Height and Gradient from Shading

Berthold K.P. Horn

Abstract: *The method described here for recovering the shape of a surface from a shaded image can deal with complex, wrinkled surfaces. Integrability can be enforced easily because both surface height and gradient are represented (A gradient field is integrable if it is the gradient of some surface height function). The robustness of the method stems in part from linearization of the reflectance map about the current estimate of the surface orientation at each picture cell (The reflectance map gives the dependence of scene radiance on surface orientation). The new scheme can find an exact solution of a given shape-from-shading problem even though a regularizing term is included. The reason is that the penalty term is needed only to stabilize the iterative scheme when it is far from the correct solution; it can be turned off as the solution is approached. This is a reflection of the fact that shape-from-shading problems are not ill-posed when boundary conditions are available, or when the image contains singular points.*

This paper includes a review of previous work on shape from shading and photoclinometry. Novel features of the new scheme are introduced one at a time to make it easier to see what each contributes. Included is a discussion of implementation details that are important if exact algebraic solutions of synthetic shape-from-shading problems are to be obtained. The hope is that better performance on synthetic data will lead to better performance on real data.

Key Words: Photoclinometry, Shape from Shading, Integrability, Smoothness constraint, Variational methods, Depth and Slope, Height and Gradient, Digital Elevation Models.

Acknowledgements: This paper describes research done in part at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124.

1. Background

The first method developed for solving a shape-from-shading problem was restricted to surfaces with special reflecting properties [Rindfleisch 66]. For the surfaces that Rindfleisch considered, profiles of the solution can be obtained by integrating along predetermined straight lines in the image. The general problem was formulated and solved later [Horn 70, 75], using the method of characteristic strip expansion [Garabedian 64] [John 78] applied to the nonlinear first-order partial differential *image irradiance equation*. When the light sources and the viewer are far away from the scene being viewed, use of the *reflectance map* makes the analysis of shape-from-shading algorithms much easier [Horn 77] [Horn & Sjoberg 79]. Several iterative schemes, mostly based on minimization of some functional containing an integral of the brightness error, arose later [Woodham 77] [Strat 79] [Ikeuchi & Horn 81] [Kirk 84, 87] [Brooks & Horn 85] [Horn & Brooks 86] [Frankot & Chellappa 88].

The new method presented here was developed in part as a response to recent attention to the question of integrability[1] [Horn & Brooks 86] [Frankot & Chellappa 88] and exploits the idea of a coupled system of equations for depth and slope [Harris 86, 87] [Horn 88]. It borrows from well-known variational approaches to the problem [Ikeuchi & Horn 81] [Brooks & Horn 85] and an existing least-squares method for estimating surface shape given a needle diagram (see [Ikeuchi 84], chapter 11 in [Horn 86], and [Horn & Brooks 86]). For one choice of parameters, the new method becomes similar to one of the first iterative methods ever developed for shape from shading on a regular grid [Strat 79], while it degenerates into another well-known method [Ikeuchi & Horn 81] for a different choice of parameters. If the brightness error term is dropped, then it becomes a surface interpolation method [Harris 86, 87]. The computational effort grows rapidly with image size, so the new method can benefit from proper multigrid implementation [Brandt 77, 80, 84] [Brandt & Dinar 79] [Hackbush 85] [Hackbush & Trottenberg 82], as can existing iterative shape-from-shading schemes [Terzopolous 83, 84] [Kirk 84, 87]. Alternatively, one can apply so-called direct methods for solving Poisson's equations [Simchony, Chellappa & Shao 89].

Experiments indicate that linear expansion of the reflectance map about the current estimate of the surface gradient leads to more rapid convergence. More importantly, this modification often allows the scheme to converge when simpler schemes diverge, or get stuck in local minima of the functional. Most existing iterative shape-from-shading methods

[1] A gradient field is integrable if it is the gradient of some surface height function.

handle only relatively simple surfaces and so could benefit from a retrofit of this idea.

The new scheme was tested on a number of synthetic images of increasing complexity, including some generated from digital terrain models of steep, wrinkled surfaces, such as a glacial cirque with numerous gullies. Shown in Figure 1(a) is a shaded view of a digital terrain model, with lighting from the Northwest. This is the input provided to the algorithm. The underlying 231 × 178 digital terrain model was constructed from a detailed contour map, shown in Figure 2, of Huntington ravine on the eastern slopes of Mount Washington in the White Mountains of New Hampshire[2]. Shown in Figure 1(b) is a shaded view of the same digital terrain model with lighting from the Northeast. This is *not* available to the algorithm, but is shown here to make apparent features of the surface that may not stand out as well in the other shaded view. Figure 1(c) shows a shaded view of the surface reconstructed by the algorithm, with lighting from the Northwest—it matches Figure 1(a) exactly. More importantly, the shaded view of the reconstructed surface with lighting from the Northeast, shown in Figure 1(d), matches Figure 1(b) exactly also[3].

With proper boundary conditions, the new scheme recovers surface orientation *exactly* when presented with noise-free synthetic scenes[4]. Previous iterative schemes do not find the exact solution, and in fact wander away from the correct solution when it is used as the initial guess. To obtain exact algebraic solutions, several details of the implementation have to be carefully thought through, as discussed in section 6. Simple surfaces are easier to process—with good results even when several of the implementation choices are not made in an optimal way. Similarly, these details perhaps may be of lesser importance for real images, where other error sources could dominate.

In the next few sections we review image formation and other elementary ideas underlying the usual formulation of the shape-from-shading problem. Photoclinometry is also briefly reviewed for the benefit of researchers in machine vision who may not be familiar with this field. We then discuss both the original and the variational approach to the shape-

[2] The gullies are steep enough to be of interest to ice-climbers.

[3] For additional examples of reconstructions from shaded images, see section 7.

[4] In the examples tried, the algorithm always recovered the underlying surface orientation exactly at every picture cell, starting from a random surface orientation field, provided that boundary information was available. Since the question of uniqueness of solutions has not been totally resolved, one cannot be quite certain that there may not be cases where a different solution might be found that happens to also fit the given image data exactly.

Figure 1. Reconstruction of surface from shaded image. See text.

Figure 2. Contour map from which the digital terrain model used to synthesize Figures 1(a) and (b) was interpolated. The surface was modeled as a thin plate constrained to pass through the contours at the specified elevations. The interpolating surface was found by solving the biharmonic equation, as described at the end of section 5.4.

from-shading problem. Readers familiar with the basic concepts may wish to skip over this material and go directly to section 5, where the new scheme is derived. For additional details see chapters 10 and 11 in *Robot Vision* [Horn 86] and the collection of papers, *Shape from Shading* [Horn & Brooks 89].

2. Review of Problem Formulation

2.1 Image Projection and Image Irradiance

For many problems in machine vision it is convenient to use a camera-centered coordinate system with the origin at the center of projection and the Z-axis aligned with the optical axis (the perpendicular from the center of projection to the image plane)[5]. We can align the X- and Y-axes with the image plane x- and y-axes. Let the *principal distance* (that is, the perpendicular distance from the center of projection to the image plane) be f, and let the image plane be reflected through the center of projection so that we avoid sign reversal of the coordinates. Then the perspective projection equations are

$$x = f\,\frac{X}{Z} \qquad \text{and} \qquad y = f\,\frac{Y}{Z}. \tag{1}$$

The shape-from-shading problem is simplified if we assume that the depth range is small compared with the distance of the scene from the viewer (which is often the case when we have a narrow field of view, that is, when we use a telephoto lens). Then we have

$$x \approx \frac{f}{Z_0}\,X \qquad \text{and} \qquad y \approx \frac{f}{Z_0}\,Y, \tag{2}$$

for some constant Z_0, so that the projection is approximately orthographic. In this case it is convenient to rescale the image coordinates so that we can write $x = X$ and $y = Y$. For work on shape from shading it is also convenient to use z, height above some reference plane perpendicular to the optical axis, rather than the distance measured along the optical axis from the center of projection.

If we ignore vignetting and other imaging system defects, then image irradiance E at the point (x, y) is related to scene radiance L at the corresponding point in the scene by [Horn 86]

$$E = L\,\frac{\pi}{4}\left(\frac{d}{f}\right)^2 \cos^4\alpha, \tag{3}$$

where d is the diameter of the lens aperture, f is the principal distance, and the off-axis angle α is given by

$$\tan\alpha = \frac{1}{f}\sqrt{x^2 + y^2}. \tag{4}$$

Accordingly, image irradiance[6] is a multiple of the scene radiance, with the factor of proportionality depending inversely on the square of the f-

[5] In photoclinometry it is customary to use an object-centered coordinate system. This is because surface shape can be computed along profiles only when strong additional constraint is provided, and such constraints are best expressed in an object-centered coordinate system. Working in an object-centered coordinate system, however, makes the formulation of the shape-from-shading problem considerably more complex (see, for example, [Rindfleisch 66]).

[6] Grey-levels are quantized estimates of image irradiance.

number[7]. If we have a narrow field of view, the dependence on the off-axis angle α can be neglected. Alternatively, we can normalize the image by dividing the observed image irradiance by $\cos^4 \alpha$ (or whatever the actual vignetting function happens to be).

We conclude from the above that what we measure in the image is directly proportional to scene radiance, which in turn depends on (a) the strength and distribution of illumination sources, (b) the surface microstructure and (c) surface orientation.

In order to be able to solve the shape from shading problem from a single image we must assume that the surface is uniform in its reflecting properties. If we also assume that the light sources are far away, then the irradiance of different parts of the scene will be approximately the same and the incident direction may be taken as constant. Finally, if we assume that the viewer is far away, then the direction to the viewer will be roughly the same for all points in the scene. Given the above, we find that scene radiance does not depend on the position in space of a surface patch, only on its orientation.

2.2 Specifying Surface Orientation

Methods for recovering shape from shading depend on assumptions about the continuity of surface height and its partial derivatives. First of all, since shading depends only on surface orientation, we must assume that the surface is continuous and that its first partial derivatives exist. Most formulations implicitly also require that the first partial derivatives be continuous, and some even require that second partial derivatives exist. The existence and continuity of derivatives lends a certain “smoothness” to the surface and allows us to construct local tangent planes. We can then talk about the local surface orientation in terms of the orientation of these tangent planes.

There are several commonly used ways of specifying the orientation of a planar surface patch, including:

- Unit surface normal $\hat{\mathbf{n}}$ [Horn & Brooks 86];
- Point on the Gaussian sphere [Horn 84];
- Surface gradient (p, q) [Horn 77];
- Stereographic coordinates (f, g) [Ikeuchi & Horn 81];

[7] The f-number is the ratio of the principal distance to the diameter of the aperture, that is, f/d.

- Dip and strike (as defined in geology)[8];
- Luminance longitude and latitude (as defined in astrogeology)[9];
- Incident and emittance angles (i and e)[10];

For our purposes here, the components of the surface gradient

$$p = \frac{\partial z}{\partial x} \qquad \text{and} \qquad q = \frac{\partial z}{\partial y}, \tag{5}$$

will be most directly useful for specifying surface orientation.

We can convert between different representations easily. For example, suppose that we are to determine the unit surface normal given the gradient components. We know that if we move a small distance δx in x, then the change in height is $\delta z = p\,\delta x$ (since p is the slope of the surface in the x direction). Thus $(1, 0, p)^T$ is a tangent to the surface. If we move a small distance δy in y, then the change in height is $\delta z = q\,\delta y$ (since q is the slope of the surface in the y direction). Thus $(0, 1, q)^T$ is also a tangent to the surface. The normal is perpendicular to all tangents, thus parallel to the cross-product of these particular tangents, that is parallel to $(-p, -q, 1)^T$. Hence a unit normal can be written in the form

$$\hat{\mathbf{n}} = \frac{1}{\sqrt{1 + p^2 + q^2}}(-p, -q, 1)^T. \tag{6}$$

Note that this assumes that the z-component of the surface normal is positive. This is not a problem since we can only see surface elements whose normal vectors point within $\pi/2$ of the direction toward the viewer—other surface elements are turned away from the viewer.

We can use the same notation to specify the direction to a collimated light source or a small portion of an extended source. We simply give the orientation of a surface element that lies perpendicular to the incident

[8] Dip is the angle between a given surface and the horizontal plane, while strike is the direction of the intersection of the surface and the horizontal plane. The line of intersection is perpendicular to the direction of steepest descent.

[9] Luminance longitude and latitude are the longitude and latitude of a point on a sphere with the given orientation, measured in a spherical coordinate system with the poles at right angles to both the direction toward the source and the direction toward the viewer.

[10] Incident and emittance angles are meaningful quantities only when there is a single source; and even then there is a two-way ambiguity in surface orientation unless additional information is provided. The same applies to luminance longitude and latitude.

light rays. So we can write[11]

$$\hat{\mathbf{s}} = \frac{1}{\sqrt{1 + p_s^2 + q_s^2}} (-p_s, -q_s, 1)^T, \tag{7}$$

for some p_s and q_s.

2.3 Reflectance Map

We can show the dependence of scene radiance on surface orientation in the form of a *reflectance map* $R(p, q)$. The reflectance map can be depicted graphically in gradient space[12] as a series of nested contours of constant brightness [Horn 77, 86].

The reflectance map may be determined experimentally by mounting a sample of the surface on a goniometer stage and measuring its brightness under the given illuminating conditions for various orientations. Alternatively, one may use the image of a calibration object (such as a sphere) for which surface orientation is easily calculated at every point. Finally, a reflectance map may be derived from a phenomenological model, such as that of a Lambertian surface. In this case one can integrate the product of the *bidirectional reflectance distribution function* (BRDF) and the given distribution of source brightness as a function of incident angle [Horn & Sjoberg 79].

An ideal Lambertian surface illuminated by a single point source provides a convenient example of a reflectance map[13]. Here the scene radiance is given by $R(p, q) = (E_0/\pi) \cos i$, where i is the incident angle (the angle between the surface normal and the direction toward the source), while E_0 is the irradiance from the source on a surface oriented perpendicular to the incident rays. (The above formula only applies when $i \le \pi/2$; the scene radiance is, of course, zero for $i > \pi/2$.) Now $\cos i = \hat{\mathbf{n}} \cdot \hat{\mathbf{s}}$, so

$$R(p, q) = \frac{E_0}{\pi} \frac{1 + p_s p + q_s q}{\sqrt{1 + p^2 + q^2}\sqrt{1 + p_s^2 + q_s^2}}, \tag{8}$$

as long as the numerator is positive, otherwise $R(p, q) = 0$.

[11] There is a small problem, however, with this method for specifying the direction toward the light source: A source may be "behind" the scene, with the direction to the source more than $\pi/2$ away from the direction toward the viewer. In this case the z-component of the vector pointing toward the light source is negative.

[12] The coordinates of *gradient space* are p and q, the slopes of the surface in the x and y direction respectively.

[13] Note that shape-from-shading methods are most definitely *not* restricted to Lambertian surfaces. Such special surfaces merely provide a convenient pedagogical device for illustrating basic concepts.

2.4 Image Irradiance Equation

We are now ready to write down the *image irradiance equation*

$$E(x,y) = \beta R(p(x,y), q(x,y)), \tag{9}$$

where $E(x,y)$ is the irradiance at the point (x,y) in the image, while $R(p,q)$ is the radiance at the corresponding point in the scene, at which $p = p(x,y)$ and $q = q(x,y)$. The proportionality factor β depends on the f-number of the imaging system (and may include a scaling factor that depends on the units in which the instrument measures brightness). It is customary to rescale image irradiance so that this proportionality factor may be dropped. If the reflectance map has a unique global extremum, for example, then the image can be normalized in this fashion, provided that a point can be located that has the corresponding surface orientation[14].

Scene radiance also depends on the irradiance of the scene and a reflectance factor (loosely called *albedo* here). These factors of proportionality can be combined into one that can be taken care of by normalization of image brightness. Then only the geometric dependence of image brightness on surface orientation remains in $R(p,q)$, and we can write the image irradiance equation in the simple form

$$E(x,y) = R(p(x,y), q(x,y)) \tag{10}$$

or

$$E(x,y) = R(z_x(x,y), z_y(x,y)), \tag{11}$$

where $p = z_x$ and $q = z_y$ are the first partial derivatives of z with respect to x and y. This is a first-order partial differential equation; one that is typically nonlinear, because the reflectance map in most cases depends nonlinearly on the gradient.

2.5 Reflectance Map Linear in Gradient

Viewed from a sufficiently great distance, the material in the maria of the moon has the interesting property that its brightness depends only on luminance longitude, being independent of luminance latitude [Hapke 63, 65]. When luminance longitude and latitude are related to the incident and emittance angles, it is found that longitude is a function of $(\cos i/\cos e)$. From the above we see that $\cos i = \hat{\mathbf{n}} \cdot \hat{\mathbf{s}}$, while $\cos e = \hat{\mathbf{n}} \cdot \hat{\mathbf{v}}$, where $\hat{\mathbf{v}} =$

[14] If there is a unique maximum in reflected brightness, it is convenient to rescale the measurements so that this extremum corresponds to $E = 1$. The same applies when there is a unique minimum, as is the case for the scanning electron microscope (SEM).

