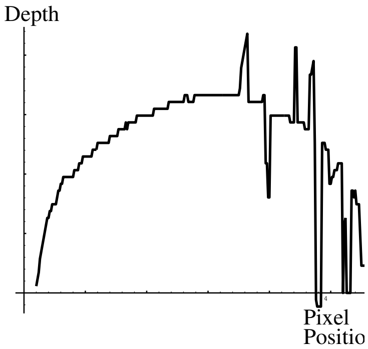
(a)



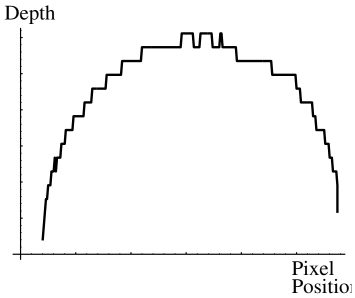
(d)



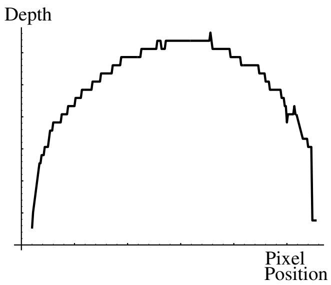
(b) $\theta_v = 8.0°$



(e) $\theta_v = 11.0°$



(c) $\theta_v = 6.0°$



(f) $\theta_v = 9.0°$

Figure 12: Stereo matching with the two objects. (a-c) Image of the object with gift wrapper surface and depth obtained along a scanline using the vergence values shown; (d-f) Image of the object with rough xerox paper surface and depth obtained along a scanline using the vergence values shown. Notice that for both surfaces, computation of depth results in large errors at higher vergence values, although a larger vergence is acceptable for the rougher surface.

If $\alpha_{i,k}$ represents the angle between the optical axes of the i^{th} and k^{th} sensors, then the following constraint ($d1$) holds:

$$\begin{aligned} \alpha_{i,i-1} &= \alpha_{i,i+1} = \alpha,\ i = 2,3,\ldots,n-1,\ n > 2 \\ \hat{v}_m \cdot \hat{k} &\geq 0 \\ \hat{v}_m \cdot \hat{j} &= 0,\ m = 1,2,\ldots,n \end{aligned} \tag{18}$$

where n denotes the number for sensors used in the system[12]. In the case when n is even, the sensor coinciding with the z-axis of the coordinate system is virtual. α is called *multiple view vergence* to distinguish it from binocular vergence. The symmetry in configuration is advantageous since it results in fewer design parameters. In order to

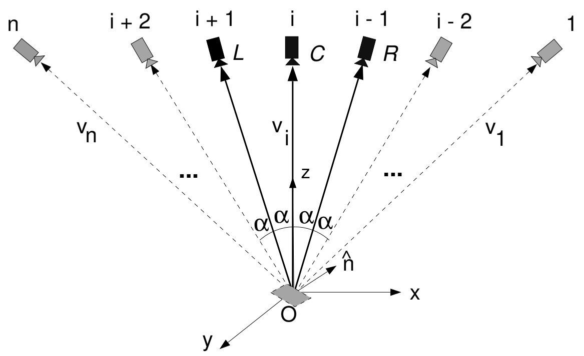


Figure 13: Layout of the multiple view configuration. A trinocular configuration is illustrated by the dark shaded cameras which are labelled as L, C and R.

be able to reconstruct an accurate depth map from multiple views, for each point in the scene at least one pair of views must provide the right depth estimate[13]. The implication is that in the presence of specular reflection, for each scene point I_{diff} must not be too large in at least one stereo pair. This constraint ($d2$) can be expressed as:

$$\exists(i,k)(|I_s^i - I_s^k| < T),\ k \neq i\,,\ 1 \leq i,k \leq n \tag{19}$$

Note that the two views which satisfy the above constraint can change from one scene point to the next. Therefore, if the constraint is satisfied for all scene points then we can design an algorithm that switches between different stereo pairs to construct a complete and accurate depth map of the scene. In addition, we are concerned with only those

[12]We have specified the constraint in terms of the angular separation between optical axes since we use orthographic projection in the following discussion. In the case of perspective projection, the constraint would be in terms of baseline distance between the sensors.

[13]In this paper we are not dealing with mismatches that could be caused by other factors like repetitive texture.

scene points in the region of stereoscopic viewing with surface normal orientations such that they are visible to all sensors and can be illuminated. Equivalently, constraint $(d3)$ is:

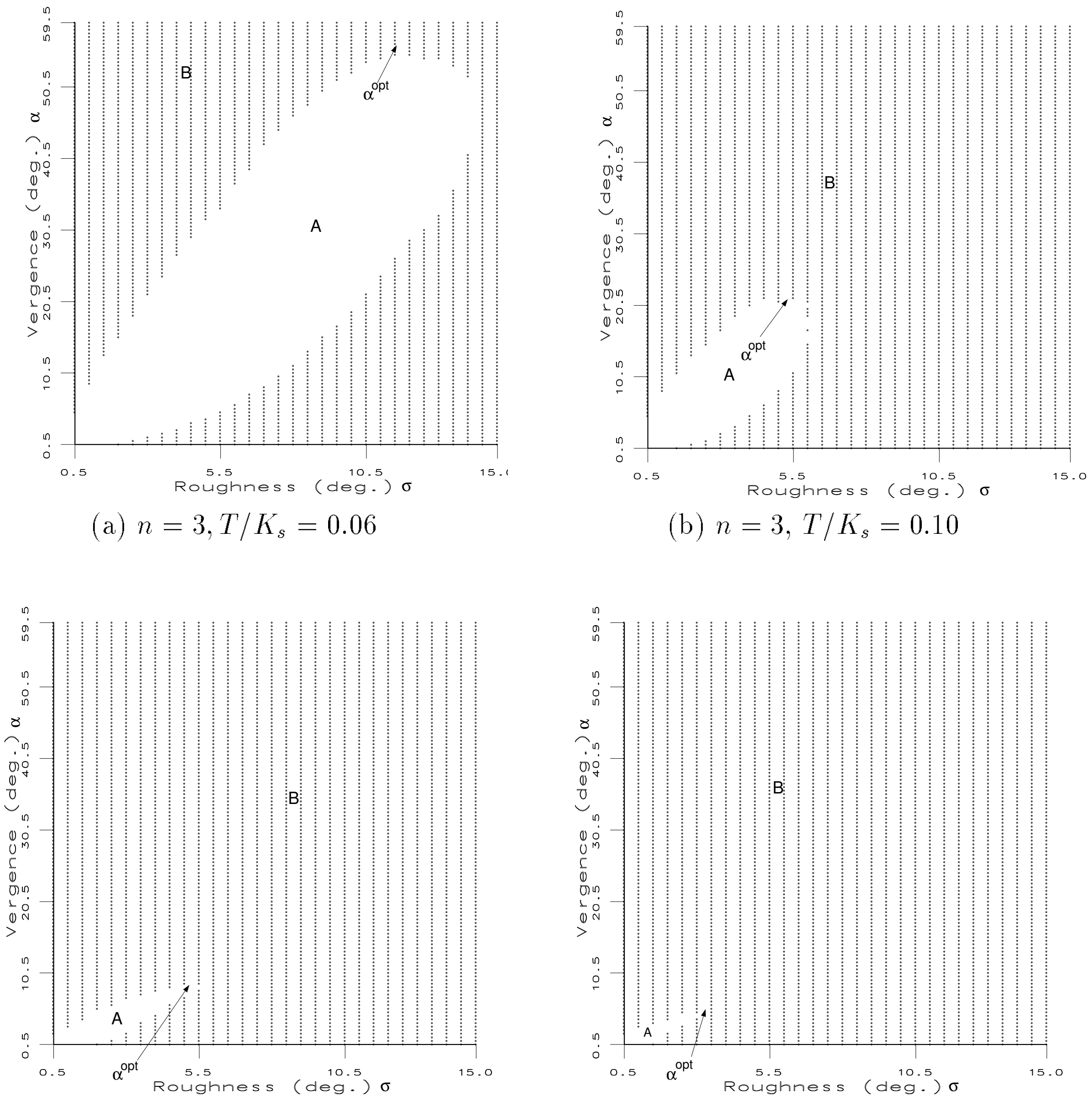
$$\begin{aligned} \hat{v}_m \cdot \hat{n} &> 0, \; m = 1, 2 \ldots, n \\ \hat{s} \cdot \hat{n} &> 0 \end{aligned} \tag{20}$$

Lastly, the unit vector constraints $(d4)$ are included:

$$||\hat{s}|| = ||\hat{n}|| = ||\hat{v}_m|| = 1, \; m = 1, 2 \ldots, n \tag{21}$$

To examine if a configuration as described above can be derived, we analyze the following problem – determine those values of α which satisfy the constraints $(d1 - d4)$. Like in the case of binocular stereo, the relative threshold T/K_s and the roughness σ are free parameters. Due to the highly nonlinear nature of the constraints, a closed-form solution for multiple view vergence as a function of the relative threshold and roughness is not feasible. Therefore, the problem is solved numerically. Figure 14 illustrates the corresponding solution space (α vs σ) for two values of n, each for two different values of T/K_s. In the case of binocular stereo, the solution space was clearly demarcated into two regions, corresponding to acceptable and unacceptable vergences for *each* σ value. In this case, however, such demarcation does not exist. Further, all $\alpha > \alpha^{opt}$ are acceptable multiple view vergence values for *any roughness value*. α^{opt} is termed as the *minimum acceptable vergence*. In other words, α^{opt} denotes that value of α beyond which it is ensured that I_{diff} does not exceed a threshold in at least one pair of views for any point with any roughness value. Two characteristics can be observed from the graphs: (a) The minimum acceptable vergence decreases with increasing T/K_s; (b) The minimum acceptable vergence decreases as more number of sensors are used. Intuitively, this is because the probability of finding a stereo pair with $I_{diff} < T$ at any point increases, for a given α. Note that the solution space defined by acceptable binocular vergences can be mapped onto the multiple view vergence solution space.

