MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo 813 March, 1985

The Variational Approach to Shape from Shading

Berthold K.P. Horn
Michael J. Brooks

Abstract: We develop a systematic approach to the discovery of parallel iterative schemes for solving the shape-from-shading problem on a grid. A standard procedure for finding such schemes is outlined, and subsequently used to derive several new ones.

The shape-from-shading problem is known to be mathematically equivalent to a non-linear first-order partial differential equation in surface elevation. To avoid the problems inherent in methods used to solve such equations, we follow previous work in reformulating the problem as one of finding a surface orientation field that minimizes the integral of the brightness error. The calculus of variations is then employed to derive the appropriate Euler equations on which iterative schemes can be based.

The problem of minimizing the integral of the brightness error term is ill posed, since it has an infinite number of solutions in terms of surface orientation fields. A previous method used a regularization technique to overcome this difficulty. An extra term was added to the integral to obtain an approximation to a solution that was as smooth as possible.

We point out here that surface orientation has to obey an integrability constraint if it is to correspond to an underlying smooth surface. Regularization methods do not guarantee that the surface orientation recovered satisfies this constraint. Consequently, we attempt to develop a method that enforces integrability, but fail to find a convergent iterative scheme based on the resulting Euler equations. We show, however, that such a scheme can be derived if, instead of strictly enforcing the constraint, a penalty term derived from the constraint is adopted. This new scheme, while it can be expressed simply and elegantly using the surface gradient, unfortunately cannot deal with constraints imposed by occluding boundaries. These constraints are crucial if ambiguities in the solution of the shape-from-shading problem are to be avoided.

Different schemes result if one uses different parameters to describe surface orientation. We derive two new schemes, using unit surface normals, that facilitate the incorporation of the occluding boundary information. These schemes, while more complex, have several advantages over previous ones.

This report describes research done 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 System Development Foundation. This work was done while Michael Brooks was a visiting scientist at MIT on leave from Flinders University.

1. Introduction

We begin by reviewing the shape-from-shading problem, its formulation as a minimization problem, and the use of the calculus of variations in deriving the partial differential equations governing the solution of the minimization problem.

1.1. Preview

The first study of the shape-from-shading problem was undertaken by Horn (1970 & 1975). There, the partial differential equation in surface elevation fundamental to the problem was converted to an equivalent set of five ordinary differential equations called the *characteristic strip* equations. Algorithms based directly on numerical solution of the discrete approximations of these equations are inherently sequential in nature and have difficulty with unavoidable noise in the image data.

Later, a method lending itself to parallel solution on a grid was developed by Strat (1979) using minimization in the discrete domain. Strat used the gradient to express surface orientation and so was unable to deal with occluding boundaries, which are known to provide crucial constraint needed to avoid ambiguity in the solution, as shown by Bruss (1983). For this reason, another approach, based on the stereographic projection of the Gaussian sphere, was explored by Ikeuchi and Horn (1981). The calculus of variations was used there for the first time in the analysis of the shape-from-shading problem. Their method depended on the use of a regularization term in the functional to be minimized.

In this paper, we carefully examine the role of the variational calculus in the derivation of iterative schemes for shape from shading. Previous methods are discussed in detail, and rationalized in terms of the new point of view, where appropriate. The application of regularization techniques to well-posed problems is called into question.

We note in particular that the surface gradient should satisfy an integrability constraint. Guided by this observation, we attempt to impose integrability in a strict sense. We are, however, unable to derive a convergent iterative scheme based on the appropriate Euler equation. We learn that such a scheme may be found if we instead incorporate a penalty term based on the integrability constraint. This we demonstrate first using the gradient to specify surface orientation, as has been customary. The resulting iterative scheme is shown to be related to that developed by Strat.

As already stated, use of the surface gradient precludes incorporation of the occluding boundary information. We overcome this difficulty by taking the novel approach of adopting surface-normal vectors directly. This leads to iterative schemes that are more complex, but manageable. We finally develop two such schemes that:

- ensure the result is (at least approximately) integrable,
- avoid the smoothing introduced by a regularizing term, and
- permit use of the known normals on the occluding boundary.

None of the previous methods combined all of these features.

1.2. The shape-from-shading problem

Monochrome images of smoothly curved surfaces with homogeneous reflecting properties commonly exhibit a variation in image irradiance, or *shading*. This is due to the interaction of four principal factors: the illumination, the shape of the surface, the reflecting characteristics of the material, and the image projection. The *shape-from-shading* problem may be regarded as that of extracting the shape information encoded in the irradiance data. It therefore entails inversion of the image-forming process.

Because a number of factors are confounded in irradiance values, the shape depicted in an image cannot be determined unless additional information is provided. Of considerable utility in this regard has been the *reflectance map* (Horn, 1977), which specifies the radiance of a surface patch as a function of its orientation. The reflectance map can be computed from the bidirectional reflectance-distribution function and the light-source arrangement (Horn & Sjoberg, 1979). Usually it is more practical to determine the reflectance map experimentally, by means of a calibration object of known shape, for example. In any case, the reflectance map encodes, inextricably, information about the reflecting properties of the surface and the distribution and intensity of the light sources.

In adopting the reflectance map, we implicitly make the assumption that, for the given scene conditions, the radiance emanating from a small surface patch is dependent only on the orientation of the patch, and not its position in space. This requires that the light sources and the viewer be distant. We also assume that the image is formed by orthographic image projection, and that the surface has homogeneous reflecting properties[1].