$(0, 0, 1)^T$ is a unit vector in the direction toward the viewer. Consequently,

$$\frac{\cos i}{\cos e} = \frac{\hat{\mathbf{n}} \cdot \hat{\mathbf{s}}}{\hat{\mathbf{n}} \cdot \hat{\mathbf{v}}} = \frac{1}{\sqrt{1 + p_s^2 + q_s^2}} (1 + p_s p + q_s q). \tag{12}$$

Thus $(\cos i / \cos e)$ depends linearly on the gradient components p and q, and we can write

$$R(p, q) = f(c\,p + s\,q), \tag{13}$$

for some function f and some coefficients c and s. Both Lommel-Seeliger's and Hapke's functions fit this mold [Minnaert 61] [Hapke 63, 65]. (For a few other papers on the reflecting properties of surfaces, see [Hapke 81, 84] [Hapke & Wells 81] and the bibliography in [Horn & Brooks 89].) We can, without loss of generality[15], arrange for $c^2 + s^2 = 1$.

If the function f is continuous and monotonic[16], we can find an inverse

$$c\,p + s\,q = f^{-1}(E(x, y)). \tag{14}$$

The slope in the image direction (c, s) is

$$m = \frac{c\,p + s\,q}{\sqrt{c^2 + s^2}} = \frac{1}{\sqrt{c^2 + s^2}} f^{-1}(E(x, y)). \tag{15}$$

We can integrate[17] out this slope along the line

$$x(\xi) = x_0 + c\,\xi \qquad \text{and} \qquad y(\xi) = y_0 + s\,\xi, \tag{16}$$

to obtain

$$z(\xi) = z_0 + \frac{1}{\sqrt{c^2 + s^2}} \int_0^{\xi} f^{-1}\Big(E\big(x(\eta), y(\eta)\big)\Big)\, d\eta. \tag{17}$$

An extension of the above approach allows one to take into account perspective projection as well as finite distance to the light source [Rindfleisch 66]. Two changes need to be made; one is that the reflectance map now is no longer independent of image position (since the directions to the viewer and the source vary significantly); and the other that the integral is for the logarithm of the radial distance from the center of projection, as opposed to distance measured parallel to the optical axis.

The above was the first shape-from-shading or photoclinometric problem ever solved in other than a heuristic fashion. The original formulation was considerably more complex than described above, as the result of the

[15] We see that $c : s = p_s : q_s$, so that the direction specified in the image by (c, s) is the direction "toward the source," that is, the projection into the image plane of the vector $\hat{\mathbf{s}}$ toward the light source.

[16] If the function f is not monotonic, there will be more than one solution for certain brightness values. In this case one may need to introduce assumptions about continuity of the derivatives in order to decide which solution to choose.

[17] The integration is, of course, carried out numerically, since the integrand is derived from image measurements and not represented as an analytic function.

use of full perspective projection, the lack of the notion of anything like the reflectance map, and the use of an object-centered coordinate system [Rindfleisch 66].

Note that we obtain profiles of the surface by integrating along predetermined straight lines in the image. Each profile has its own unknown constant of integration, so there is a great deal of ambiguity in the recovery of surface shape. In fact, if $z(x, y)$ is a solution, so is

$$\overline{z}(x, y) = z(x, y) + g(s\,x - c\,y) \tag{18}$$

for an arbitrary function g! This is true because

$$\overline{z}_x = z_x + s\,g'(s\,x - c\,y) \quad \text{and} \quad \overline{z}_y = z_y - c\,g'(s\,x - c\,y), \tag{19}$$

so

$$c\,\overline{p} + s\,\overline{q} = c\,p + s\,q, \tag{20}$$

where $\overline{p} = \overline{z}_x$ and $\overline{q} = \overline{z}_y$. It follows that $R(\overline{p}, \overline{q}) = R(p, q)$. This ambiguity can be removed if an *initial curve* is given from which the profiles can be started. Such an initial curve is typically not available in practice. Ambiguity is not restricted to the special case of a reflectance map that is linear in the gradient: Without additional constraint shape-from-shading problems typically do not have a unique solution.

2.6 Low Gradient Terrain and Oblique Illumination

If we are looking at a surface where the gradient (p, q) is small, we can approximate the reflectance map using series expansion

$$R(p, q) \approx R(0, 0) + p\,R_p(0, 0) + q\,R_q(0, 0). \tag{21}$$

This approach does not work when the reflectance map is rotationally symmetric, since the first-order terms then drop out[18]. If the illumination is oblique, however, we can apply the method in the previous section to get a first estimate of the surface. Letting $c = R_p(0, 0)$, $s = R_q(0, 0)$ and

$$f^{-1}(E(x, y)) = E(x, y) - R(0, 0), \tag{22}$$

we find that

$$z(\xi) = z_0 + \frac{1}{\sqrt{R_p^2(0, 0) + R_q^2(0, 0)}} \int_0^{\xi} \Big(E(x(\eta), y(\eta)) - R(0, 0)\Big)\, d\eta. \tag{23}$$

(For a related frequency domain approach see [Pentland 88].)

One might imagine that the above would provide a good way to get initial conditions for an iterative shape from shading method. Unfortunately, this is not very helpful, because of the remaining ambiguity in the

[18] The reflectance map is rotationally symmetric, for example, when the source is where the viewer is, or when an extended source is symmetrically distributed about the direction toward the viewer.

direction at right angles to that of profile integration. Iterative methods already rapidly get adequate variations in height along "down-sun profiles," but then struggle for a long time to try to get these profiles tied together in the direction at right angles.

The above also suggests that errors in gradients of a computed solution are likely to be small in the direction towards or away "from the source" and large in the direction at right angles. It should also be clear that it is relatively easy to find solutions for slowly undulating surfaces (where p and q remain small) with oblique illumination (as in [Kirk 87]). It is harder to deal with cases where the surface gradient varies widely, and with cases where the source is near the viewer (see also the discussion in section 7.3).

3. Brief Review of Photoclinometry

Photoclinometry is the recovery of surface slopes from images [Wilhelms 64] [Rindfleisch 66] [Lambiotte & Taylor 67] [Watson 68] [Lucchitta & Gambell 70] [Tyler, Simpson & Moore 71] [Rowan, McCauley & Holm 71] [Bonner & Small 73] [Wildey 75] [Squyres 81] [Howard, Blasius & Cutt 82]. Many papers and abstracts relating to this subject appear in places that may seem inaccessible to someone working in machine vision [Davis, Soderblom, & Eliason 82] [Passey & Shoemaker 82] [Davis & McEwen 84] [Davis & Soderblom 83, 84] [Malin & Danielson 84] [Wilson *et al.* 84] [McEwen 85] [Wilson *et al.* 85] (For additional references see *Shape from Shading* [Horn & Brooks 89]). Superficially, *photoclinometry* may appear to be just another name for *shape from shading*. Two different groups of researchers independently tackled the problem of recovering surface shape from spatial brightness variations in single images. Astrogeologists and workers in machine vision became aware of each other's interests only a few years ago. The underlying goals of the two groups are related, but there are some differences in approach that may be worthy of a brief discussion.

3.1 Photoclinometry versus Shape from Shading

- First, photoclinometry has focused mostly on profile methods (photoclinometrists now refer to existing shape-from-shading methods as *area-based* photoclinometry, as opposed to *profile-based*). This came about in large part because several of the surfaces of interest to the astrogeologist have reflecting properties that allow numerical integration along predetermined lines in the image, as discussed above in section 2.5 [Rindfleisch 66]. Later, a similar profile integration approach was applied to other kinds of surfaces, by using strong assumptions about local surface geometry instead. The assumption that the surface is locally cylindrical leads to such a profile integration scheme [Wildey 86], for example. More commonly, however, it has been assumed that the cross-track slope is zero, in a suitable object-centered coordinate system [Squyres 81]. This may be reasonable when one is considering a cross-section of a linearly extended feature, like a ridge, a graben, or a central section of a rotationally symmetric feature like a crater.

- The introduction of constraints that are easiest to express in an object-centered coordinate system leads away from use of a camera-centered coordinate system and to complex coordinate transformations that tend to obscure the underlying problem. A classic paper on photoclinometry [Rindfleisch 66] is difficult to read for this reason, and as a result had little impact on the field. On the other hand, it must be acknowledged that this paper dealt properly with perspective projection, which is important when the field of view is large. In all but the earliest work on shape from shading [Horn 70, 75], the assumption is made that the projection is approximately orthographic. This simplifies the equations and allows introduction of the reflectance map.
- The inherent ambiguity of the problem does not stand out as obviously when one works with profiles, as it does when one tries to fully reconstruct surfaces. This is perhaps why workers on shape from shading have been more concerned with ambiguity, and why they have emphasized the importance of *singular points* and *occluding boundaries* [Bruss 82] [Deift & Sylvester 81] [Brooks 83] [Blake, Zisserman & Knowles 85] [Saxberg 88].
- The recovery of shape is more complex than the computation of a set of profiles. Consequently much of the work in shape from shading has been restricted to simple shapes. At the same time, there has been extensive testing of shape from shading algorithms on synthetic data. This is something that is important for work on shape from shading, but makes little sense for the study of simple profile methods, except to test for errors in the procedures used for inverting the photometric function.
- Shape-from-shading methods easily deal with arbitrary collections of collimated light sources and extended sources, since these can be accommodated in the reflectance map by integrating the BRDF and the

source distribution. In astrogeology there is only one source of light (if we ignore mutual illumination or interflection between surfaces), so methods for dealing with multiple sources or extended sources were not developed.

- Calibration objects are used both in photoclinometry and shape from shading. In photoclinometry the data derived is used to fit parameters to phenomenological models such as those of Minnaert, Lommel and Seeliger, Hapke, and Lambert. In work on shape from shading the numerical data is at times used directly without further curve fitting. The parameterized models have the advantage that they permit extrapolation of observations to situations not encountered on the calibration object. This is not an issue if the calibration object contains surface elements with all possible orientations, as it will if it is smooth and convex.

- Normalization of brightness measurements is treated slightly differently too. If the imaging device is linear, one is looking for a single overall scale factor. In photoclinometry this factor is often estimated by looking for a region that is more or less flat and has known orientation in the object-centered coordinate system. In shape from shading the brightness of singular points is often used to normalize brightness measurements instead. The choice depends in part on what is known about the scene, what the shapes of the objects are (that is, are singular points or occluding boundaries imaged) and how the surface reflects light (that is, is there a unique global extremum in brightness).

- Finally, simple profiling methods usually only require continuity of the surface and existence of the first derivative (unless there is an ambiguity in the inversion of the photometric function whose resolution requires that neighboring derivatives are similar). Most shape from shading methods require continuous first derivatives and the existence of second derivatives (In some cases use is made of the equality of the second cross-derivatives taken in different order, that is, $z_{xy} = z_{yx}$). This means that these methods do not work well on scenes composed of objects that are only piecewise smooth, unless appropriately modified[19] (but see [Malik & Maydan 89]).

[19] Methods for recovering the shapes of polyhedral objects using shading on the faces and the directions of the projections of the edges into the image are discussed in [Sugihara 86] and [Horn 86].

3.2 Profiling Methods

We have seen in section 2.5 how special photometric properties sometimes allow one to calculate a profile by integration along predetermined straight lines in the image. The other approach commonly used in photoclinometry to permit simple integration is to make strong assumptions about the surface shape, most commonly that, in a suitable object-centered coordinate system, the slope of the surface is zero in a direction at right angles to the direction in which the profile is being computed. Local surface orientation has two degrees of freedom. The measured brightness provides one constraint. A second constraint is needed to obtain a solution for surface orientation. A known tangent of the surface can provide the needed information. Two common cases are treated in astrogeology:

(a) features that appear to be linearly extended (such as some ridges and grabens), in a direction presumed to be “horizontal” (that is, in the average local tangent plane);

(b) features that appear to be rotationally symmetric (like craters), with symmetry axis presumed to be “vertical” (that is, perpendicular to the average local tangent plane).

In each case, the profile is taken “across” the feature, that is, in a direction perpendicular to the intersection of the surface with the average local tangent plane. Equivalently, it is assumed that the cross-track slope is zero in the object-centered coordinate system.

One problem with this approach is that we obtain a profile in a plane containing the viewer and the light source, not a “vertical” profile, one that is perpendicular to the average local tangent plane. One way to deal with this is to iteratively adjust for the image displacement resulting from fluctuations in height on the surface, using first a scan that really is just a straight line in the image, then using the estimated profile to introduce appropriate lateral displacements into the scan line, and so on [Davis & Soderblom 84].

It turns out that the standard photoclinometric profile approach can be easily generalized to arbitrary tangent directions, ones that need not be perpendicular to the profile, and also to nonzero slopes. All that we need to assume is that the surface can locally be approximated by a (general) cylinder, that is, a surface generated by sweeping a line, the *generator*, along a curve in space. Suppose the direction of the generator is given by the vector $\mathbf{t} = (a, b, c)^T$. Note that at each point on the surface, a line parallel to the generator is tangent to the surface. Then, since the normal is perpendicular to any tangent, we have $\mathbf{t} \cdot \mathbf{n} = 0$ at every point on the surface, or just

$$a\,p + b\,q = c. \tag{24}$$

This, together with the equation $E = R(p, q)$, constitutes a pair of equations in the two unknowns p and q. There may, however, be more than one solution (or perhaps none) since one of the equations is nonlinear. Other means must be found to remove possible ambiguity arising from this circumstance. Under appropriate oblique lighting conditions, there will usually only be one solution for most observed brightness values.

From the above we conclude that we can recover surface orientation locally if we assume that the surface is cylindrical, with known direction of the generator. We can integrate out the resulting gradient in any direction we please, not necessarily across the feature. Also, the generator need not lie in the average local tangent plane; we can deal with other situations, as long as we know the direction of the generator in the camera-centered coordinate system. Further generalizations are possible, since any means of providing one more constraint on p and q will do.

In machine vision too, some workers have used strong local assumptions about the surface to allow direct recovery of surface orientation. For example, if the surface is assumed to be locally spherical, the first two partial derivatives of brightness allow one to recover the surface orientation [Pentland 84] [Lee & Rosenfeld 85]. Alternatively, one may assume that the surface is locally cylindrical [Wildey 84, 86] to resolve the ambiguity present locally in the general case.

4. Review of Shape from Shading Schemes

4.1 Characteristic Strips

The original solution of the general shape from shading problem [Horn 70, 75] uses the method of characteristic strip expansion for first order partial differential equations [Garabedian 64] [John 78]. The basic idea is quite easy to explain using the reflectance map [Horn 77, 86]. Suppose that we are at a point $(x, y, z)^T$ on the surface and we wish to extend the solution a small distance in some direction by taking a step δx in x and δy in y. We need to compute the change in height δz. This we can do if we know the components of the gradient, $p = z_x$ and $q = z_y$, because

$$\delta z = p\,\delta x + q\,\delta y. \tag{25}$$

So, as we explore the surface, we need to keep track of p and q in addition to x, y and z. This means that we also need to be able to compute the changes in p and q when we take the step. This can be done using

$$\delta p = r\,\delta x + s\,\delta y \qquad \text{and} \qquad \delta q = s\,\delta x + t\,\delta y, \tag{26}$$

where $r = z_{xx}$, $s = z_{xy} = z_{yx}$ and $t = z_{yy}$ are the second partial derivatives of the height. It seems that we need to now keep track of the second derivatives also, and in order to do that we need the third partial derivatives, and so on.

To avoid this infinite recurrence, we take another tack. Note that we have not yet used the image irradiance equation $E(x, y) = R(p, q)$. To find the brightness gradient we differentiate this equation with respect to x and y and so obtain

$$E_x = r\,R_p + s\,R_q \qquad \text{and} \qquad E_y = s\,R_p + t\,R_q. \tag{27}$$

At this point we exploit the fact that we are free to choose the direction of the step $(\delta x, \delta y)$. Suppose that we pick

$$\delta x = R_p\,\delta\xi \qquad \text{and} \qquad \delta y = R_q\,\delta\xi, \tag{28}$$

then, from equations (26) & (27) we have

$$\delta p = E_x\,\delta\xi \qquad \text{and} \qquad \delta q = E_y\,\delta\xi. \tag{29}$$

This is the whole "trick." We can summarize the above in the set of ordinary differential equations

$$\begin{gathered} \dot{x} = R_p, \qquad \dot{y} = R_q, \qquad \dot{z} = p\,R_p + q\,R_q \\ \dot{p} = E_x, \qquad \dot{q} = E_y, \end{gathered} \tag{30}$$

where the dot denotes differentiation with respect to ξ, a parameter that varies along a particular solution curve (the equations can be rescaled to make this parameter be arc length). Note that we actually have more than a mere *characteristic curve*, since we also know the orientation of the surface at all points in this curve. This is why a particular solution is called a *characteristic strip*. The projection of a characteristic curve into the image plane is called a *base characteristic* [Garabedian 64] [John 78].

The base characteristics are predetermined straight lines in the image only when the ratio $\dot{x} : \dot{y} = R_p : R_q$ is fixed, that is when the reflectance map is linear in p and q. In general, one *cannot* integrate along arbitrary curves in the image. Also, an initial curve is needed from which to sprout the characteristics strips.