It is necessary to keep α^{opt} small in order to be able to adopt a reasonable α configuration; one where the range of stereoscopic viewing is sufficiently large. This can be achieved by using more number of sensors. However, this is not desirable due to two reasons. First, the computational cost of depth reconstruction using more number of views can outweigh the benefit of using lower multiple view vergence [Dhond and Aggarwal-1991]. Second, the implementation of the depth reconstruction algorithm increases in complexity. Therefore, we prefer a trinocular configuration ($n = 3$) at the expense of using a higher α. Apart from the advantage that a multiple view configuration can be used for all rough surfaces, the depth resolution obtained from such a configuration is much better than it's binocular counterpart for smooth surfaces. The reason is that α^{opt} is much greater than ${\theta_v}^{opt}$ for smooth surfaces.



(a) $n = 3, T/K_s = 0.06$

(b) $n = 3,\ T/K_s = 0.10$

(c) $n = 4, T/K_s = 0.06$

(d) $n = 4, T/K_s = 0.10$

Figure 14: Graphs illustrating acceptable vergences in multiple view configurations. (a-b) Three view configuration ($n = 3$); (c-d) Four view configuration ($n = 4$). The regions marked B depict acceptable vergence values and the regions marked A depict unacceptable vergence values. α^{opt} is the minimum acceptable vergence as defined in section 9.

The exact determination of the breaking relative threshold value is also not so crucial since we can choose a large α which is valid for all roughness values.

10 Reconstruction Algorithm

This section describes a algorithm for matching three views obtained using a configuration with $\alpha > \alpha^{opt}$. For convenience, we label the images as the left (L), right (R) and center (C) (see Figure 13). The three designated stereo image pairs are, (L, R), (C, R) and (L, C). The essence of the algorithm lies in determining which of the three stereo pairs provides a "good" depth estimate for any point in the scene. The algorithm suitably switches between the three stereo pairs in order to appraise correct matches.

To find a match for a pixel, the algorithm searches in a prespecified pixel range S in the related stereo image, using stereo operators NCC or SSD_m(16). To evaluate if a match is good, two confidence tests, $C1$ and $C2$, are used which are explained below. Both tests found to be sufficiently capable of detecting wrong matches.

- $C1$: Compare the SSD_m(or NCC) value obtained with a predefined threshold. If the SSD_m value is larger (or NCC is smaller), then accept the match. Otherwise, the match is not good enough. The reason is that at a wrong match, texture and shading are different, and hence similarity between the windows is expected to be poor.

- $C2$: If x_b is the current match in image I_2 for pixel x_a in I_1, then reverse the search and find the corresponding pixel for x_b by searching in I_1. This match must coincide with x_a. If not then x_b is an incorrect match for x_a. This works well in detecting mismatches due to occlusion too [Hannah-1989].

In Appendix D, depth evaluation at a point using trinocular stereo is explained. It relates the corresponding coordinates in the three images for any projected scene point. Therefore, given matching points in one stereo pair, the corresponding point in the third image can be located. This constraint is used to check consistency of matches. The outline of the algorithm is now given.

Algorithm:

(1) Initialize the current stereo pair to (L, R). The reason for choosing this pair is that it yields maximum vergence thereby providing good depth resolution.

(2) Choose a pixel x_L in L for which correspondence is to be established, and perform the following steps:

(2.1) Compute the variance of intensities in the window centered at x_L. If the variance is below a threshold, it implies lack of texture surrounding the pixel and hence it cannot be matched. Discard this pixel, and go to step 2 to process the next pixel.

(2.2) Find the corresponding pixel in R. Note that the computed correspondence may or may not be correct, due to specular reflection or occlusion. We deal with specular reflection and occlusion in identical fashion. Using confidence tests $C1$ and $C2$, evaluate the goodness of match. If the match is good, compute depth and go to step 2 for processing the next pixel. If not, then it implies that the current stereo pair (L, R) cannot be used for matching pixel x_L, and hence perform the following steps:

(2.2.1) Set (L, C) as the current stereo pair, and find the corresponding pixel (for x_L) in C. Again, evaluate the confidence of matching using $C1$ and $C2$. If the match is good, then compute the corresponding pixel x_R in image R by transformation[14]. Evaluate depth using x_L and x_R and go to step 2 to continue stereo processing. If the match is not good, then the current stereo pair has also failed to establish correspondence, and hence perform the following steps:

(2.2.1.1) Set (C, R) as the current stereo pair. The correct match for x_L must be found using this pair because the other two pairs failed to do so. Find that pixel x_C in C within the range $(x_L - S, x_L + S)$ which matches well with a pixel x_R in R, *and* together map onto x_L when transformed into the image coordinate system of L. Thus, we establish consistent correspondence for x_L in the three images. Compute depth using x_L and x_R and go to step 2. If no such consistent correspondence can be established, then we report that depth cannot be computed at point x_L, and go back to step 2 for processing the next pixel.