It turns out that direct numerical implementations of the above equations do not yield particularly good results, since the paths of the characteristics are affected by noise in the image brightness measurements and errors tend to accumulate along their length. In particularly bad cases, the base characteristics may even cross, which does not make any sense in terms of surface shape. It is possible, however, to grow characteristic strips in parallel and use a so-called *sharpening* process to keep neighboring characteristics consistent by enforcing the conditions $z_t = p\,x_t + q\,y_t$ and $E(x,y) = R(p,q)$ along curves connecting the tips of characteristics advancing in parallel [Horn 70, 75]. This greatly improves the accuracy of the solution, since the computation of surface orientation is tied more closely to image brightness itself rather than to the brightness gradient. This also makes it possible to interpolate new characteristic strips when existing ones spread too far apart, and to remove some when they approach each other too closely.

4.2 Rotationally Symmetric Reflectance Maps

One can get some idea of how the characteristics explore a surface by considering the special case of a rotationally symmetric reflectance map, as might apply when the light source is at the viewer (or when dealing with scanning electron microscope (SEM) images). Suppose that

$$R(p,q) = f(p^2+q^2), \tag{31}$$

then

$$R_p = 2p\,f'(p^2+q^2) \qquad \text{and} \qquad R_q = 2q\,f'(p^2+q^2), \tag{32}$$

and so the directions in which the base characteristics grow are given by

$$\dot{x} = k\,p \qquad \text{and} \qquad \dot{y} = k\,q, \tag{33}$$

for some k. That is, in this case the characteristics are curves of steepest ascent or descent on the surface. The extrema of surface height are sources and sinks of characteristic curves. In this case, these are the points where the surface has maxima in brightness.

This example illustrates the importance of so-called singular points. At most image points, as we have seen, the gradient is not fully constrained by image brightness. Now suppose that $R(p,q)$ has a unique global maximum[20], that is

$$R(p,q) < R(p_0,q_0) \qquad \text{for all} \qquad (p,q) \neq (p_0,q_0). \tag{34}$$

A *singular point* (x_0,y_0) in the image is a point where

$$E(x_0,y_0) = R(p_0,q_0). \tag{35}$$

At such a point we may conclude that $(p,q) = (p_0,q_0)$. Singular points in general are sources and sinks of characteristic curves. Singular points provide strong constraint on possible solutions [Horn 70, 75] [Bruss 82] [Brooks 83] [Saxberg 88].

[20] The same argument applies when the unique extremum is a minimum, as it is in the case of scanning electron microscope (SEM) images.

The *occluding boundary* is the set of points where the local tangent plane contains the direction toward the viewer. It has been suggested that occluding boundaries provide strong constraint on possible solutions [Ikeuchi & Horn 81] [Bruss 82]. As a consequence there has been interest in representations for surface orientation that behave well near the occluding boundary, unlike the gradient, which becomes infinite [Ikeuchi & Horn 81] [Horn & Brooks 86]. Recently there has been some question as to how much constraint occluding boundaries really provide, given that singular points appear to already strongly constrain the solution [Brooks 83] [Saxberg 88].

4.3 Existence and Uniqueness

Questions of existence and uniqueness of solutions of the shape-from-shading problem have still not been resolved entirely satisfactorily. With an initial curve, however, the method of characteristic strips does yield a unique solution, assuming only continuity of the first derivatives of surface height (see Haar's theorem on pg. 145 in [Courant & Hilbert 62] or [Bruss 82]). The question of uniqueness is more difficult to answer when an initial curve is not available. One problem is that it is hard to say anything completely general that will apply to all possible reflectance maps. More can be said when specific reflectance maps are chosen, such as ones that are linear in the gradient [Rindfleisch 66] or those that are rotationally symmetric [Bruss 82].

It has recently been shown that there exist *impossible shaded images*, that is, images that do not correspond to any surface illuminated in the specified way [Horn, Szeliski & Yuille 89]. It may turn out that almost all images with multiple singular points are impossible in this sense [Saxberg 88]. This is an important issue, because it may help explain how our visual system sometimes determines that the surface being viewed cannot possibly be uniform in its reflecting properties. One can easily come up with smoothly shaded images, for example, that do not yield an impression of shape, instead appearing as flat surfaces with spatially varying reflectance or surface “albedo.” (See also Figure 10 in section 7.2).

4.4 Variational Formulations

As discussed above in section 2.4, in the case of a surface with constant albedo, when both the observer and the light sources are far away, surface radiance depends only on surface orientation and not on position in space and the image projection can be considered to be orthographic[21]. In this case the image irradiance equation becomes just

$$E(x,y) = R\big(p(x,y), q(x,y)\big), \tag{36}$$

where $E(x,y)$ is the image irradiance at the point (x,y), while $R(p,q)$, the *reflectance map*, is the (normalized) scene radiance of a surface patch with orientation specified by the partial derivatives

$$p = \frac{\partial z}{\partial x} \quad \text{and} \quad q = \frac{\partial z}{\partial y}, \tag{37}$$

of surface height $z(x,y)$ above some reference plane perpendicular to the optical axis.

The task is to find $z(x,y)$ given the image $E(x,y)$ and the reflectance map $R(p,q)$. Additional constraints, such as boundary conditions and singular points, are needed to ensure that there is a unique solution [Bruss 82] [Deift & Sylvester 81] [Blake, Zisserman & Knowles 85] [Saxberg 88]. If we ignore *integrability*[22], some versions of the problem of shape from shading may be considered to be ill-posed[23], that is, there is not a unique solution $\{p(x,y), q(x,y)\}$ that minimizes the brightness error

$$\iint \big(E(x,y) - R(p,q)\big)^2 \, dx\, dy. \tag{38}$$

In fact the error can be made equal to zero for an infinite number of choices for $\{p(x,y), q(x,y)\}$. We can pick out one of these solutions by finding the one that minimizes some functional such as a measure of "departure from smoothness"

$$\iint (p_x^2 + p_y^2 + q_x^2 + q_y^2) \, dx\, dy, \tag{39}$$

while satisfying the constraint $E(x,y) = R(p,q)$. Introducing a Lagrange multiplier $\lambda(x,y)$ to enforce the constraint, we find that we have to minimize

$$\iint \big((p_x^2 + p_y^2 + q_x^2 + q_y^2) + \lambda(x,y)\,(E - R)\big) \, dx\, dy. \tag{40}$$

[21] The shape-from-shading problem can be formulated and solved when the viewer and the light sources are not at a great distance [Rindfleisch 66] [Horn 70, 75], but then scene radiance depends on position as well as surface orientation, and the notion of a reflectance map is not directly applicable.

[22] A gradient-field (or needle diagram) $\{p(x,y), q(x,y)\}$ is integrable if there exists some surface height function $z(x,y)$ such that $p(x,y) = z_x(x,y)$ and $q(x,y) = z_y(x,y)$, where the subscripts denote partial derivatives.

[23] Generally, a small patch of a shaded image is infinitely ambiguous. Also, without integrability, the problem of recovering a gradient field is generally ill-posed. But if we impose integrability, and provide suitable boundary conditions, then the shape-from-shading problem is definitely *not* ill-posed [Bruss 82] [Deift & Sylvester 81] [Brooks 83] [Blake, Zisserman & Knowles 85] [Saxberg 88].

The Euler equations are

$$\Delta p + \lambda(x,y)\, R_p = 0 \qquad \text{and} \qquad \Delta q + \lambda(x,y)\, R_q = 0. \tag{41}$$

After elimination of the Lagrange multiplier $\lambda(x,y)$, we are left with the pair of equations

$$R_q\,\Delta p = R_p\,\Delta q \qquad \text{and} \qquad E(x,y) = R(p,q). \tag{42}$$

Unfortunately, no convergent iterative scheme has been found for this constrained variational problem [Horn & Brooks 86] (compare [Wildey 75]).

We can approach this problem in a quite different way using the "departure from smoothness" measure in a penalty term [Ikeuchi & Horn 81], looking instead for a minimum of[24]

$$\iint \Big(\big(E(x,y) - R(p,q)\big)^2 + \lambda\,(p_x^2 + p_y^2 + q_x^2 + q_y^2) \Big)\, dx\, dy. \tag{43}$$

It should be pointed out that a solution of this "regularized" problem is *not* a solution of the original problem, although it may be close to some solution of the original problem [Brooks 85]. In any case, this variational problem leads to the following coupled pair of second-order partial differential equations:

$$\begin{aligned} \lambda\,\Delta p &= -\big(E(x,y) - R(p,q)\big)\, R_p(p,q) \\ \lambda\,\Delta q &= -\big(E(x,y) - R(p,q)\big)\, R_q(p,q) \end{aligned} \tag{44}$$

Using a discrete approximation of the Laplacian operator[25]

$$\{\Delta f\}_{kl} \approx \frac{\kappa}{\epsilon^2}(\overline{f}_{kl} - f_{kl}), \tag{45}$$

where $\overline{f}$ is a local average[26] of f, and ϵ is the spacing between picture cells, we arrive at the set of equations

$$\begin{aligned} \kappa\lambda'\, p_{kl} &= \kappa\lambda'\, \overline{p}_{kl} + \big(E(x,y) - R(p,q)\big)R_p(p,q) \\ \kappa\lambda'\, q_{kl} &= \kappa\lambda'\, \overline{q}_{kl} + \big(E(x,y) - R(p,q)\big)R_q(p,q) \end{aligned} \tag{46}$$

[24] Note that λ here is *not* a Lagrange multiplier, but a factor that balances the relative contributions of the brightness error term and the term measuring departure from smoothness. That is, there is no absolute *constraint* imposed here, only a penalty term added that increases with departure from smoothness.

[25] There are several methods for approximating the Laplacian operator, including five-point and nine-point approximations. It is well known that, while the nine-point approximation involves more computation, its lowest-order error term has a higher order than that of the five-point approximation [Horn 86].

[26] Here $\kappa = 4$ when the local average $\overline{f}_{kl}$ is computed using the four edge-adjacent neighbors, while $\kappa = 10/3$, when $1/5$ of the average of the corner-adjacent neighbors is added to $4/5$ of the average of the edge-adjacent neighbors (see also section 6.2).

where $\lambda' = \lambda/\epsilon^2$. This immediately suggests the iterative scheme

$$\begin{aligned} p_{kl}^{(n+1)} &= \overline{p}_{kl}^{(n)} + \frac{1}{\kappa\lambda'}\Big(E(x,y) - R(p^{(n)},q^{(n)})\Big)R_p(p^{(n)},q^{(n)}) \\ q_{kl}^{(n+1)} &= \overline{q}_{kl}^{(n)} + \frac{1}{\kappa\lambda'}\Big(E(x,y) - R(p^{(n)},q^{(n)})\Big)R_q(p^{(n)},q^{(n)}) \end{aligned} \tag{47}$$

where the superscript denotes the iteration number[27].

From the above it may appear that $R(p,q)$, $R_p(p,q)$, and $R_q(p,q)$ should be evaluated using the "old" values of p and q. It turns out that the numerical stability of the scheme is somewhat enhanced if they are evaluated instead at the local average values, $\overline{p}$ and $\overline{q}$ [Ikeuchi & Horn 81].

One might hope that the correct solution of the original shape-from-shading problem provides a fixed point for the iterative scheme. This is not too likely, however, since we are solving a modified problem that includes a penalty term. Consequently, an interesting question one might ask about an algorithm such as this, is whether it will "walk away" from the correct solution of the original image irradiance equation $E(x,y) = R(p,q)$ when this solution is provided as an initial condition [Brooks 85]. The algorithm described here does just that, since it can trade off a small amount of brightness error against an increase in surface smoothness. At the solution, we have $E(x,y) = R(p,q)$, so that the right hand sides of the two coupled partial differential equations (equations (44)) are zero. This implies that if the solution of the modified problem is to be equal the solution of the original problem then the Laplacians of p and q must be equal to zero. This is the case for very few surfaces, just those for which

$$\Delta z(x,y) = k, \tag{48}$$

for some constant k. While this includes all harmonic functions, it excludes most real surfaces, for which adjustments away from the correct shape are needed to assure equality of the left and right hand sides of equations (44) describing the solution of the modified problem. In general, this approach produces solutions that are too smooth, with the amount of distortion depending on the choice of the parameter λ. For related reasons, this algorithm does well only on simple smooth shapes, and does not perform well on complex, wrinkled surfaces.

4.5 Recovering Height from Gradient

In any case, we are also still faced with the problem of dealing with the lack of *integrability*, that is the lack of a surface $z(x,y)$ such that $p(x,y) =$

[27] These equations are solved iteratively because the system of equations is so large and because of the fact that the reflectance map $R(p,q)$ is typically nonlinear.

$z_x(x,y)$ and $q(x,y) = z_y(x,y)$[28]. At the very least, we should try to find the surface $z(x,y)$ that has partial derivatives z_x and z_y that come closest to matching the computed $p(x,y)$ and $q(x,y)$, by minimizing

$$\iint \left((z_x - p)^2 + (z_y - q)^2\right) dx\, dy. \tag{49}$$

This leads to the Poisson equation

$$\Delta z = p_x + q_y. \tag{50}$$

Using the discrete approximation of the Laplacian given above (equation (45)) yields

$$\frac{\kappa}{\epsilon^2} z_{kl} = \frac{\kappa}{\epsilon^2} \overline{z}_{kl} - (p_x + q_y), \tag{51}$$

a set of equations that suggests the following iterative scheme:

$$z_{kl}^{(n+1)} = \overline{z}_{kl}^{(n)} - \frac{\epsilon^2}{\kappa} \left(\{p_x\}_{kl}^{(n)} + \{q_y\}_{kl}^{(n)} \right), \tag{52}$$

where the terms in braces are numerical estimates of the indicated derivatives at the picture cell (k,l).

The so-called *natural boundary conditions*[29] here are just

$$c\, z_x + s\, z_y = c\, p + s\, q, \tag{53}$$

where (c,s) is a normal to the boundary.

Another way of dealing with the integrability issue is to try and directly minimize

$$\iint \left((E(x,y) - R(p,q))^2 + \lambda\, (p_y - q_x)^2 \right) dx\, dy. \tag{54}$$

This leads to the coupled partial differential equations [Horn & Brooks 86]

$$\begin{aligned} \lambda\, (p_{yy} - q_{xy}) &= -(E(x,y) - R(p,q)) R_p, \\ \lambda\, (q_{xx} - p_{yx}) &= -(E(x,y) - R(p,q)) R_q. \end{aligned} \tag{55}$$

This set of equations can also be discretized by introducing appropriate finite difference approximations for the second partial derivatives p_{yy}, q_{xx} and the cross derivatives of p and q. An iterative scheme is suggested once one isolates the center terms of the discrete approximations of p_{yy} and q_{xx}. This is very similar to the method developed by Strat, although he arrived at his scheme directly in the discrete domain [Strat 79]. His iterative scheme avoids the excessive smoothing of the one described early, but appears to be less stable, in the sense that it diverges under a wider set of circumstances.

[28] The resulting gradient field is likely not to be integrable because we have not enforced the condition $p_y = q_x$, which corresponds to $z_{xy} = z_{yx}$.

[29] Natural boundary conditions arise in variational problems where no boundary conditions are explicitly imposed [Courant & Hilbert 53].

5. New Coupled Height and Gradient Scheme

The new shape-from-shading scheme will be presented through a series of increasingly more robust variational methods. We start with the simplest, which grows naturally out of what was discussed in the previous section.

5.1 Fusing Height and Gradient Recovery

One way of fusing the recovery of gradient from shading with the recovery of height from gradient, is to represent both gradient (p, q) and height z in one variational scheme and to minimize the functional

$$\iint \Big((E(x,y) - R(p,q))^2 + \mu((z_x - p)^2 + (z_y - q)^2) \Big)\, dx\, dy. \tag{56}$$

Note that, as far as $p(x,y)$ and $q(x,y)$ are concerned, this is an ordinary calculus problem (since no partial derivatives of p and q appear in the integrand). Differentiating the integrand with respect to $p(x,y)$ and $q(x,y)$ and setting the result equal to zero leads to

$$\begin{aligned} p &= z_x + \frac{1}{\mu}(E - R)R_p, \\ q &= z_y + \frac{1}{\mu}(E - R)R_q. \end{aligned} \tag{57}$$

Now $z(x,y)$ does not occur directly in $(E(x,y) - R(p,q))$ so we actually just need to minimize

$$\iint ((z_x - p)^2 + (z_y - q)^2)\, dx\, dy, \tag{58}$$

and we know from the previous section that the Euler equation for this variational problem is just

$$\Delta z = p_x + q_y. \tag{59}$$

We now have one equation for each of p, q and z.

These three equations are clearly satisfied when $p = z_x$, $q = z_y$ and $E = R$. That is, if a solution of the original shape-from-shading problem exists, then it satisfies this system of equations exactly (which is more than can be said for some other systems of equations obtained using a variational approach, as pointed out in section 4.4). It is instructive to substitute the expressions obtained for p and q in $p_x + q_y$:

$$\begin{aligned} p_x + q_y = z_{xx} + z_{yy} + \frac{1}{\mu}\Big(&(E - R)(R_{pp}p_x + R_{pq}(p_y + q_x) + R_{qq}q_y) \\ &- (R_p^2 p_x + R_pR_q(p_y + q_x) + R_q^2 q_y) + (E_xR_p + E_yR_q)\Big). \end{aligned} \tag{60}$$

Since $\Delta z = (p_x + q_y)$, we note that the three equations above are satisfied when

$$(R_p^2 p_x + R_p R_q (p_y + q_x) + R_q^2 q_y) - (E_x R_p + E_y R_q) \tag{61}$$
$$= (E - R)(R_{pp} p_x + R_{pq}(p_y + q_x) + R_{qq} q_y).$$

This is exactly the equation obtained at the end of section 4.2 in [Horn & Brooks 86], where an attempt was made to directly impose integrability using the constraint $p_y = q_x$. It was stated there that no convergent iterative scheme had been found for solving this complicated nonlinear partial differential equation directly. The method presented in this section provides an indirect way of solving this equation.

Note that the natural boundary conditions for z are once again

$$c\, z_x + s\, z_y = c\, p + s\, q, \tag{62}$$

where (c, s) is a normal to the boundary.

The coupled system of equations above for p, q (equation (57)) and z (equation (59)) immediately suggests an iterative scheme

$$\begin{aligned}
p_{kl}^{(n+1)} &= \{z_x\}_{kl}^{(n)} + \frac{1}{\mu}(E - R)R_p, \\
q_{kl}^{(n+1)} &= \{z_y\}_{kl}^{(n)} + \frac{1}{\mu}(E - R)R_q, \\
z_{kl}^{(n+1)} &= \overline{z}_{kl}^{(n)} - \frac{\epsilon^2}{\kappa}\left(\{p_x\}_{kl}^{(n+1)} + \{q_y\}_{kl}^{(n+1)}\right),
\end{aligned} \tag{63}$$

where we have used the discrete approximation of the Laplacian for z introduced in equation (45). This new iterative scheme works well when the initial values given for p, q and z are close to the solution. It will converge to the exact solution if it exists; that is, if there exist a discrete set of values $\{z_{kl}\}$ such that $\{p_{kl}\}$ and $\{q_{kl}\}$ are the discrete estimate of the first partial derivatives of z with respect to x and y respectively and $E_{kl} = R(p_{kl}, q_{kl})$.

In this case the functional we wish to minimize can actually be reduced to zero. It should be apparent that for this to happen, the discrete estimator used for the Laplacian must match the sum of the convolution of the discrete estimator of the x derivative with itself and the convolution of the discrete estimator of the y derivative with itself. (This and related matters are taken up again in section 6.2.)

The algorithm can easily be tested using synthetic height data z_{kl}. One merely estimates the partial derivatives using suitable discrete difference formulae, and then uses the resulting values p_{kl} and q_{kl} to compute the synthetic image $E_{kl} = R(p_{kl}, q_{kl})$. This construction guarantees that there will be an exact solution. If a real image is used, there is no guarantee that there is an exact solution, and the algorithm can at best find a good discrete approximation of the solution of the underlying continuous problem. In this case the functional will in fact not be reduced exactly to zero. In some cases the residue may be quite large. This may be the result of aliasing introduced when sampling the image, as discussed in section 6.5, or because in fact the image given could not have arisen from shading on a homogeneous surface with the reflectance properties and lighting as encoded in the reflectance map—that is, it is an *impossible shaded image* [Horn, Szeliski & Yuille 89].

The iterative algorithm described in this section, while simple, is not very stable, and has a tendency to get stuck in local minima, unless one is close to the exact solution, particularly when the surface is complex and the reflectance map is not close to linear in the gradient. It has been found that the performance of this algorithm can be improved greatly by linearizing the reflectance map. It can also be stabilized by adding a penalty term for departure from smoothness. This allows one to come close to the correct solution, at which point the penalty term is removed in order to prevent it from distorting the solution. We first treat the linearization of the reflectance map.

5.2 Linearization of Reflectance Map

We can develop a better scheme than the one described in the previous section, while preserving the apparent linearity of the equations, by approximating the reflectance map $R(p,q)$ locally by a linear function of p and q. There are several options for choice of reference gradient for the series expansion, so let us keep it general for now at (p_0, q_0)[30]. We have

$$R(p,q) \approx R(p_0,q_0) + (p-p_0)\,R_p(p_0,q_0) + (q-q_0)\,R_q(p_0,q_0) + \cdots \tag{64}$$

Again, gathering all of the terms in p_{kl} and q_{kl} on the left hand sides of the equations, we now obtain

$$\begin{aligned}(\mu + R_p^2)\,p_{kl} + R_pR_q\,q_{kl} &= \mu z_x + (E - R + p_0R_p + q_0R_q)R_p,\\ R_qR_p\,p_{kl} + (\mu + R_q^2)\,q_{kl} &= \mu z_y + (E - R + p_0R_p + q_0R_q)R_q,\end{aligned} \tag{65}$$

while the equation for z remains unchanged. (Note that now R, R_p and R_q denote quantities evaluated at the reference gradient (p_0, q_0)).

It is convenient to rewrite these equations in terms of quantities rel-

[30] The reference gradient will, of course, be different at every picture cell, but to avoid having subscripts on the subscripts, we will simple denote the reference gradient at a particular picture cell by (p_0, q_0).

ative to the reference gradient (p_0, q_0). Let

$$\begin{aligned} \delta p_{kl} &= p_{kl} - p_0 \qquad \text{and} \qquad \delta q_{kl} = q_{kl} - q_0, \\ \delta z_x &= z_x - p_0 \qquad \text{and} \qquad \delta z_y = z_y - q_0. \end{aligned} \tag{66}$$

This leads to

$$\begin{aligned} (\mu + R_p^2)\,\delta p_{kl} + R_p R_q\,\delta q_{kl} &= \mu\,\delta z_x + (E - R)R_p, \\ R_p R_q\,\delta p_{kl} + (\mu + R_q^2)\,\delta q_{kl} &= \mu\,\delta z_y + (E - R)R_q. \end{aligned} \tag{67}$$

(The equations clearly simplify somewhat if we choose (z_x, z_y) for the reference gradient (p_0, q_0).) We can view the above as a pair of linear equations for δp_{kl} and δq_{kl}. The determinant of the 2×2 coefficient matrix,

$$D = \mu(\mu + R_p^2 + R_q^2), \tag{68}$$

is always positive, so there is no problem with singularities. The solution is given by

$$\begin{aligned} D\,\delta p_{kl} &= (\mu + R_q^2)\,A - R_p R_q\,B, \\ D\,\delta q_{kl} &= (\mu + R_p^2)\,B - R_q R_p\,A, \end{aligned} \tag{69}$$

where

$$\begin{aligned} A &= \mu\,\delta z_x + (E - R)R_p, \\ B &= \mu\,\delta z_y + (E - R)R_q. \end{aligned} \tag{70}$$

This leads to a convenient iterative scheme where the new values are given by

$$p_{kl}^{(n+1)} = p_0^{(n)} + \delta p_{kl}^{(n)} \quad \text{and} \quad q_{kl}^{(n+1)} = q_0^{(n)} + \delta q_{kl}^{(n)}, \tag{71}$$

in terms of the old reference gradient and the increments computed above. The new version of the iterative scheme does not require a great deal more computation than the simpler scheme introduced in section 4.5, since the partial derivatives R_p and R_q are already needed there.

5.3 Incorporating Departure from Smoothness Term

We now introduce a penalty term for departure from smoothness, effectively combining the iterative method of [Ikeuchi & Horn 81] for recovering p and q from $E(x,y)$ and $R(p,q)$, with the scheme for recovering z given p and q discussed in section 4.5. (For the moment we do not linearize the reflectance map; this will be addressed in section 5.6). We look directly for a minimum of

$$\begin{aligned} \iint \Big((E(x,y) - R(p,q))^2 \\ + \lambda\,(p_x^2 + p_y^2 + q_x^2 + q_y^2) \\ + \mu\big((z_x - p)^2 + (z_y - q)^2\big)\Big)\,dx\,dy. \end{aligned} \tag{72}$$

The Euler equations of this calculus of variations problem lead to the following coupled system of second-order partial differential equations:

$$\begin{aligned}\lambda\,\Delta p &= -(E-R)R_p - \mu(z_x - p),\\ \lambda\,\Delta q &= -(E-R)R_q - \mu(z_y - q),\\ \Delta z &= p_x + q_y.\end{aligned} \tag{73}$$

A discrete approximation of these equations can be obtained using the discrete approximation of the Laplacian operator introduced in equation (45):

$$\begin{aligned}\frac{\kappa\lambda}{\epsilon^2}\,(\overline{p}_{kl} - p_{kl}) &= -(E-R)R_p - \mu(z_x - p_{kl}),\\ \frac{\kappa\lambda}{\epsilon^2}\,(\overline{q}_{kl} - q_{kl}) &= -(E-R)R_q - \mu(z_y - q_{kl}),\\ \frac{\kappa}{\epsilon^2}\,(\overline{z}_{kl} - z_{kl}) &= p_x + q_y.\end{aligned} \tag{74}$$

where E, R, R_p, and R_q are the corresponding values at the picture cell (k,l), while z_x, z_y, p_x and q_y are discrete estimates of the partial derivative of z, p and q there. We can collect the terms in p_{kl}, q_{kl} and z_{kl} on one side to obtain

$$\begin{aligned}(\kappa\lambda' + \mu)\,p_{kl} &= \left(\kappa\lambda'\,\overline{p}_{kl} + \mu z_x\right) + (E-R)R_p,\\ (\kappa\lambda' + \mu)\,q_{kl} &= \left(\kappa\lambda'\,\overline{q}_{kl} + \mu z_y\right) + (E-R)R_q,\\ \frac{\kappa}{\epsilon^2}\,z_{kl} &= \frac{\kappa}{\epsilon^2}\,\overline{z}_{kl} - (p_x + q_y),\end{aligned} \tag{75}$$

where $\lambda' = \lambda/\epsilon^2$. These equations immediately suggest an iterative scheme, where the right hand sides are computed using the current values of the z_{kl}, p_{kl}, and q_{kl}, with the results then used to supply new values for the unknowns appearing on the left hand sides

From the above it may appear that $R(p,q)$, $R_p(p,q)$, and $R_q(p,q)$ should be evaluated using the "old" values of p and q. One might, on the other hand, argue that the local average values $\overline{p}$ and $\overline{q}$, or perhaps even the gradient estimates z_x and z_y, are more appropriate. Experimentation suggests that the scheme is most stable when the local averages $\overline{p}$ and $\overline{q}$ are used.

The above scheme contains a penalty term for departure from smoothness, so it may appear that it cannot to converge to the exact solution. Indeed, it appears as if the iterative scheme will "walk away" from the correct solution when it is presented with it as initial conditions, as discussed in section 4.4. It turns out, however, that the penalty term is needed only to prevent instability when far from the solution. When we come close to the solution, λ' can be reduced to zero, and so the penalty term drops out. It is tempting to leave the penalty term out right from the start, since this simplifies the equations a great deal, as shown in section 5.1. The contribution from the penalty term does, however, help damp out instabilities when far from the solution and so should be included. This is particularly important with real data, where one cannot expect to find an exact solution.

Note, by the way, that the coupled second order partial differential equations above (equation (76)) are eminently suited for solution by coupled resistive grids [Horn 88].

5.4 Relationship to Existing Techniques

- Recently new methods have been developed that combine the iterative scheme discussed in section 4.4 for recovering surface orientation from shading with a projection onto the subspace of integrable gradients [Frankot & Chellappa 88] [Shao, Simchony & Chellappa 88]. The approach there is to alternately take one step of the iterative scheme [Ikeuchi & Horn 81] and to find the "nearest" integrable gradient. This gradient is then provided as initial conditions for the next step of the iterative scheme, ensuring that the gradient field never departs far from integrability. The integrable gradient closest to a given gradient field can be found using orthonormal series expansion and the fact that differentiation in the spatial domain corresponds to multiplication by frequency in the transform domain [Frankot & Chellappa 88].
- Similar results can be obtained by using instead the method described in section 4.5 for recovering the height $z(x, y)$ that best matches a given gradient. The resulting surface can then be (numerically) differentiated to obtain initial values for $p(x, y)$ and $q(x, y)$ for the next step of the iterative scheme [Shao, Simchony & Chellappa 88].
- Next, note that we obtain the scheme of [Ikeuchi & Horn 81] (who ignored the integrability problem) discussed in section 4.4, if we drop the departure from integrability term in the integrand—that is, when $\mu = 0$. If we instead remove the departure from smoothness term in the integrand—that is, when $\lambda = 0$—we obtain something reminiscent of the iterative scheme of [Strat 79], although Strat dealt with the integrability issue in a different way.
- Finally, if we drop the brightness error term in the integrand, we obtain the scheme of [Harris 86, 87] for interpolating from depth and slope. He minimizes

$$\iint \Big(\lambda\,(p_x^2 + p_y^2 + q_x^2 + q_y^2) + ((z_x - p)^2 + (z_y - q)^2)\Big)\, dx\, dy. \quad (76)$$

> and arrives at the Euler equations
>
> $$\begin{aligned}\lambda\,\Delta p &= -(z_x - p),\\ \lambda\,\Delta q &= -(z_y - q),\\ \Delta z &= p_x + q_y.\end{aligned} \tag{77}$$
>
> Now consider that
>
> $$\Delta(\Delta z) = \Delta(p_x + q_y). \tag{78}$$
>
> Since application of the Laplacian operator and differentiation commute we have
>
> $$\Delta(\Delta z) = (\Delta p)_x + (\Delta q)_y, \tag{79}$$
>
> or
>
> $$\lambda\,\Delta(\Delta z) = -(z_{xx} - p_x) - (z_{yy} - q_y), \tag{80}$$
>
> and so
>
> $$\lambda\,\Delta(\Delta z) = -\Delta z + (p_x + q_y) = 0. \tag{81}$$
>
> So his method solves the biharmonic equation for z, by solving a coupled set of second-order partial differential equations. It does it in an elegant, stable way that permits introduction of constraints on both height z and gradient (p, q). This is a good method for interpolating from sparse depth and surface orientation data.

The biharmonic equation has been employed to interpolate digital terrain models (DTMs) from contour maps. Such DTMs were used, for example, in [Horn 81] [Sjoberg & Horn 83]. The obvious implementations of finite difference approximations of the biharmonic operator, however, tend to be unstable because some of the weights are negative, and because the corresponding coefficient matrix lacks diagonal dominance. Also, the treatment of boundary conditions is complicated by the fact that the support of the biharmonic operator is so large. The scheme described above circumvents both of these difficulties—it was used to interpolate the digital terrain model[31] used for the example illustrated by Figure 1.

5.5 Boundary Conditions & Nonlinearity of Reflectance Map

So far we have assumed that suitable boundary conditions are available, that is, the gradient is known on the boundary of the image region to which the computation is to be applied. If this is not the case, the solution is likely not to be unique. We may nevertheless try to find some solution by

[31] The new shape-from-shading algorithm, of course, works equally well on synthetic shaded images of digital terrain models obtained by other means, such as one of the Les Diablerets region of Switzerland used in [Horn & Bachman 78].

imposing so-called *natural* boundary conditions [Courant & Hilbert 53]. The natural boundary conditions for the variational problem described here can be shown to be

$$c\,p_x + s\,p_y = 0 \quad \text{and} \quad c\,q_x + s\,q_y = 0 \tag{82}$$

and

$$c\,z_x + s\,z_y = c\,p + s\,q \tag{83}$$

where (c, s) is a normal to the boundary. That is, the normal derivative of the gradient is zero and the normal derivative of the height has to match the slope in the normal direction computed from the gradient.

In the above we have approximated the original partial differential equations by a set of discrete equations, three for every picture cell (one each for p, q and z). If these equations were linear, we could directly apply all the existing theory relating to convergence of various iterative schemes and how one solves such equations efficiently, given that the corresponding coefficient matrices are sparse[32]. Unfortunately, the equations are in general not linear, because of the nonlinear dependence of the reflectance map $R(p, q)$ on the gradient. In fact, in deriving the above simple iterative scheme, we have treated $R(p, q)$, and its derivatives, as constant (independent of p and q) during any particular iterative adjustment of p and q.

5.6 Local Linear Approximation of Reflectance Map

In section 5.2 we linearized the reflectance map in order to counteract the tendency of the simple iterative scheme developed in section 5.1 to get stuck in local minima. We now do the same for the more complex scheme described in section 5.3. We again use

$$R(p,q) \approx R(p_0,q_0) + (p - p_0)\,R_p(p_0,q_0) + (q - q_0)\,R_q(p_0,q_0) + \cdots \tag{84}$$

Gathering all of the terms in p_{kl} and q_{kl} on the left hand sides of the equations, we now obtain

$$\begin{aligned}(\lambda'' + R_p^2)\,p_{kl} + R_pR_q\,q_{kl} \\ &= (\kappa\lambda'\overline{p}_{kl} + \mu z_x) + (E - R + p_0R_p + q_0R_q)R_p,\\ R_qR_p\,p_{kl} + (\lambda'' + R_q^2)\,q_{kl} \\ &= \left(\kappa\lambda'\overline{q}_{kl} + \mu z_y\right) + (E - R + p_0R_p + q_0R_q)R_q,\end{aligned} \tag{85}$$

while the equation for z remains unchanged. (Note that here R, R_p and R_q again denote quantities evaluated at the reference gradient (p_0, q_0)). In the above we have abbreviated $\lambda'' = \kappa\lambda' + \mu$.