The worst-case algorithmic complexity can be estimated as follows. Let N be the total number of pixels for which depth is being estimated, S be the search range, and M be the size of the window used by the similarity measure. Let k_1 be the fraction of points for which the first stereo pair fails to match correctly and hence the second stereo pair has to be used, and k_2 be the fraction of points for which the second pair fails to match too and hence the third stereo pair has to be used. Then the complexity is given by, $O(2SM^2N(1 + k_1 + Sk_1k_2))$, $k_1, k_2 < 1$. It can be seen that the complexity increases with the increase in number of switches between stereo pairs. However, the increase is only linear. If only confidence test $C1$ is used, then the complexity reduces by half. Of course, while this would make the algorithm run faster, its reliability is also reduced.

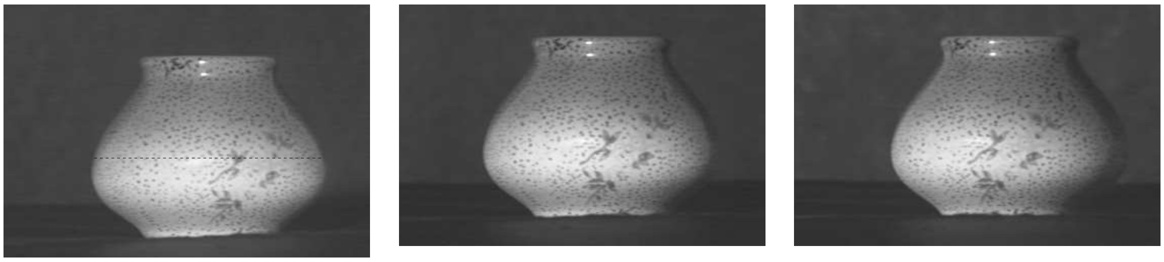
The algorithm is simple to implement and uses regular correspondence operators. It is parallelizable since finding the correspondence for one pixel is independent of others. Note that we have not used any surface continuity assumption either implicitly or explicitly. In the next section, we show the performance of this algorithm when used with real stereo images.

11 Experiments

In this section, we present trinocular stereo experiments with objects of different roughness. Here we *do not* estimate surface roughness as required in the case of binocular stereo. Figure 10 shows the photograph of the experimental stereo setup used. As with the case of experiments on binocular stereo (section 7), different vergence values are obtained by moving the camera about a calibrated center where objects are placed.

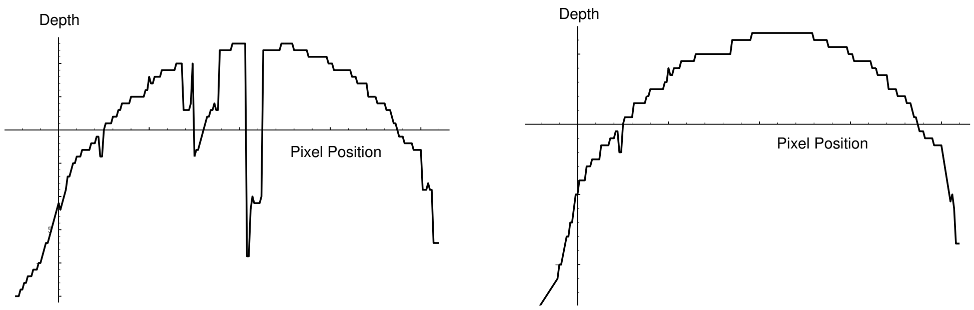
Figure 15 shows trinocular stereo images of a small vase whose surface roughness varies. The upper half of the vase is smoother than it's lower half. The images were obtained using multiple vergence value $\alpha = 10°$, i.e the binocular vergence with the left and right images is $20°$. We chose this value of α by experimentation. We used a single distant light source to capture the images, in order to keep the experiments consistent with our theory.

We first illustrate the performance of the reconstruction algorithm on a single scanline shown in image L. Notice that the specular region shifts differently in the image space in comparison to the neighbouring texture. Figure 16(a) shows depth computed along the scanline using images L and R, with a fixed window correlation algorithm. Figure 16(b) shows the depth along the same scanline using our algorithm. It can be seen that our algorithm works well. Figure 17 shows the entire depth map from two different views. Notice that matches computed within the entire highlight region in the middle of the vase is accurate. The blank regions observed in the top view implies depth could not be determined at these regions. The reason is either due to lack of texture or since matches were not satisfactory (as measured by the confidence tests). We now use the same configuration to obtain the depth map for an object of different roughness. Figure 18 shows trinocular stereo images of the egg-shaped object. Figure 19 compares our algorithm with naive binocular stereo matching along a single scanline. Finally, the depth map computed using our algorithm is shown in Figure 20. Again, it can be seen that in the computed depth is quite accurate.