[32] See [Lee 88] for a proof of convergence of an iterative shape-from-shading scheme.

It is convenient to rewrite these equations in terms of quantities defined relative to the reference gradient:

$$\begin{aligned} \delta p_{kl} &= p_{kl} - p_0 \quad &&\text{and} \quad &\delta q_{kl} &= q_{kl} - q_0 \\ \delta \overline{p}_{kl} &= \overline{p}_{kl} - p_0 \quad &&\text{and} \quad &\delta \overline{q}_{kl} &= \overline{q}_{kl} - q_0 \\ \delta z_x &= z_x - p_0 \quad &&\text{and} \quad &\delta z_y &= z_y - q_0 \end{aligned} \tag{86}$$

This yields

$$\begin{aligned} (\lambda'' + R_p^2)\, \delta p_{kl} + R_p R_q\, \delta q_{kl} &= \kappa\lambda'\, \delta\overline{p}_{kl} + \mu\, \delta z_x + (E - R) R_p, \\ R_p R_q\, \delta q_{kl} + (\lambda'' + R_q^2)\, \delta q_{kl} &= \kappa\lambda'\, \delta\overline{q}_{kl} + \mu\, \delta z_y + (E - R) R_q. \end{aligned} \tag{87}$$

(The equations clearly simplify somewhat if we choose either $\overline{p}$ and $\overline{q}$ or z_x and z_y for the reference gradient p_0 and q_0.) We can view the above as a pair of linear equations for δp_{kl} and δq_{kl}. The determinant of the 2×2 coefficient matrix

$$D = \lambda''(\lambda'' + R_p^2 + R_q^2) \tag{88}$$

is always positive, so there is no problem with singularities. The solution is given by

$$\begin{aligned} D\, \delta p_{kl} &= (\lambda'' + R_q^2)\, A - R_p R_q\, B, \\ D\, \delta q_{kl} &= (\lambda'' + R_p^2)\, B - R_q R_p\, A, \end{aligned} \tag{89}$$

where

$$\begin{aligned} A &= \kappa\lambda'\, \delta\overline{p}_{kl} + \mu\, \delta z_x + (E - R) R_p, \\ B &= \kappa\lambda'\, \delta\overline{q}_{kl} + \mu\, \delta z_y + (E - R) R_q. \end{aligned} \tag{90}$$

This leads to a convenient iterative scheme where the new values are given by

$$p_{kl}^{(n+1)} = p_0^{(n)} + \delta p_{kl}^{(n)} \quad \text{and} \quad q_{kl}^{(n+1)} = q_0^{(n)} + \delta q_{kl}^{(n)}, \tag{91}$$

in terms of the old reference gradient and the increments computed above. It has been determined empirically that this scheme converges under a far wider set of circumstances than the one presented in the previous section.

Experimentation with different reference gradients, including the old values of p and q, the local average $\overline{p}$ and $\overline{q}$, as well as z_x and z_y showed that the accuracy of the solution and the convergence is affected by this choice. It became apparent that if we do not want the scheme to "walk away" from the correct solution, then we should use the old value of p and q for the reference p_0 and q_0.

6. Some Implementation Details

6.1 Derivative Estimators and Staggered Grids

In one dimension, it is well-known from numerical analysis that the best finite difference estimators of even derivatives have odd support, while the best estimators of odd derivatives have even support. These estimators are “best” in the sense that their lowest-order error terms have a small coefficient and that they do not attenuate the higher frequencies as much as the alternative ones. A good estimator of the second derivative of z, for example, is

$$\{z_{xx}\}_k \approx \frac{1}{\epsilon^2}(z_{k-1} - 2z_k + z_{k+1}), \tag{92}$$

while a good estimator of the first derivative of z is just

$$\{z_x\}_k \approx \frac{1}{\epsilon}(z_{k+1} - z_k). \tag{93}$$

Note that the latter, like other estimators with even support for odd derivatives, gives an estimate valid at the point midway between samples.

This suggests that one should use staggered grids. That is, the arrays containing sampled values of p and q (and hence image brightness E) should be offset by $1/2$ picture cells in both x and y from those for z (see Figure 3). This also means that if the image is rectangular and contains $n \times m$ picture cells, then the array of heights should be of size $(n+1) \times (m+1)$. Appropriate two-dimensional estimators for the first partial derivatives of z then are (see also [Horn & Schunck 81]):

$$\begin{aligned} \{z_x\}_{k,l} &\approx \frac{1}{2\epsilon}(z_{k,l+1} - z_{k,l} + z_{k+1,l+1} - z_{k+1,l}) \\ \{z_y\}_{k,l} &\approx \frac{1}{2\epsilon}(z_{k+1,l} - z_{k,l} + z_{k+1,l+1} - z_{k,l+1}) \end{aligned} \tag{94}$$

These can be conveniently shown graphically in the form of the stencils:

$$\frac{1}{2\epsilon}\begin{array}{|c|c|}\hline -1 & +1 \\ \hline -1 & +1 \\ \hline \end{array} \quad \text{and} \quad \frac{1}{2\epsilon}\begin{array}{|c|c|}\hline +1 & +1 \\ \hline -1 & -1 \\ \hline \end{array}$$

The results obtained apply to the point $(k + 1/2, l + 1/2)$ in the grid of discrete values of z; or the point (k, l) in the (offset) discrete grid of values of p and q. Similar schemes can be developed for the first partial derivatives of p and q needed in the algorithms introduced here, with the offsets now acting in the opposite direction.

6.2 Discrete Estimators of the Laplacian

We also need to obtain local averages based on discrete approximations of the Laplacian operators. We could simply use one of the stencils

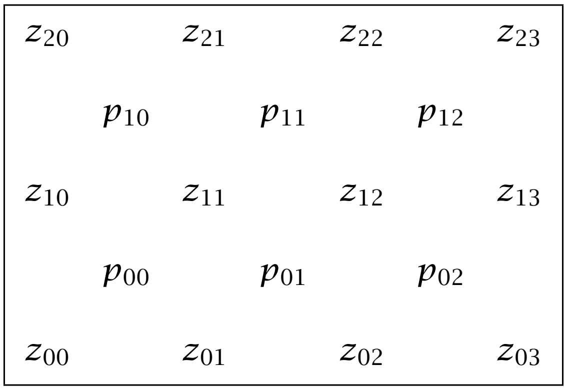
z_{20} z_{21} z_{22} z_{23}
p_{10} p_{11} p_{12}
z_{10} z_{11} z_{12} z_{13}
p_{00} p_{01} p_{02}
z_{00} z_{01} z_{02} z_{03}

Figure 3. It is convenient to have the discrete grid for p, q (and hence for the image E itself) offset by $1/2$ picture cell in x and $1/2$ picture cell in y from the grid for z.

$$\frac{4}{\epsilon^2}\begin{array}{|c|c|c|}\hline & \frac{1}{4} & \\ \hline \frac{1}{4} & -1 & \frac{1}{4} \\ \hline & \frac{1}{4} & \\ \hline \end{array} \quad \text{or} \quad \frac{2}{\epsilon^2}\begin{array}{|c|c|c|}\hline \frac{1}{4} & & \frac{1}{4} \\ \hline & -1 & \\ \hline \frac{1}{4} & & \frac{1}{4} \\ \hline \end{array}\ .$$

The second, diagonal, form has a higher coefficient on the lowest-order error term than the first, edge-adjacent form, and so is usually not used by itself. The diagonal form is also typically not favored in iterative schemes for solving Poisson's equations, since it does not suppress certain high frequency components. We can write a stencil for a linear combination of the edge-adjacent and the diagonal versions in the form

$$\frac{4}{(a+1)\,\epsilon^2}\begin{array}{|c|c|c|}\hline \frac{a}{4} & \frac{1-a}{4} & \frac{a}{4} \\ \hline \frac{1-a}{4} & -1 & \frac{1-a}{4} \\ \hline \frac{a}{4} & \frac{1-a}{4} & \frac{a}{4} \\ \hline \end{array}\ .$$

A judiciously chosen weighted average, namely one for which $a = 1/5$, is normally preferred, since this combination cancels the lowest-order error term.

If we wish to prevent the iterative scheme from "walking away" from the solution, however, we need to make our estimate of the Laplacian consistent with repeated application of our estimators for the first partial derivatives. That is, we want our discrete estimate of Δz to be as close as possible to our discrete estimate of

$$(z_x)_x + (z_y)_y. \tag{95}$$

It is easy to see that the sum of the convolution of the discrete estimator for the x-derivative with itself and the convolution of the discrete estimator for the y-derivative with itself yields the diagonal pattern. So, while the diagonal pattern is usually not favored because it leads to less stable iterative schemes, it appears to be desirable here to avoid inconsistencies between discrete estimators of the first and second partial derivatives. Experimentation with various linear combinations bears this out. The edge-adjacent stencil is very stable and permits over-relaxation (SOR) with $\alpha = 2$ (see next section), but leads to some errors in the solution with noisefree input data. The diagonal form is less stable and requires a reduced value for α, but allows the scheme to converge to the exact algebraic solution to problems that have exact solutions.

The incipient instability inherent in use of the diagonal form is a reflection of the fact that if we think of the discrete grid as a checkerboard, then the "red" and the "black" squares are decoupled[33]. That is, updates of red squares are based only on existing values on red squares, while updates of black squares are based only on existing values on black squares. Equivalently, note that there is no change in the incremental update equations when we add a discrete function of the form

$$\delta z_{kl} = (-1)^{k+l} \tag{96}$$

to the current values of the height. The reason is that the estimators of the first derivatives and the diagonal form of the Laplacian estimator are completely insensitive to components of this specific (high) spatial frequency[34]. Fortunately, the iterative update cannot inject components of this frequency either, so that if the average of the values of the "red" cells initially matches the average of the values of the "black" cells, then it will continue to do so. The above has not turned out to be an important issue, since the iteration appears to be stable with the diagonal form of

[33] The "red" and "black" squares are the cells for which the sum of the row and column indexes are even and odd respectively.

[34] It may appear that this difficulty stems from the use of staggered grids. The problem is even worse when aligned grids are used, however, because the discrete estimator of the Laplacian consistent with simple central difference estimators of the first partial derivatives has a support that includes only cells that are 2ϵ away from the center. And this form of the Laplacian operator is known to be badly behaved. We find that there are *four* decoupled subsets of cells in this case.

the average, that is, for $a = 1$, when the natural boundary conditions are implemented with care.

6.3 Boundary Conditions

The boundary conditions have also to be dealt with properly to assure consistency between first- and second-order derivative estimators. In a simple rectangular image region, the natural boundary conditions for z could be implemented by simply taking the average of the two nearest values of the appropriate gradient component and multiplying by ϵ to obtain an offset from the nearest value of z in the interior of the grid. That is, for $1 \le k < n$ and $1 \le l < m$, we could use

$$\begin{aligned} z_{k,0} &= z_{k,1} - \frac{\epsilon}{2}(p_{k-1,0} + p_{k,0}) \\ z_{k,m} &= z_{k,m-1} + \frac{\epsilon}{2}(p_{k-1,m-1} + p_{k,m-1}) \\ z_{0,l} &= z_{1,l} - \frac{\epsilon}{2}(q_{0,l-1} + q_{0,l}) \\ z_{n,l} &= z_{n-1,l} + \frac{\epsilon}{2}(q_{n-1,l-1} + q_{n-1,l}) \end{aligned} \tag{97}$$

on the left, right, bottom and top border of a rectangular image region (the corners are extrapolated diagonally from the nearest point in the interior using both components of the gradient). But this introduces a connection between the “red”and the “black” cells, and so must be in conflict with the underlying discrete estimators of the derivatives that are being used.

One can do better using offsets from cells in the interior that lie in diagonal direction from the ones on the boundary. That is, for $2 \le k < n - 1$ and $2 \le l < m - 1$, we use

$$\begin{aligned} z_{k,0} &= \frac{1}{2}(z_{k-1,1} - \epsilon(p_{k-1,0} - q_{k-1,0}) + z_{k+1,1} - \epsilon(p_{k,0} + q_{k,0})) \\ z_{k,m} &= \frac{1}{2}(z_{k-1,m-1} + \epsilon(p_{k-1,m-1} + q_{k-1,m-1}) + z_{k+1,m-1} + \epsilon(p_{k,m-1} - q_{k,m-1})) \\ z_{0,l} &= \frac{1}{2}(z_{1,l-1} + \epsilon(p_{0,l-1} - q_{0,l-1}) + z_{1,l+1} - \epsilon(p_{0,l} + q_{0,l})) \\ z_{n,l} &= \frac{1}{2}(z_{n-1,l-1} + \epsilon(p_{n-1,l-1} + q_{n-1,l-1}) + z_{n-1,l+1} - \epsilon(p_{n-1,l} - q_{n-1,l})) \end{aligned} \tag{98}$$

on the left, right, bottom and top border of a rectangular image region. The corners are again extrapolated diagonally from the nearest point in the interior using both components of the gradient. Note that, in this scheme, one point on each side of the corner has to be similarly interpolated, because only one of the two values needed by the above diagonal template lies in the interior of the region.

If the surface gradient is not given on the image boundary, then natural boundary conditions must be used for p and q as well. The natural boundary condition is that the normal derivatives of p and q are zero. The simplest implementation is perhaps, for $1 \leq k < n - 1$ and $1 \leq l < m - 1$,

$$\begin{aligned} p_{k,0} &= p_{k,1} \\ p_{k,m-1} &= p_{k,m-2} \\ p_{0,l} &= p_{1,l} \\ p_{n-1,l} &= p_{n-2,l} \end{aligned} \tag{99}$$

and similarly for q (points in the corner are copied from the nearest neighbor diagonally in the interior of the region). It may be better to again use a different implementation, where the values for points on the boundary are computed from values at interior cells that have the same "color." That is, for $2 \leq k < n - 2$ and $2 \leq l < m - 2$,

$$\begin{aligned} p_{k,0} &= \frac{1}{2}(p_{k-1,1} + p_{k+1,1}), \\ p_{k,m-1} &= \frac{1}{2}(p_{k-1,m-2} + p_{k+1,m-2}), \\ p_{0,l} &= \frac{1}{2}(p_{1,l-1} + p_{1,l+1}), \\ p_{n-1,l} &= \frac{1}{2}(p_{n-2,l-1} + p_{n-2,l+1}), \end{aligned} \tag{100}$$

and similarly for q. As before, the corner points, and one point on each side of the corner have to be copied diagonally, without averaging, since only one of the two values needed lies in the interior of the region.

6.4 Iterative Schemes and Parallelism

There are numerous iterative schemes for solution of large sparse sets of equations, among them:

- Gauss-Seidel—with replacement—sequential update;
- Jacobi—without replacement—parallel update;
- Successive Over-Relaxation;
- Kazmarz relaxation;
- Line relaxation.

Successive over-relaxation (SOR) makes an adjustment from the old value that is α times the correction computed from the basic equations. That is, for example,

$$z_{kl}^{(n+1)} = z_{kl}^{(n)} + \alpha \left(\tilde{z}_{kl}^{(n)} - z_{kl}^{(n)} \right) \tag{101}$$

where $\tilde{z}_{kl}^{(n)}$ is the "new" value calculated by the ordinary scheme without over-relaxation. When $\alpha > 1$, this amounts to moving further in the direction of the adjustment than suggested by the basic equations. This can speed up convergence, but also may lead to instability[35]. The Gauss-Seidel method typically can be sped up in this fashion by choosing a value for α close to two—the scheme becomes unstable for $\alpha > 2$. Unfortunately the Gauss-Seidel method does not lend itself to parallel implementation.

The Jacobi method is suited for parallel implementation, but successive over-relaxation cannot be applied directly—the scheme diverges for $\alpha > 1$. This greatly reduces the speed of convergence. Some intuition may be gained into why successive over-relaxation cannot be used in this case, when it is noted that the neighbors of a particular cell, the ones on which the future value of the cell is based, are changed in the same iterative step as the cell itself. This does not happen if we use the Gauss-Seidel method, which accounts for its stability. This also suggests a modification of the Jacobi method, where the parallel update of the cells is divided into sequential updates of subsets of the cells. Imagine coloring the cells in such a way that the neighbors of a given cell used in computing its new value have a different color from the cell itself. Now it is "safe" to update all the cells of one color in parallel (for an analogous solution to a problem in binary image processing, see chapter 4 of [Horn 86]).

Successive over-relaxation can be used with this modified Jacobi method. If local averages are computed using only the four edge-adjacent neighbors of a cell, then only two colors are needed (where the colors are assigned according to whether $i + j$ is even or odd—see Figure 4). Each step of the iteration is carried out in two sub-steps, one for each of the cells of one color. The above shows that the improved convergence rates of successive over-relaxation can be made accessible to parallel implementations.

When the illumination of the surface is oblique (light source away from the viewer), $R(p, q)$ will tend to be locally approximately linear. This means that the gradient of $R(p, q)$ will point in more or less the same direction over some region of the image. The effect of this is that influences on the adjustments of the estimated gradient tend to be much smaller along a direction at right angles to the direction "away from the light source," than they are along other directions. This can be seen most

[35] Conversely, if the basic method has a tendency to be unstable, then one can "under-relax"—that is, use a value $\alpha < 1$.

z_{40}	z_{41}	z_{42}	z_{43}	z_{44}	z_{45}	z_{46}
z_{30}	z_{31}	z_{32}	z_{33}	z_{34}	z_{35}	z_{36}
z_{20}	z_{21}	z_{22}	z_{23}	z_{24}	z_{25}	z_{26}
z_{10}	z_{11}	z_{12}	z_{13}	z_{14}	z_{15}	z_{16}
z_{00}	z_{01}	z_{02}	z_{03}	z_{04}	z_{05}	z_{06}

Figure 4. The modified Jacobi method operates on subsets of cells with different "colors" at different times. In the simplest case, there are only two colors, one for the cells where the sum of the indexes is even, the other for the cells where the sum of the indexes is odd.

easily when the coordinate system is aligned with the direction toward a single light source in such a way that the reflectance map has bilateral symmetry with respect to the axis $q = 0$. Then R_q will be small, at least for gradients near the p-axis. In this case the coefficients on the diagonal of the 2×2 matrix may be very different in magnitude. This is analogous to a system of equations being much stiffer in one direction than another, and suggests that the convergence rate may be lower in this case. A possible response to this difficulty is the use of line relaxation.

6.5 Aliasing, and How to Avoid It

Discrete samples can represent a continuous waveform uniquely only if the continuous waveform does not contain frequency components above the Nyquist rate ($\omega_0 = \pi/\epsilon$, where ϵ is the spacing between samples). If a waveform is sampled that contains higher frequency components, these make contributions to the sampled result that are not distinguishable from low frequency components. If, for example, we have a component at frequency $\omega_0 < \omega < 2\omega_0$, it will make the same contributions as a component at frequency $2\omega_0 - \omega$. This is what is meant by *aliasing*. Ideally, the continuous function to be sampled should first be lowpass filtered. Filtering after sampling can only suppress desirable signal components along with aliased information.

Numerical estimation of derivatives is weakly *ill-posed*. The continuous derivative operator multiplies the amplitude of each spatial frequency component by the frequency, thus suppressing low frequencies and accentuating higher frequencies. Any corruption of the higher frequencies is noticeable, particularly if most of the signal itself is concentrated at lower frequencies. This means that we have to be careful how we estimate derivatives and how we sample the image.

Suppose, for example, that we have an image of a certain size, but that we would like to run our shape-from-shading algorithm on a smaller version, perhaps to obtain a result in a reasonable amount of time, or to cheaply provide useful initial values for iteration on the finer grid. It would be quite wrong to simply sub-sample the original image. Simple block-averaging is better, although frequency analysis shows that the response of a block-averaging filter first drops to zero only at twice the Nyquist frequency. It is better to use a cubic spline approximation of the ideal

$$\frac{\sin(\pi x/\epsilon)}{(\pi x/\epsilon)} \tag{102}$$

response for filtering before sub-sampling [Rifman & McKinnon 74] [Bernstein 76] [Keys 81] [Abdou & Young 82]. There is nothing specific in the above relating to shape-from-shading; these are considerations that apply generally to machine vision.

Similar notions apply to processing of the surface itself. If we have a digital terrain model of a certain resolution and want to generate a lower resolution shaded image from it, we need to first filter and sample the digital terrain model. Otherwise the result will be subject to aliasing, and some features of the shaded image will not relate in a recognizable way to features of the surface.

Finally, in creating synthetic data it is not advisable to compute the surface gradient on a regular discrete set of points and then use the reflectance map to calculate the expected brightness values. At the very least one should perform this computation on a grid that is much finer than the final image, and then compute block averages of the result to simulate the effect of finite sensing element areas—just as is done in computer graphics to reduce aliasing effects[36].

(This hints at an interesting problem, by the way, since the brightness associated with the average surface orientation of a patch is typically not quite equal to the average brightness of the surface, since the reflectance map is not linear in the gradient. This means that one has to use a reflectance map appropriate to the resolution one is working at—the

[36] One can obtain good synthetic data, however, with an exact algebraic solution, by sampling the height on a regular discrete set of points and then estimating the derivatives numerically, as discussed in section 5.1. This was done here to generate most of the examples shown in section 7.

reflectance map depends on the optical properties of the micro-structure of the surface, and what is micro-structure depends on what scale one is viewing the surface at.)

6.6 Measuring the Quality of Reconstruction

There are many ways of accessing the quality of the solution surface generated. Not all are useful:

- In the case of a synthetic image obtained from a surface model, the best test of the output of a shape-from-shading algorithm is comparison of the surface orientations of the computed result with those of the underlying surface. One can either compute the root-mean-square deviation of the directions of the computed surface normals from the true surface normals, or just the root-mean-square difference in the gradients themselves.
- Shading is a function of the surface gradient and thus most sensitive to higher spatial frequencies. Conversely, in the presence of noise and reconstruction errors, we expect that the lower spatial frequencies will not be recovered as well. This makes pointwise comparison of the heights of the computed surface with that of the original surface somewhat less useful, since errors in the lower spatial frequencies will affect this result strongly. Also, errors in height will be a function of the width of the region over which one has attempted to recover height from gradient.
- Comparison of an “image” obtained by making brightness a function of height with a similar “image” obtained from the original surface is usually also not very useful, since such a representation is not sensitive to surface orientation errors, only gross errors in surface height. Also, people generally find such displays quite hard to interpret.
- Oblique views of “wire-meshes” or “block-diagrams” defined by the discrete known points on the surface may be helpful to get a qualitative idea of surface shape, but can be misleading and are difficult to compare. If the shape-from-shading scheme is working anything like it is supposed to, the differences between the solution and the true surface are likely to be too small to be apparent using this mode of presentation.
- Comparing the original image with an image obtained under the same lighting conditions from the solution for the gradient (p, q) is not useful, since the brightness error is reduced very quickly with most iterative schemes. Also, a “solution” can have gradient field $\{p_{kl}, q_{kl}\}$

that yields exactly the correct image when illuminated appropriately, yet it may not even be integrable. In fact, the "surface" may yield an arbitrary second image when illuminated from a different direction unless p and q are forced to be consistent (that is, unless $p_y = q_x$) as discussed at the end of section 7.3.

- Slightly better is comparison of the original image with an image obtained under the same lighting conditions using numerical estimates of (z_x, z_y). But, unless the image is corrupted, or the assumptions about the reflecting properties of the surface and the lighting are incorrect, this synthetic image too will soon be very close to the original image.
- If the underlying surface is known, shaded views of the solution and the original surface, produced under lighting conditions *different* from those used to generate the input to the algorithm, are worth comparing. This is a useful test, that immediately shows up shortcomings of the solution method. It also is a graphic way of portraying the progress of the iteration—one that is easier to interpret than a set of numbers representing the state of the computation.
- Various measures of departure from integrability may be computed. Perhaps most useful are comparisons of numerical estimates of (z_x, z_y) with (p, q). Slightly less useful is the difference $(p_y - q_x)$ of the solution, since the height z may still not have converged to the best fit to p and q, even when the gradient itself is almost integrable.

6.7 When to Stop Iterating

As is the case with many iterative processes, it is difficult to decide when to stop iterating. If we knew what the underlying surface was, we could just wait for the gradient of the solution to approach that of the surface. But, other than when we test the algorithm on synthetic images, we do not know what the surface is, otherwise we would probably not be using a shape-from-shading method in the first place! Some other test quantities include:

- The brightness error

$$\iint \left(E(x,y) - R(p,q)\right)^2 \, dx \, dy \tag{103}$$

should be small. Unfortunately this error becomes small after just a few iterations, so it does not yield a useful stopping criterion.

- A slightly different brightness error measure
$$\iint (E(x,y) - R(z_x, z_y))^2 \, dx\, dy \tag{104}$$
is a bit more useful, for while it approaches the above when an exact solution is obtained, it lags behind until the gradient of z equals (p,q). When an exact solution is not possible, there will continue to be small differences between the gradient of z and (p,q), which means that this error measure does not tend to zero.
- The departure from smoothness
$$\iint (p_x^2 + p_y^2 + q_x^2 + q_y^2) \, dx\, dy \tag{105}$$
often drops as the solution is approached, but does not constitute a particularly good indicator of approach to the solution. In particular, when one comes close to the solution, one may wish to reduce the parameter λ, perhaps even to zero, in which case further iterations may actually *reduce* smoothness in order to better satisfy the remaining criteria.
- One of the measures of lack of integrability
$$\iint (p_y - q_x)^2 \, dx\, dy \tag{106}$$
is also not too useful, since it can at times become small, or stop changing significantly, even when z is still inconsistent with p and q.
- Another measure of lack of integrability
$$\iint ((z_x - p)^2 + (z_y - q)^2) \, dx\, dy \tag{107}$$
does appear to be very useful, since it drops slowly and often keeps on changing until the iteration has converged.
- One can also keep track of the rate of change of the solution with iterations
$$\iint \left(\frac{dp}{dt}\right)^2 + \left(\frac{dq}{dt}\right)^2 \, dx\, dy. \tag{108}$$
One should not stop until this has become small. In most cases it helps to continue for a while after the above measures stop changing rapidly since the solution often continues to adjust a bit.

Some of the implementation details given above may appear to be extraneous. However, when all of these matters are attended to, then the iterative algorithm will not “walk away” from the solution, and it will find the solution, to machine precision, given exact data (and assuming that boundary conditions for p and q are given, and that λ' is reduced to zero as the solution is approached). Convergence to the exact solution will not occur when something is amiss, such as a mismatch between the discrete estimators of the first derivative and the discrete estimator of the Laplacian. It is not yet clear how significant all of this is when one works with real image data, where there is no exact solution, and where the error introduced by incorrect implementation detail may be swamped by errors from other sources.

7. Some Experimental Results

The new algorithm has been applied to a number of synthetic images of simple shapes (such as an asymmetrical Gaussian, a sum of Gaussian blobs, and a sum of low frequency sinusoidal gratings) generated with a number of different reflectance maps (including one linear in p and q, Lambertian with oblique illumination, and a rotationally symmetric one). These synthetic images were small (usually 64×64 picture cells) in order to keep the computational time manageable. Typically the surface normals would be within a degree or two of the correct direction after a few hundred iterations. With appropriate boundary conditions, the computed shape would eventually be the same (to machine precision) as the shape used to generate the synthetic image. In each case, the brightness error decreased rapidly, while the integrability of the estimated gradient decreased much more slowly.

7.1 Graphical Depiction of Solution Process

For help in debugging the algorithm, and for purposes of determining a good schedule for adjusting the parameters μ and λ', it is useful to print out the diagnostic measurements discussed in sections 6.6 and 6.7. But it is hard to tell exactly what is going on just by looking at a large table of numbers such as that shown in Figure 5. It is important to also provide some graphic depiction of the evolving shape as it is computed. To make shaded images of the reconstructed surface useful, however, they must be illuminated from a direction *different* from the direction of illumination used for the original input image[37]. Shown in Figure 6, is such a sequence of shaded images generated during the reconstruction of the surface of a polyhedral object, starting from a random field of surface orientations.

[37] The test illumination should be quite different from the illumination used to generate the original image—preferrably lying in a direction that differs from the original source direction by as much as $\pi/2$

Figure 5. Diagnostic trace of various error measures. This sequence of results corresponds to the reconstruction of the sharp-edged crater shape shown in Figure 7. This kind of presentation is important, but must be supplemented by some graphic depiction of the evolving solution surface.

Here the image provided to the algorithm[38] corresponded to illumination form the Northwest, while illumination from the Northeast was used to display the reconstruction. Note how the edges become sharper as λ', controlling the contribution of the penalty term for departure from smoothness, is made smaller and smaller. This example illustrates the algorithm's ability to deal with surfaces that have discontinuities in surface orientation.

Because of the interest in application to astrogeology, a crater-like shape was also reconstructed, as shown in Figure 7. In this case, the algorithm rapidly found a shape that was generally correct, except for flaws in places on the rim of the crater in the Northeast and Southwest. These are areas where there is little contrast between the inside and the outside of the crater in the input image[39]. It took the algorithm a considerable number of additional iterations to determine the correct continuation of the shape computed in other image areas.

7.2 Emergent Global Organization

Often progress towards the correct solution is not as uneventful. Frequently small internally consistent solution patches will establish themselves, with discontinuities in surface orientation where these patches adjoin. Also, conical singularities form that tend to move along the boundaries between such regions as the iterative solution progresses. Conversely, boundaries between solution patches often form along curves connecting conical singularities that form earlier. After a large enough number of iterations, patches of local organization tend to coalesce and lead to emergent global organization. This can be observed best when λ' is smaller than what it would normally be for rapid convergence. In Figures 8 and 9, for example, are shown a sequence of shapes leading finally to a spherical cap on a planar surface. Within some regions, solution surface patches quickly establish themselves that individually provide good matches to corresponding parts of the input image. The borders between these internally consistent regions provide error contributions that the algorithm slowly reduces by moving the boundaries and incrementally changing the shapes within each of the regions. Too rapid a reduction of λ' can remove the incentive to reduce the creases and kinks and to freeze the solution in a state where some unnecessary discontinuities remain.

[38] The input image is not shown, but is just like the last image in the sequence shown, except that left and right are reversed.

[39] Again, the input image is not shown, but is like the last image in the sequence shown, except that left and right are reversed.

Figure 6. Reconstruction of a portion of a truncated hexahedron from a shaded image. The dihedral angle between each pair of the three outer surfaces is $\pi/2$. This example illustrates the algorithm's ability to deal with surfaces that have large discontinuities in surface orientation.

Figure 7. Reconstruction of a crater-like shape. Points on the rim in the Northeast and the Southwest correspond to places in the input image where there is least contrast between the inside and the outside, since the direction of the incident illumination is parallel to the rim there.

If, for example, λ' were to be set to zero with a “solution” consisting of a spherical cap with an inner disk inverted, as in the right hand image of the middle row of Figure 9, there would be no incentive to further reduce the length of the circular discontinuity, and the smooth solution for this part of the image would not be found.

The algorithm was also applied to impossible shaded images. Suppose, for example, that we are dealing with a Lambertian surface illuminated by a source near the viewer and that there is a dark smudge in the middle of a large planar region facing us (which appears brightly lit). It turns out that there is no surface with continuous first derivatives that could give rise to a shaded image with a simply connected, bounded dark region in the middle of a bright region [Horn, Szeliski & Yuille 89]. In Figure 10 we see what happens when the algorithm attempts to find a solution. Patches grow within which the solution is consistent with the image, but which yield discontinuities at boundaries between patches. Conical singularities sit astride these boundaries. For all random initial conditions tried, the algorithm eventually eliminates all but one of these conical singularities. The computed surface is in fact a “solution,” if one is willing to allow such singularities.

The graphical method of presenting the progress of the iterative solutions illustrated above was very helpful in debugging the program and in determining reasonable schedules for reduction of the parameters λ' and μ. Shown in Figure 11 are some examples of what happens when things go wrong. In the top row are shown instabilities arising in the solution for the crater-like shape, near the points where there is low contrast between the inside and the outside of the crater—that is, where there is no local evidence for curvature. These instabilities can be suppressed by reducing λ' more slowly. In the middle row are shown patterns resulting from various programming errors. Finally, in the bottom row is shown the propagation of an instability from a free boundary when λ' is set to zero. It appears that the process is not stable without the regularizer when the boundary is completely free. This is not too surprising, since the problem in this case may be underdetermined.

In the past, shape-from-shading algorithms have often been “tested” by verifying that the computed gradient field actually generates something close to the given input image. To show just how dangerous this is, consider Figure 12, which demonstrates a new non-iterative method for recovering a “surface” given a shaded image. In Figure 12(a), is the input to the algorithm, while Figure 12(c) is what the gradient field that is constructed by this algorithm looks like when illuminated in the same

Figure 8. Emergent global organization of local nonlinear iterative process. Internally consistent solutions arise in certain image patches with discontinuities at the borders between regions. The boundaries between these patches move, and the solutions within the patches adjust in order to reduce the sum of the error terms.

Figure 9. Emergent global organization of local nonlinear iterative process. Neighboring patches coalesce, as conical singularities are absorbed by coalescing with other singularities, or by being pushed towards a contour where the surface orientation is discontinuous. The final solution is a spherical cap resting on a plane.

Figure 10. What happens when the algorithm is confronted with an impossible shaded image? The input image here (not shown) is a circularly symmetric dark smudge in a uniformly bright surround. The light source is assumed to be near the viewer. The algorithm finds a “solution” with a single conical singularity.