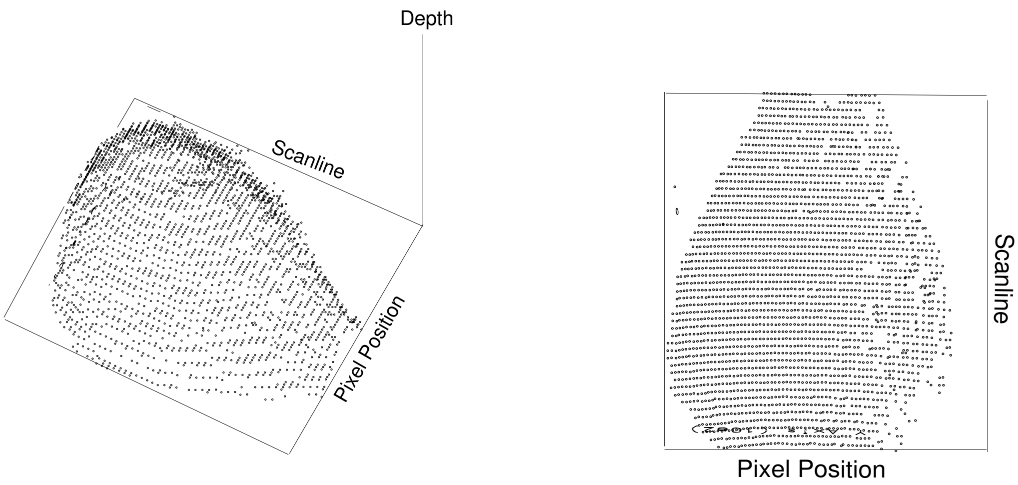
(a) Left image (L) (b) Center image (C) (c) Right image (R)

Figure 15: Trinocular stereo images of a small vase, obtained using multiple view vergence $\alpha = 10°$. The images are gamma corrected to enhance visual contrast between the specular and diffuse regions. One scanline is shown near the center of the left image using a thin dotted line. Notice that the highlight region shifts differently relative to the neighbouring texture.



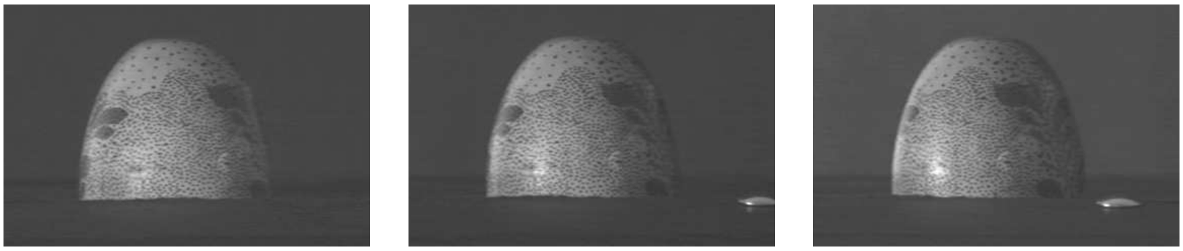
(a) (b)

Figure 16: Depth determination for the vase shown in Figure 15 along a scanline, (a) using views L and R; (b) using our reconstruction algorithm which uses all three views.



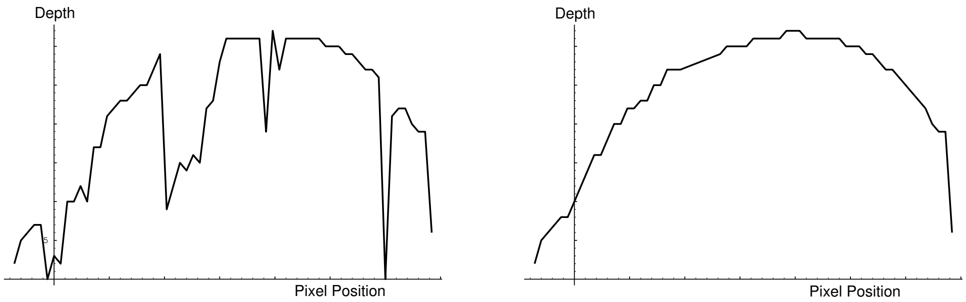
(a) Side view (b) Top view

Figure 17: The depth map of the vase computed using our reconstruction algorithm displayed from two viewpoints. The map has *not been postprocessed* using smoothing!. Notice that depth is computed accurately within the highlight region in the center of the vase.



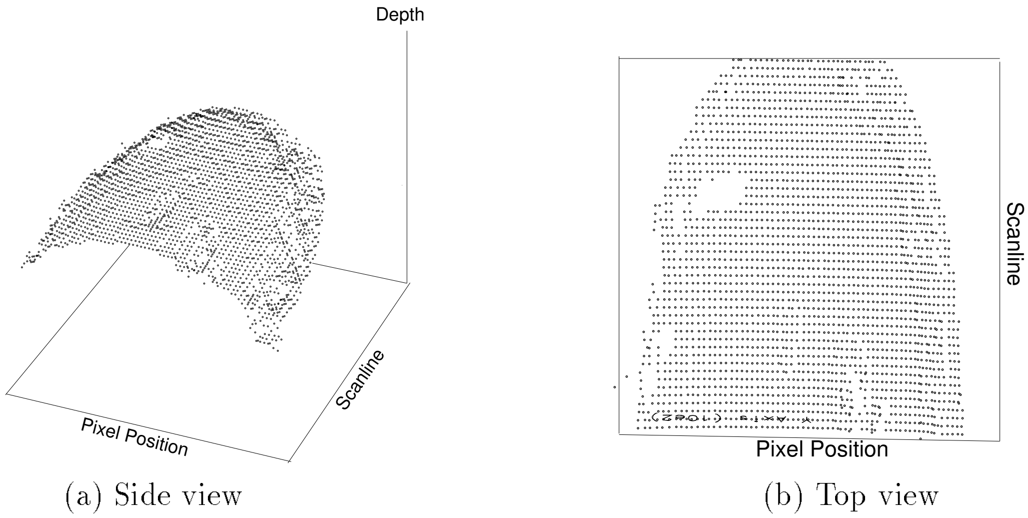
(a) Left image (L) (b) Center image (C) (c) Right image (R)

Figure 18: Trinocular stereo images of an egg-shaped object, obtained using a multiple view vergence $\alpha = 10°$. The images are gamma corrected to enhance contrast during display.



(a) (b)

Figure 19: Depth determination for the egg shaped object along a scanline, (a) using views L and R; (b) using our reconstruction algorithm which uses all three views.



(a) Side view (b) Top view

Figure 20: The depth map of the egg-shaped object computed using our reconstruction algorithm. The depth map is not postprocessed.

12 Discussion and Conclusion

We conclude our paper by summarizing its main results and contributions :

- We developed a physically based approach for reliable stereo in the presence of specular reflection. We showed that choosing appropriate stereo configurations to eliminate mismatches during correspondence is practical. To our knowledge, this is the first approach which makes explicit use of reflectance models for stereo.

- A scene independent binocular stereo solution was obtained by minimizing intensity differences at corresponding points while maximizing depth resolution. The solution was shown to be a function of surface roughness. Hence, this configuration is usable in indoor structured environments where roughness can be assessed.

- Multiple view stereo configurations (those using three or more views) were derived to obviate the need for surface roughness measurement. It implies that these configurations can be used in scenes containing different objects with varying reflectance properties.

- We developed a simple algorithm for reconstructing accurate depth maps from three views of a scene influenced by specular reflection.

In future work, we plan to generalize our reconstruction algorithm for using any number of views. We also intend to perform experiments on outdoor scenes. Determination of the relative threshold depends on the diffuse texture of the surface. The reason is that shading information from current low intensity resolution sensors does not provide enough discriminatory power for correspondence. However, we feel that with the advent of high resolution sensors, the details of shading will suffice for matching. The relative threshold will then be based on shading variation only, and hence its determination could become tractable. The problem can be formulated statistically based on probabilistic assumptions of local surface characteristics (like in shape from shading [Zheng and Chellappa-1991]).

Appendix

A Depth Resolution and Vergence

We derive the expression for depth z of any point $P(x, z)$ in the scene with respect to the world coordinate system when the cameras (Figure 3) tilt equally about the z axis. Let x_l and x_r be the projections of the point on the left and right image planes. The distance O_1O and O_2O are represented by L_1 and L_2, respectively. The depth of point P is given by:

$$z = \frac{(L_1 + L_2) - (x_l - x_r)(\sec\frac{\beta}{2})}{(2tan\frac{\beta}{2})}, \ \ \beta < \pi \qquad (22)$$

Two useful quantities to estimate depth resolution are the *absolute range error* [Verri and Torre-1986] and the *expected range error* [Rodriguez and Aggarwal-1988]. In this paper, only the absolute error is discussed. The absolute error in depth Δz is given by:

$$|\Delta z| = |\frac{\partial z}{\partial x_l}|\Delta x_l + |\frac{\partial z}{\partial x_r}|\Delta x_r + |\frac{\partial z}{\partial \beta}|\Delta \beta \tag{23}$$

Δx_l and Δx_r represent errors due to matching inaccuracies and quantization (as explained in section 3). $\Delta\beta$ represents error in camera vergence due to mechanical defects and improper calibration of the cameras. If we assume that error in camera vergence is negligible, then $\Delta\beta = 0$. Further, if wrong matches are eliminated, then depth errors are primarily due to quantization. Using (22) and (23), the depth error can be expressed as:

$$\frac{|\Delta z|}{(\Delta x_l + \Delta x_r)} = \frac{1}{2sin\frac{\beta}{2}} = \frac{1}{2sin\frac{\theta_v}{2}} \tag{24}$$

where the term, $(\Delta x_l + \Delta x_r)$, represents correspondence error due to quantization. From (24), it follows that the absolute depth error is inversely proportional to vergence. In other words, depth resolution increases with increasing vergence.