Figure 11. Graphical depiction of instabilities and the effects of programming errors. In the top row are shown instabilities resulting from too rapid reduction of the penalty term for departure from smoothness. The middle row shows the results of various programming errors. The bottom row shows waves of instability propagating inwards from a free boundary.

Figure 12. New one-step shape-from-shading algorithm. See text!

way as the original surface. Now Figure 12(b) shows what the original surface looks like when illuminated from another direction. As a test, we should check whether the computed gradient field looks the same under these illuminating conditions. But behold, it does not! In Figure 12(d) we see what we get when we use the other illuminating condition. The "trick" here is that the problem of shape from shading is heavily underconstrained if we are only computing a gradient field and not enforcing integrability. There are many solutions and we can, in fact, impose additional constraints. The underlying gradient field here was computed by solving the photometric stereo equations [Woodham 78, 79, 80a, 89] for the two images in Figures 12(c) and (d) under the two assumed lighting conditions[40].

The new algorithm has also been applied to synthetic images generated from more complicated surfaces such as digital terrain models (DTMs) used earlier in research on interpretation of satellite images of hilly terrain [Horn & Bachman 78] [Sjoberg & Horn 83] and in automatic generation of shaded overlays for topographic maps [Horn 81]. These synthetic images were somewhat larger (the one used for Figure 1 is of size 231×178, for example). In this case, the simple algorithm, presented in section 5.3, using a regularizing term would often get trapped in a local minimum of the error function after a small number of iterations, while the modified algorithm presented in section 5.6, exploiting the linearization of the reflectance map, was able to proceed to a solution to machine precision after a few thousand iterations. Most of the surface normals typically were already within a degree or so of the correct direction after a few hundred iterations.

The closeness of approach to the true solution depends on several of the implementation details discussed earlier. In particular, it was helpful to use the old values of p and q for the reference point in the linearization of $R(p,q)$, rather than any of the other choices suggested earlier. Also, it helps to use the diagonal averaging scheme in the iteration for height rather than the one based on edge-adjacent neighbors.

7.3 Real Shaded Images

The new algorithm has also been applied to a few real images, mostly

[40] There is no guarantee that there is a solution of the photometric stereo problem for surface orientation, given two arbitrary brightness values, since the two equations are non-linear. In the particular case shown here, the dynamic range of the two images was such that a solution could be found at all but about a hundred picture cells.

aerial photographs and images taken by spacecraft. Shown here are the results obtained from a 108×128 patch of a 1024×1024 SPOT satellite image (CNES—Central National Experimential Station, France) taken in 1988 of the Huntsville, Alabama region. The ground resolution of such images, that is, the spacing between picture cells projected on the surface, is 10 meter. The area of interest is Monte Sano State Park, a tree-covered hilly region east of Huntsville.

The algorithm was run with free boundary conditions on both height and gradient. With real data, there typically is no exact solution, and the error terms cannot be reduced to zero. Correspondingly, with free boundary conditions, the iteration is not stable when the regularizer is removed completely, so there is a limit on how small one can make λ'. One side-effect of this is that the reconstructed surface is somewhat smoother than the real surface and consequently the vertical relief is attenuated somewhat. The actual vertical relief here, for example, is about 250 m, while the overall relief in the reconstruction is a bit less than 200 m.

At the time of this experiment, the viewing geometry and the light source position were not available, nor was information on atmospheric conditions or sensor calibration. The atmospheric scatter component was estimated by looking in regions that appear to be shadowed, where the reflected light component is expected to be small [Woodham 80b] [Horn & Sjoberg 83]. The product of illumination, surface albedo and camera sensitivity, was estimated by looking in regions that appeared to be turned to more or less face the light source. Unfortunately the range of grey-levels in the region of interest was rather small (23–42), since the sensor had been adjusted to so that it could cover the full dynamic range needed for the adjacent urban area, which was much brighter (21–149)[41]. Also, comparison of the left and right images indicates that there may be a certain degree of aliasing in these images.

The light source elevation was estimated by assuming that the average brightness of the image was approximately equal to the cosine of the angle between the average local surface normal and the light source direction. The polar angle of the light source (90° minus the elevation above the horizon) can then be found if one assumes further that the average local surface normal is approximately vertical. For this image, this method yielded a polar angle of about 65°, or an elevation of 25°.

The light source azimuth, that is, the projection of the direction toward the light source into the image plane, was first estimated to be about 60° clockwise from the x-axis of the image, based on the directions of

[41] The mapping finally chosen took a grey-level of 22 into 0.0 and a grey-level of 43 into 1.0 normalized surface radiance.

what appear to be shadows of tall buildings in the downtown area of Huntsville, as well as some other image features. Attempts to use Pentland's method [Pentland 84] for estimation of the source azimuth failed, as did Lee and Rosenfeld's refinement of that method [Lee & Rosenfeld 85]. A reasonable direction was found by instead computing the axis of least inertia through the origin [Horn 86] of a scattergram of the brightness gradient (E_x, E_y). There is a two way ambiguity in the result (corresponding to the usual convex versus concave interpretations of a surface) that can be resolved by other methods. Despite the crude nature of the scattergram, resulting from the coarse quantization of image irradiance measurements, an acceptable azimuth of between 60° and 65° was found in this fashion.

Finally, it was possible to refine this estimate of the azmiuth by running the shape from shading algorithm for various source azimuths and recording the remaining solution errors after many iterations. There was a broad minimum near an azimuth of 65°. This method of estimating the source azimuth, while computationally expensive, seems to be relatively reliable, since there does not appear to be a systematic deformation of the surface that can compensate for a change in azimuth of the light source while yielding a similar shaded image. Unfortunately the same cannot be said of the elevation angle of light source position, since tilting the surface about an axis perpendicular to the light source position, in such a way as to maintain the angle between the average surface normal and the direction to the light source, produces a similar shaded image—at least to first order.

Shown in Figure 13 is a registered stereo-pair of SPOT images of the Monte Sano State Park region. Note that the light comes from the lower right (not the upper left, as is common in artistic renderings of sculpted surfaces). The stereo pair is shown here so that the reader can get a better idea of the actual surface shape. The algorithm, when presented with the left image of the pair, calculates a shape used to generate the synthetic stereo pair in Figure 14. (The vertical relief has been exaggerated slightly in the computation of the synthetic stereo pair in order to partially compensate for the attenutation of vertical relief mentioned earlier[42].)

Another way of presenting the resulting shape is as a contour map. Shown in Figure 15(a) is a portion of the USGS 7.5' quadrangle of the Huntsville Alabama area, with the area covered by the left satellite photograph outlined, while Figure 15(b) shows a contour map derived from a

[42] While the base-to-height ratio in the satellite images appears to be about 0.5, it was assumed to be 0.75 for purposes of the computation of the synthetic stereo pair.

Figure 13. A stereo pair of SPOT satellite images of Monte Sano State park east of Hunstville, Alabama. The left subimage of 108×128 picture cells is used as input for the shape from shading algorithm.

Figure 14. A synthetic stereo pair computed from the solution obtained by the new shape from shading algorithm.

smoothed form of the solution obtained by the shape from shading algorithm. This is not a comparison that is likely to be flattering to the shape from shading algorithm, since we know that it is not very good at recovering the lower spatial frequencies. Conversely, the shape from shading algorithm finds a lot of detailed surface undulations that cannot be represented in a contour map. For this reason the surface must be smoothed or "generalized" before contours can be drawn.

For want of a better assumption, the spacecraft was at first assumed

Figure 15. (a) A portion of the USGS topographic map of the Huntsville Alabama area covering Monte Sano State park, with the approximate area covered by the left satellite image outlined. The rectangle is 3543' by 4200' (1080 m by 1280 m) and the contour interval is 20' (6.1 m). (b) Contour map derived from a smoothed version of the solution obtained by the shape from shading algorithm from the left satellite image.

to be vertically above the region of interest when the image was taken. Judging from lateral displacements of surface features it appears, however, that the left image was actually taken from a position that is about 15° away from the nadir, in the direction of the negative x-axis of the image coordinate system (and the right image from a position roughly the same amount in the other direction). This means that the computed result really applies to a tilted coordinate system. But more importantly, there is a distortion introduced by a poor estimate of the source direction occassioned by the assumption that average surface normal is parallel to the z-axis in the camera coordinate system. Attempts were made to compensate for this by estimating the source direction based on the assumption that the average surface normal was tilted 15° in the camera coordinate system. The reconstruction produced in this fashion was then rotated in the xz-plane to bring it back into alignment with local vertical. While the result produced in this way was better in certain ways (the lateral displacement of terrain features was greatly reduced), it was worse in others (including a small tilt of the result in the y direction). The moral is that to obtain quantitatively meaningful results, one needs to know accurately where the light source is in the camera coordinate system—and, if the result is to be related to some external coordinate system, then the camera position and attitude in that coordinate system needs to be known also.

The algorithm has, by the way, also been applied to some images of Mars taken by the Viking Orbiter. But since the "ground truth" is not (yet) available in the case of the Mars images, it is not possible to say much about the accuracy of the recovered surface orientation field.

7.4 Rating the Difficulty of Shape-from-Shading Problems

Experiments with synthetic shaded images suggests that certain shape-from-shading problems are relatively easy, while others are quite difficult. First of all, *basso-relievo*[43] surfaces (those with only low slopes) are easy to deal with (see also section 2.6) in comparison with *alto-relievo* surfaces (those with steep slopes). The digital terrain model used for the experiment illustrated in Figure 1 falls in the latter category, since the sides of the glacial cirque are steep and the individual gullies steeper still.

Typically the brightness of a surface patch increases the more it is turned towards the light source. If it is turned too far, however, it becomes so steep that its brightness once again decreases. There is a qualitative difference between shape-from-shading problems where none of the surface patches are turned that far, and those where some surface patches are so steep as to have reduced brightness. In the latter case, there appears to be a sort of two-way ambiguity locally about whether a patch is dark because it has not been turned enough to face the light source or whether it has been turned too far. This ensures that simplistic schemes will get trapped in local minima where patches of the solution have quite the wrong orientation. Similarly, the more sophisticated scheme described here takes many more iterations to unkink the resulting creases.

The transition between the two situations depends on where the light source is. The difficulty is reduced when the illumination is oblique (see also section 2.6). Conversely, the problem is more severe when the light source is at the viewer, in which case brightness decreases with slope independent of the direction of the surface gradient. This explains why the algorithm took longer to find the solution in the case of the spherical cap (Figure 8) since it was illuminated by a source near the viewer. It was more straightforward to find the solutions for the truncated hexa-

[43] For more regarding the terms *basso-relievo*, *mezzo-relievo* and *alto-relievo*, see [Koenderink & van Doorn 80].

hedron and the crater-like surface (Figures 6 and 7), both of which were illuminated obliquely. The above dichotomy is related to another factor: problems where the relevant part of the reflectance map is nearly linear in gradient are considerably easier to deal with than those in which the reflectance map displays strong curvatures of iso-brightness contours.

Smooth surfaces, particularly when convex, can be recovered easily. Surfaces with rapid undulations and wrinkles, such as the digital terrain model surface (Figure 1) are harder. Discontinuities in surface orientation are even more difficult to deal with. Note that, with the exception of the digital terrain model, all of the examples given here involve surfaces that have some curves along which the surface orientation is not continuous. The spherical cap, for example, lies on a planar surface, with a discontinuity in surface orientation where it touches the plane.

Problems where boundary conditions are not available, and where there are no occluding boundaries or singular points, are ill-posed, in the sense that an infinite variety of surfaces could have given rise to the observed shading. Not too surprisingly these tend to lead to instabilities in the algorithm, particularly when one attempts to reduce the penalty term for departure from smoothness. In these cases instabilities can be damped out to some extent by enforcing the image irradiance equation on the boundary by iterative adjustment of the gradient computed from the discrete approximation of the natural boundary conditions for p and q. But results have not been promising enough to be worth discussing here in more detail.

The number of iterations to converge to a good solution appears to grow almost quadratically with image size (number of rows or columns). This is because some effects have to “diffuse” across the image. This means that the total amount of computation grows almost with the fourth power of the (linear) image size. It is well known that ordinary iterative schemes for solving elliptic partial differential equations quickly damp out higher spatial frequency errors, while low frequency components are removed very slowly. One way to deal with this problem is to use computation on coarser grids to reduce the low spatial frequency components of the error. This is the classic multigrid approach [Brandt 77, 80, 84] [Hackbush 85] [Hackbush & Trottenberg 82]. It is clear that a true multigrid implementation (as opposed to a simple pyramid scheme)[44] would be required to pursue this approach further on larger images. This is mostly to cut down on the computational effort, but can also be expected to reduce

[44] A naive approach has one solve the equations on a coarse grid first, with the results used as initial conditions for a finer grid solution after interpolation. True multigrid methods are more complex, but also have much better properties.

even further the chance of getting caught in a local minimum of the error function. Implementation, however, is not trivial, since the equations are nonlinear, and because there are boundary conditions. Both of these factors complicate matters, and it is known that poor implementation can greatly reduce the favorable convergence rate of the basic multigrid scheme [Brandt 77, 80, 84].

Alternatively, one may wish to apply so-called direct methods for solving Poisson's equations [Simchony, Chellappa & Shao 89].

8. Conclusion

The original approach to the general shape-from-shading problem requires numerical solution of the characteristic strip equations that arise from the first-order nonlinear partial differential equation that relates image irradiance to scene radiance [Horn 70, 75]. Variational approaches to the problem instead minimize the sum of the brightness error and a penalty term such as a measure of departure from smoothness. These yield second-order partial differential equations whose discrete approximation on a regular grid can be conveniently solved by classic iterative techniques from numerical analysis. Several of these methods, however, compute surface orientation, not height, and do not ensure that the resulting gradient field is integrable [Ikeuchi & Horn 81] [Brooks & Horn 85]. One thus has, as a second step, to find a surface whose gradient comes closest to the estimated gradient field in a least-squares sense (see [Ikeuchi 84], chapter 11 in [Horn 86], and [Horn & Brooks 86]).

The two steps can be combined, and the accuracy of the estimated surface shape improved considerably, by alternately taking one step of the iteration for recovering surface orientation from brightness, and one step of the iteration that recovers the surface that best fits the current estimate of the surface gradient. This idea can be formalized by setting up a variational problem involving both the surface height above a reference plane and the first partial derivatives thereof. The resulting set of three coupled Euler equations can be discretized and solved much as the two coupled equations are in the simpler methods that only recover surface orientation.

Such an iterative scheme for recovering shape from shading has been implemented. The new scheme recovers height and gradient at the same time. Linearization of the reflectance map about the local average surface orientation greatly improves the performance of the new algorithm and could be used to improve the performance of existing iterative shape-from-shading algorithms. The new algorithm has been successfully applied to complex wrinkled surfaces; even surfaces with discontinuities in the gradient.

9. Acknowledgements

Michael Caplinger kindly provided pointers to relevant publications on photoclinometry, in particular the Ph.D. thesis of R.L. Kirk at the California Institute of Technology. Michael Brooks drew my attention to the fact that existing iterative schemes “walk away” from the correct solution and that regularization does not “pick one of the infinite number of solutions” of an ill-posed problem, as sometimes mistakenly stated. Bradford Washburn supplied a detailed topographic map of the Mt. Washington area of the White Mountains in New Hampshire from which digital terrain models used in the experiments here were interpolated. Steve Ahlmer and Michael Caplinger provided real shaded images of terrain on Earth and on Mars. Joel Moses found a way for the Electrical Engineering and Computer Science Department to help cover the cost of the author’s travel to the photoclinometry workshop held at Arizona State University, February 12–13, 1989. This workshop provided much of the impetus for the work reported here. Michael Brooks, Eric Grimson and Teresa Ehling provided additional helpful comments.

10. References

Abdou, I.E. & K.Y. Wong (1982) “Analysis of Linear Interpolation Schemes for Bi-Level Image Applications,” *IBM Journal of Research and Development*, Vol. 26, No. 6, pp. 667–686, November (see Appendix).

Bernstein, R. (1976) “Digital Image Processing of Earth Observation Sensor Data,” *IBM Journal of Research and Development*, pp. 40-57, January (see Appendix).

Blake, A., A. Zisserman & G. Knowles (1985) “Surface Descriptions from Stereo and Shading,” *Image & Vision Computing*, Vol. 3, No. 4, pp. 183–191. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Bonner, W.J. & R.A. Schmall (1973) “A Photometric Technique for Determining Planetary Slopes from Orbital Photographs,” U.S. Geological Survey Professional Paper 812-A, pp. 1–16.

Brandt, A. (1977) “Multi-level Adaptive Solutions to Boundary-Value Problems,” *Mathematics of Computation*, Vol. 31, No. 138, April, pp. 333–390.

Brandt, A. (1980) “Stages in Developing Multigrid Solutions,” in *Numerical Methods for Engineering*, Absi, E., R. Glowinski, P. Lascaux, H. Veysseyre (eds.), Dunod, Paris, pp. 23–44.