B Relationship between Field Vergence and Baseline

First, we derive the expression for point vergence at a point $P(x, z)$ (Figure 4). Let α_1 be the angle between the projected ray from P to the left image plane and the vertical, and α_2 be the angle between the projected ray from P to the right image plane and the vertical, i.e $\alpha_1 + \alpha_2 = \theta_v$. The distance O_1O and O_2O are denoted by L_1 and L_2, respectively. The point vergence at point $P(x, z)$ is given by:

$$\begin{aligned} tan\theta_v &= \frac{tan\alpha_1 + tan\alpha_2}{1 - tan\alpha_1 tan\alpha_2} && (25) \\ tan\alpha_1 &= \frac{L_1 + x}{z},\ tan\alpha_2 = \frac{L_2 - x}{z} \\ L_1 + L_2 &= 2L && (26) \end{aligned}$$

Substituting (26) in the expression for point vergence, we get:

$$\theta_v = \arctan\frac{2Lz}{(z^2 + x(L_1 - L_2) + x^2 - L_1L_2)} \tag{27}$$

If the cameras tilt equally about the z axis, then $L_1 = L_2$, and the above equation reduces to:

$$\theta_v = \arctan\frac{2Lz}{(z^2 + x^2 - L^2)} \tag{28}$$

Field vergence $\bar{\theta}_v$ can be obtained by integrating point vergences over a suitable workspace as explained in section 3. Using (6), it can be observed that field vergence is a function of a single parameter, namely the baseline.

C A Statistical Interpretation

The problem of minimizing I_{diff} can be viewed in a statistical sense. To do that, we use the following model for the images in one-dimension:

$$\begin{aligned} I_1(x) &= I_d(x) + I_{s1}(\hat{n}(x)) \\ I_2(x) &= I_d(x + d(x)) + I_{s2}(\hat{n}(x)) \end{aligned} \tag{29}$$

where $I_d(x)$ is the signal corresponding to the diffuse component, $I_s(n(x))$ is the component corresponding to the specular component, $\hat{n(x)}$ is the surface normal at point x, d is the displacement between the images. To simplify discussion, we neglect noise and perspective distortion effects. In contrast to [Matthies-1992], we consider the specular component, but neglect noise.

The specular intensity difference between corresponding points in a window is modelled as a uniformly distributed random variable, as given by (30). The chosen viewing directions decide the threshold T. Note that we have chosen to model I_{diff} rather than $\hat{n}$ because no assumption can be made regarding the surface normal in any window.

$$\begin{aligned} f(I_{diff} = |I_{s1} - I_{s2}|) &= \frac{1}{2T}, \; -T \leq I_{diff} \leq T \\ &= 0, \; otherwise \end{aligned} \tag{30}$$

The variance σ_s^2 of I_{diff} is given by $\sigma_s^2 = \frac{T^2}{3}$.

To compute the disparity at any point x_i in the left image, we use the absolute value of the differences (or squared difference) between point intensities, with the left window centered at x_i. The intensity differences between two windows being matched is expressed as ([Matthies-1992]):

$$\begin{aligned} \hat{e}(x_i; d) &= [e(x_i + \Delta x_1; d), ..., e(x_i + \Delta x_n; d)] \\ e(x_i + \Delta x_j) &= |I_1(x_i + \Delta x_j - d) - I_2(x_i + \Delta x_j)| \end{aligned} \tag{31}$$

where Δx_j represents pixels in the windows, d is the displacement between the windows, and n is the size of the window. Under the distribution model adopted , the conditional p.d.f of $\hat{e}$ at the right match estimated by d_0 (also called the likelihood function), is:

$$f(\hat{e}|d = d_0) = \Pi_{i=1}^{n} e_i \tag{32}$$

The uncertainty of the disparity estimate is directly given by:

$$E(d_0^2) = \sigma_s^2 = \frac{T^2}{3} \tag{33}$$

To increase the reliability of the disparity estimate, T must be minimized. Or in other words minimizing T, maximizes the likelihood function f (32) [Matthies-1992].

D Depth Determination using Trinocular Stereo

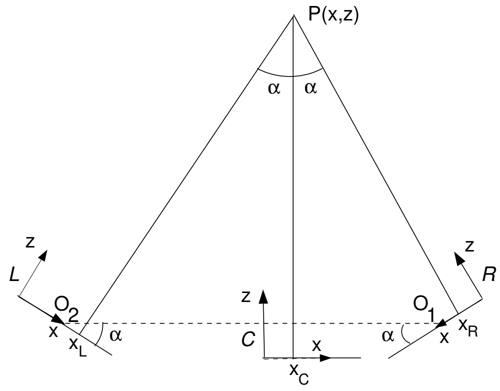


Figure 21: Correspondence of a scene point in the three images under orthographic projection.