Brandt, A. (1984) *Multigrid Techniques: 1984 Guide with Applications to Fluid Dynamics*, Monograph available as GMD-Studie No. 85, from GMD-F1T, Postfach 1240, D-2505, St. Augustin 1, West Germany.

Brandt, A. & N. Dinar (1979) "Multigrid solutions of elliptic flow problems," in Parter, S.V. (ed.) *Numerical Methods for PDE*, Academic Press, New York, NY.

Brooks, M.J. (1983) "Two Results Concerning Ambiguity in Shape from Shading," *Proceedings of the National Conference on Artificial Intelligence*, Washington, D.C., August 22–26, pp. 36–39.

Brooks, M.J. (1985) Personal communication.

Brooks, M.J. & B.K.P. Horn (1985) "Shape and Source from Shading," *Proceedings of the International Joint Conference on Artificial Intelligence*, Los Angeles, CA, August 18–23, pp. 932–936. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Bruss, A.R. (1982) "The Eikonal Equation: Some Results Applicable to Computer Vision," *Journal of Mathematical Physics*, Vol. 23, No. 5, pp. 890–896, May. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Courant, R. & D. Hilbert (1953) *Methods of Mathematical Physics*, Volume I, John Wiley & Sons, New York, NY.

Courant, R. & D. Hilbert (1962) *Methods of Mathematical Physics*, Volume II, John Wiley & Sons, New York, NY.

Davis, P.A. & A.S. McEwen (1984) "Photoclinometry: Analysis of Inherent Errors and Implications for Topographic Measurement," *Lunar and Planetary Science Conference* XV, 12–16 March, pp. 194–195.

Davis, P.A. & L.A. Soderblom (1983) "Rapid Extraction of Relative Topography from Viking Orbiter Images: II. Application to Irregular Topographic Features," in *Reports on Planetary Geology Program* (1983), NASA Technical Memorandum 86246, pp. 29–30.

Davis, P.A. & L.A. Soderblom (1984) "Modeling Crater Topography and Albedo from Monoscopic Viking Orbiter Images: I. Methodology," *Journal of Geophysical Research*, Vol. 89, No. B11, October, pp. 9449–9457.

Davis, P.A., L.A. Soderblom & E.M. Eliason (1982) "Rapid Estimation of Martian Topography from Viking Orbiter Image Photometry," in *Reports on Planetary Geology Program* (1982), NASA Technical Memorandum 85127, pp. 331–332.

Deift, P. & Sylvester, J. (1981) "Some Remarks on the Shape-from-Shading Problem in Computer Vision," *Journal of Mathematical Analysis and Applications*, Vol. 84, No. 1, pp. 235–248, November.

Frankot, R.T. & R. Chellappa (1988) "A Method for Enforcing Integrability in Shape from Shading Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 4, pp. 439–451, July. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Garabedian, P.R. (1964) *Partial Differential Equations*, John Wiley & Sons, New York.

Hackbush, W. (1985) *Multigrid Methods and Applications*, Springer Verlag, Berlin.

Hackbush, W. & U. Trottenberg (eds.) (1982) *Multigrid Methods*, Springer Verlag, Berlin.

Hapke, B.W. (1963) "A Theoretical Photometric Function for the Lunar Surface," *Journal of Geophysical Research*, Vol. 68, No. 15, pp. 4571–4586, August.

Hapke, B.W. (1965) "An Improved Theoretical Lunar Photometric Function," *Astronomical Journal*, Vol. 71, No. 5, pp. 333–339, June.

Hapke, B.W. (1981) "Bidirectional Reflectance Spectroscopy: (1) Theory," *Journal of Geophysical Research*, Vol. 86, No. B4, April, pp. 3039–3054.

Hapke, B.W. (1984) "Bidirectional Reflectance Spectroscopy: (3) Correction for Macroscopic Roughness," *Icarus*, Vol. 59, pp. 41–59.

Hapke, B.W. & E. Wells (1981) "Bidirectional Reflectance Spectroscopy: (2) Experiments and Observations," *Journal of Geophysical Research*, Vol. 86, No. B4, April, pp. 3055–3060.

Harris, J.G. (1986) "The Coupled Depth/Slope Approach to Surface Reconstruction," S.M. Thesis, Department of Electrical Engineering and Computer Science, MIT. Also Technical Report 908, Artificial Intelligence Laboratory, MIT, Cambridge, MA.

Harris, J.G. (1987) "A New Approach to Surface Reconstruction: The Coupled Depth/Slope Model," *Proceedings of the International Conference on Computer Vision*, London, England, June 8–11, pp. 277–283.

Horn, B.K.P. (1970) "Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View," Ph.D. Thesis, Department of Electrical Engineering, MIT. Also Technical Report TR-79, Project MAC, MIT, Cambridge, MA. Also Technical Report TR-232, Artificial Intelligence Laboratory, MIT, Cambridge, MA.

Horn, B.K.P. (1975) "Obtaining Shape from Shading Information," Chapter 4 in *The Psychology of Computer Vision*, P.H. Winston (ed.), McGraw Hill, New York, NY, pp. 115–155. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Horn, B.K.P. (1977) "Understanding Image Intensities (sic)," *Artificial Intelligence*, Vol. 8, No. 2, pp. 201–231, April. Also in (1987) *Readings in Computer Vision*, Fischler, M.A. & O. Firschein (eds.), Kaufmann, pp. 45–60.

Horn, B.K.P. (1981) "Hill Shading and the Reflectance Map," *Proceedings of the IEEE*, Vol. 69, No. 1, pp. 14–47, January. Also (1982) *Geo-Processing*, Vol. 2, No. 1, pp. 65-146, October. Also (1979) "Automatic Hill-Shading and the Reflectance Map," *Image Understanding Workshop*, Palo Alto, CA, April 24–25, pp. 79–120.

Horn, B.K.P. (1984) "Extended Gaussian Images," *Proceedings of the IEEE*, Vol. 72, No. 12, pp. 1671–1686, December.

Horn, B.K.P. (1986) *Robot Vision*, MIT Press, Cambridge, MA & McGraw-Hill, New York, NY.

Horn, B.K.P. (1988) "Parallel Analog Networks for Machine Vision," Memo 1071, Artificial Intelligence Laboratory, MIT, Cambridge, MA, December.

Horn, B.K.P. & B.L. Bachman (1978) "Using Synthetic Images to Register Real Images with Surface Models," *Communications of the ACM*, Vol. 21, No. 11, pp. 914–924, November. Also "Registering Real Images Using Synthetic Images," in *Artificial Intelligence: An MIT Perspective* (Volume II), Winston, P.H. & R.H. Brown (eds.), MIT Press, Cambridge, MA, pp. 129–160.

Horn, B.K.P. & M.J. Brooks (1986) "The Variational Approach to Shape from Shading," *Computer Vision, Graphics and Image Processing*, Vol. 33, No. 2, pp. 174–208, February. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Horn, B.K.P. & M.J. Brooks (eds.) (1989) *Shape from Shading*, MIT Press, Cambridge, MA, June.

Horn, B.K.P. & B.G. Schunck (1981) "Determining Optical Flow," *Artificial Intelligence*, Vol. 17, No. 1–3, August 1981, pp. 185–203. Also (1980) Memo 572, Artificial Intelligence Laboratory, MIT, Cambridge, MA, April.

Horn, B.K.P. & R.W. Sjoberg (1979) "Calculating the Reflectance Map," *Applied Optics*, Vol. 18, No. 11, pp. 1770–1779, June. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Horn, B.K.P., R. Szeliski & A.L. Yuille (1989) "Impossible Shaded Images," submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Howard, A.D., K.R. Blasius, & J.A. Cutt (1982) "Photoclinometric Determination of the Topography of the Martian North Polar Cap," *Icarus*, Vol. 50, pp. 245–258.

Ikeuchi, K. (1984) "Reconstructing a Depth Map from Intensity Maps," *International Conference on Pattern Recognition*, Montreal, Canada, July 30–August 2, pp. 736–738. Also (1983) "Constructing a Depth Map from Images," Memo 744, Artificial Intelligence Laboratory, MIT, Cambridge, MA, August.

Ikeuchi, K. & B.K.P. Horn (1981) "Numerical Shape from Shading and Occluding Boundaries," *Artificial Intelligence*, Vol. 17, No. 1–3, pp. 141–184, August. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

John, F. (1978) *Partial Differential Equations*, Springer Verlag, Berlin.

Keys, R.G. (1981) "Cubic Convolution Interpolation for Digital Image Processing," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 29, No. 6, pp. 1153–1160, December.

Kirk, R.L. (1984) "A Finite-Element Approach to Two-Dimensional Photoclinometry," brief abstract in *Bulletin of the American Astronomical Society*, Vol. 16, No. 3, pg. 709.

Kirk, R.L. (1987) "A Fast Finite-Element Algorithm for Two-Dimensional Photoclinometry," Part III of Ph.D. Thesis, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA.

Koenderink, J.J. & A.J. van Doorn (1980) "Photometric Invariants Related to Solid Shape," *Optica Acta*, Vol. 27, No. 7, pp. 981–996. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Lambiotte, J.J. & G.R. Taylor (1967) "A Photometric Technique for Deriving Slopes from Lunar Orbiter Photography," *Proceedings of the Conference on the Use of Space Systems for Planetary Geology and Geophysics*, Boston, MA, May 25–27.

Lee, C.-H. & A. Rosenfeld (1985) "Improved Methods of Estimating Shape from Shading using the Light Source Coordinate System," *Artificial Intelligence*, Vol. 26, No. 2, pp. 125–143. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Lee, D. (1988) "Algorithms for Shape from Shading and Occluding Boundaries," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 5–9, Ann Arbor, MI, pp. 478–485. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Lucchitta, B.K. & N.A. Gambell (1970) "Evaluation of Photoclinometric Profile Determination," in *Analysis of Apollo 8 Photographs and Visual Observations*, NASA SP-201, National Aeronautics and Space Administration, pp. 51–59.

Malik, J. & D. Maydan (1989) "Recovering Three Dimensional Shape from a Single Image of Curved Objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 6, pp. 555–566, June. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Malin, M.C. & G.E. Danielson (1984) "Topography on Ganymede Derived from Photoclinometry," in *Reports on Planetary Geology Program* (1983), NASA Technical Memorandum 86246, pp. 29–30.

McEwen, A.S. (1985) "Topography and Albedo of Ius Chasma, Mars," *Lunar and Planetary Science Conference* XVI, 11–15 March, pp. 528–529.

Mingolla, E. & J.T. Todd (1986) "Perception of Solid Shape from Shading," *Biological Cybernetics*, Vol. 53, pp. 137–151. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Minnaert, M. (1961) "Photometry of the Moon," Chapter 6 in Volume 3 of *Planets and Satellites: The Solar System*, Kuiper, G.P. & B.M. Middlehurst (eds.), University of Chicago Press, Chicago, IL, pp. 213–248.

Passey, Q.R. & E.M. Shoemaker (1982) "Craters and Basins on Ganymede and Callisto: Morphological Indicators of Crustal Evolution," in Morrison, D. (ed.) *Satellites of Jupiter*, University of Arizona Press, Tucson, pp. 379–434.

Pentland, A.P. (1984) "Local Shading Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 2, pp. 170–187, March. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Pentland, A.P. (1988) "Shape Information from Shading: A Theory about Human Perception," Technical Report 103, Vision Sciences, MIT Media Laboratory, MIT, Cambridge, MA, May.

Rifman, S.S.& D.M. McKinnon (1974) "Evaluation of Digital Correction Techniques—for ERTS images," Report Number E74-10792, TRW Systems Group, July (see Chapter 4). Also Final Report TRW 20634-6003-TU-00, NASA Goddard Space Flight Center.

Rindfleisch, T. (1966) "Photometric Method for Lunar Topography," *Photogrammetric Engineering*, Vol. 32, No. 2, pp. 262–277, March. Also (1965) "A Photometric Method for Deriving Lunar Topographic Information," Technical Report 32-786, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, September.

Rowan, L.C., J.F. McCauley & E.A. Holm (1971) "Lunar Terrain Mapping and Relative Roughness Analysis," U.S. Geological Survey Professional Paper 599-G, pp. 1-32.

Saxberg, B.V.H. (1988) "A Modern Differential Geometric Approach to Shape from Shading," Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA.

Shao, M., T. Simchony & R. Chellappa (1988) "New Algorithms for Reconstruction of a 3-D Depth Map from One of More Images," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 5–9, Ann Arbor, MI, pp. 530–535.

Simchony, T., R. Chellappa & M. Shao (1989) "Direct Analytical Methods for Solving Poisson Equations in Computer Vision Problems," unpublished report, University of Southern California. Also in *IEEE Computer Society Workshop on Computer Vision*, Miami Beach, Florida, November.

Sjoberg, R.W. & B.K.P. Horn (1983) "Atmospheric Effects in Satellite Imaging of Mountainous Terrain," *Applied Optics*, Vol. 22, No. 11, pp. 1702–1716, June.

Squyres, S.W. (1981) "The Topography of Ganymede's Grooved Terrain," *Icarus*, Vol. 46, pp. 156–168.

Strat, T. (1979) "A Numerical Method for Shape from Shading for a Single Image," S.M. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA.

Sugihara, K. (1986) *Machine Interpretation of Line Drawings*, Chapter 10, MIT Press, Cambridge, MA.

Terzopoulos, D. (1983) "Multilevel Computational Processes for Visual Surface Reconstruction," *Computer Vision, Graphics and Image Processing*, Vol. 24, pp. 52–96. Also (1982) "Multi-level Reconstruction of Visual Surfaces: Variational Principles and Finite Element Representation, Memo 671, Artificial Intelligence Laboratory, MIT, Cambridge, MA, April.

Terzopoulos, D. (1984) "Multigrid Relaxation Methods and the Analysis of Lightness, Shading, and Flow," Memo 803, Artificial Intelligence Laboratory, MIT, Cambridge, MA, October. Also chapter 10 in *Image Understanding 84*, Ullman, S. & W. Richards (eds.), Ablex Publishing Corporation, Norwood, NJ, pp. 225–262.

Tyler, G.L., R.A. Simpson & H.J. Moore (1971) "Lunar Slope Distributions: Comparison of Bi-Static Radar and Photographic Results," *Journal of Geophysical Research*, Vol. 76, No. 11, pp. 2790–2795.

Watson, K. (1968) "Photoclinometry from Spacecraft Images," U.S. Geological Survey Professional Paper 599-B, pp. 1–10.

Wildey, R.L. (1975) "Generalized Photoclinometry for Mariner 9," *Icarus*, Vol. 25, pp. 613–626.

Wildey, R.L. (1984) "Topography from Single Radar Images," *Science*, Vol. 224, pp. 153-156, April.

Wildey, R.L. (1986) "Radarclinometry for the Venus Radar Mapper," *Photogrammetric Engineering and Remote Sensing*, Vol. 52, No. 1, pp. 41–50, January. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Wilhelms, D.E. (1964) "A Photometric Technique for Measurement of Lunar Slopes," in *Astrogeological Studies Annual Progress Report, Part D: Studies for Space Flight Program*. U.S. Geological Survey Open-File Report, pp. 1–12, May, NASA Catalog Number N66 35597.

Wilson, L., M.A. Brown, E.M. Parmentier & J.W. Head (1984) "Theoretical Aspects of Photoclinometry Terrain Profiling on the Galilean Satellites," in *Reports on Planetary Geology Program* (1983), NASA Technical Memorandum 86246, pp. 27–28.

Wilson, L., J.S. Hampton, & H.C. Balen (1985) "Photoclinometry of Terrestrial & Planetary Surfaces," *Lunar and Planetary Science Conference* XVI, 11–15 March, pp. 912–913.

Woodham, R.J. (1977) "A Cooperative Algorithm for Determining Surface Orientation from a Single View," *International Joint Conference on Artificial Intelligence*, Cambridge, MA, August 22–25, pp. 635–641.

Woodham, R.J. (1978) "Photometric Stereo: A Reflectance Map Technique for Determining Surface Orientation from a Single View," *Image Understanding Systems & Industrial Applications*, *Proceedings of the Society of Photo-Optical Instrumentation Engineers*, Vol. 155, pp. 136–143.

Woodham, R.J. (1979) "Analyzing Curved Surfaces using Reflectance Map Techniques," in *Artificial Intelligence: An MIT Perspective* (Volume II), Winston, P.H. & R.H. Brown (eds.), MIT Press, Cambridge, MA, pp. 161–184.

Woodham, R.J. (1980a) "Photometric Method for Determining Surface Orientation from Multiple Images," *Optical Engineering*, Vol. 19, No. 1, January-February, pp. 139–144. Also in (1989) *Shape from Shading*, Horn, B.K.P. & M.J. Brooks (eds.), MIT Press, Cambridge, MA.

Woodham, R.J. (1980b) "Using terrain Digital Data to Model Image Formation in Remote Sensing," *Image Processing for Guidance*, SPIE, Vl. 238, pp. 361-369.

Woodham, R.J. (1989) "Determining Surface Curvature with Photometric Stereo," *Proceedings IEEE Conference on Robotics and Automation*, 14–19 May.