Here we relate the x-coordinates of the projections of a point on the three stereo images. Figure 21 shows orthographic projections of a point $P(x,z)$ on the images L,R and C. The projections are denoted by x_L, x_R and x_C, respectively. Using simple algebra, it can be shown that the depth of P in the world coordinate system (which coincides with the image coordinate system C) obtained using the stereo pair indicated in superscripts, is given by:

$$\begin{aligned} z^{(L,R)} &= \frac{B + \frac{x_R - x_L}{cos\alpha}}{2\, tan\alpha} \\ z^{(L,C)} &= \frac{B + 2(x_C - \frac{x_L}{cos\alpha})}{2\, tan\alpha} \\ z^{(R,C)} &= \frac{B + 2(\frac{x_R}{cos\alpha} - x_C)}{2\, tan\alpha} \end{aligned} \tag{34}$$

where B refers to the distance between the viewpoints $(O_1 O_2)$. Equating z in the above relations, we relate the x-coordinates of the three projections:

$$x_L = 2x_C\, cos\alpha - x_R \tag{35}$$

References

[Barnard and Fischler, 1982] S. T. Barnard and M. A. Fischler. Computational stereo. *ACM Computing Surveys*, 14(4):553–572, December 1982.

[Blake, 1985] A. Blake. Specular stereo. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 973–976, 1985.

[Brelstaff and Blake, 1988] G. Brelstaff and A. Blake. Detecting specular reflections using lambertian constraints. *Proceedings of the IEEE Computer Society International Conference on Computer Vision*, pages 297–302, 1988.

[Ching *et al.*, 1993] Wee-Soon Ching, Peng-Seng Toh, Kap-Luk, and Meng-Hwa Er. Robust vergence with concurrent detection of occlusion and specular highlights. *Proceedings of the IEEE Computer Society International Conference on Computer Vision*, pages 384–394, 1993.

[Dhond and Aggarwal, 1991] U. R. Dhond and J. K. Aggarwal. A cost-benefit analysis of a third camera for stereo correspondence. *International Journal of Computer Vision*, 6:39–58, 1991.

[Diner and Fender, 1993] D. B. Diner and D. H. Fender. *Human Engineering in Stereoscopic Viewing Devices*. Plenum Press, 1993.

[Hannah, 1989] M. J. Hannah. A system for digitial stereo image matching. *Photogrammetric Engg. and Remote Sensing*, 55(12):1765–1770, 1989.

[Healey and Binford, 1988] G. Healey and T. O. Binford. Local shape from specularity. *Computer Vision, Graphics, and Image Processing*, 42:62–86, 1988.

[Ito and Ishii, 1986] M. Ito and A. Ishii. Range and shape measurement using three-view stereo analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9–14, 1986.

[Kaplan, 1959] W. Kaplan. *Advanced Calculus*. Addison-Wesley, 1959.

[Matthies, 1992] L. Matthies. Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. *International Journal of Computer Vision*, 8(1):71–91, 1992.

[Nayar *et al.*, 1991] S. K. Nayar, K. Ikeuchi and T. Kanade. Surface reflection: Physical and geometrical perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):611–634, July 1991.

[Nayar *et al.*, 1993] S. K. Nayar, Xi-Sheng Fang and T. Boult. Removal of specularities using color and polarization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 583–590, 1993.

[Okutomi and Kanade, 1993] M. Okutomi and T. Kanade. A multiple-basline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4), April 1993.

[Oren and Nayar, 1994] M. Oren and S. K. Nayar. Generalization of the lambertian model and implications for machine vision. *Proceedings of the European Conference on Computer Vision*, pages 269–280, 1994.

[Panton, 1978] D. A. Panton. A flexible approach to digital stereo mapping. *Photogrammetric Engg. and Remote Sensing*, 44(12):1499–1512, 1978.

[Pietikainen and Harwood, 1986] M. Pietikainen and D. Harwood. Depth from three camera stereo. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2–8, 1986.

[Press *et al.*, 1989] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C.* Cambridge University Press, 1989.

[Rodriguez and Aggarwal, 1988] J. J. Rodriguez and J. K. Aggarwal. Quantization error in stereo imaging. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 153–158, 1988.

[Smith, 1986] G. B. Smith. Stereo integral equation. *Proceedings of the AAAI*, pages 689–694, 1986.

[Torrance and Sparrow, 1967] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America*, 57:1105–1114, 1967.

[Verri and Torre, 1986] A. Verri and V. Torre. Absolute depth estimate in stereopsis. *Journal of the Optical Society of America*, 3:297–299, March 1986.

[Wolff and Angelopoulou, 1994] L.B. Wolff and E. Angelopoulou. 3-d stereo using photometric ratios. *Proceedings of the European Conference on Computer Vision*, pages 247–258, 1994.

[Yachida *et al.*, 1986] M. Yachida, Y. Kitamura and M. Kimachi. Trinocular vision: New approach for correspondence problem. *Proceedings of the Eighth International Conference on Pattern Recognition*, pages 27–31, 1986.

[Zheng and Chellappa, 1991] Qinfen Zheng and R. Chellappa. Estimation of illuminant direction, albedo, and shape from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):680–702, July 1991.