Formally, given an image, E, and a reflectance map, R, the shape-from-shading problem may be regarded as that of recovering a smooth surface, z, satisfying the *image irradiance equation*

$$E(x,y) = R\big(z_x(x,y), z_y(x,y)\big)$$

over some domain Ω of the image. Any given conditions on z on the boundary $\partial\Omega$ of the region Ω should also be satisfied. Here z_x and z_y denote the first partial derivatives of z with respect to x and y respectively. Since these derivatives will be used frequently to specify surface orientation, it is convenient to introduce the short-hand notation

$$p = \frac{\partial z}{\partial x} \qquad \text{and} \qquad q = \frac{\partial z}{\partial y}.$$

The *gradient* of the surface z at the point (x,y) is just $\big(p(x,y), q(x,y)\big)$. The gradient points in the direction of steepest ascent and has magnitude equal to the slope in that direction. It is further useful to note that a normal of the once-differentiable surface, z, at $\big(x,y,z(x,y)\big)^T$ can be written

$$\mathbf{n} = \big(-p(x,y), -q(x,y), 1\big)^T.$$

This follows from the fact that $\big(1,0,p(x,y)\big)^T$ and $\big(0,1,q(x,y)\big)^T$ are tangent vectors and that the normal must be parallel to their cross-product. For many purposes one can use

[1] These restrictive assumptions were not exploited in the original work on shape from shading. The problem formulation, however, is much easier to comprehend if the reflectance map is introduced, and that can be done only if these additional constraints are imposed.

either the surface gradient or the normal to specify surface orientation. Each has its own advantages, as we shall see.

It is customary to choose the direction of projection to be parallel to the z-axis. On the *occluding boundary*, the direction of projection is tangent to the surface. That is, the normal is orthogonal to a unit vector $\hat{\mathbf{z}}$, parallel to the z-axis. Thus we note that at least one of p and q become unbounded on the occluding boundary.

1.3. Employing the variational calculus

Suppose we seek, over some domain, a smooth surface satisfying various constraints. It is useful to obtain from the given constraints a non-negative expression that measures the departure of a particular surface from a satisfactory solution. We may then search for a surface that minimizes the expression. As the value of the expression depends on the choice of surface, or function, it is termed a *functional.*

The search for a function that minimizes an integral expression is the major concern of the *calculus of variations* (Courant & Hilbert, 1953). Here, we find the valuable result that the extrema of functionals must satisfy an associated *Euler equation.* This equation can usually be determined in a straightforward way from the functional. We can, as a result, transform our surface-recovery problem from one of minimizing a functional, to one of solving one or more partial differential equations. Some of the relevant mathematical details are presented in the Appendix of this paper.

In seeking a surface that best matches the aforementioned constraints, we require a global minimum of the corresponding functional. However, Euler equations only specify conditions on extremal values. We shall make the strong assumption in this paper that a solution to the Euler equation constitutes a global minimum of the functional, satisfying the constraints optimally. We shall as a result be deluded if we encounter a surface that gives rise to either a local minimum, a local maximum, or an inflexion point in the functional, for it too will satisfy the Euler equation[2]. The assumption here is difficult to avoid, given that we shall be dealing with functionals involving a reflectance map whose analytic form may not be known in advance.

Let us suppose that we obtain from an Euler equation a surface that generates a global minimum of the appropriate functional. It may be that the constraints on which the functional was originally based are satisfied exactly by this function. However, this need not be so. Problems can readily be formulated for which there are no perfect solutions. But here we find a very important property of this approach: the surface that best matches the constraints will generate a global minimum of the functional. This is important to vision problems as they typically involve images that are noisy. Exact solutions may not exist in this situation. For example, in the presence of noise, there may not be a smooth surface that satisfies the image irradiance equation $E(x,y) = R(p,q)$ exactly. There will, however, be a surface that minimizes the integral of the square of the difference between $E(x,y)$ and $R(p,q)$[3].

---

[2] Because of the use of expressions that are unbounded above, we shall not encounter solutions generating global maxima.

[3] The integral of the square of the difference may have a lower bound that is not attained by any surface. In that case, a surface may be found for which the integral is arbitrarily close to that lower bound.

It is important to observe that there are typically an infinite number of surfaces satisfying the Euler equation. Without further constraint, we do not have a well-posed problem. In some cases the original problem includes *boundary conditions* that, taken together with the resulting partial differential equations, lead to a unique solution. In the case where the unknown function is unconstrained on the boundary, the calculus of variations itself provides so-called *natural boundary conditions* (see Appendix).

Care must be taken when formulating the functional to ensure that it provides sufficient constraint, for otherwise there may be an infinite number of solutions even with boundary conditions. Such a difficulty may be remedied by the addition of a suitable *regularization* term (Poggio & Torre, 1984). This is discussed in more detail in the Appendix.

1.4. A procedure for deriving iterative schemes

We now consider a way of deriving iterative schemes for recovering surface shape. In the event that we seek a surface, z, best satisfying various requirements over Ω, we do the following:

(1) Select a functional, F, non-negative over Ω, such that

$$I(z) = \iint_{\Omega} F(x, y, z, \ldots)\, dx\, dy$$

constitutes a measure of the departure of z from an ideal solution.

(2) Absorb into F any constraint that z should satisfy over Ω, using Lagrangian multipliers if appropriate.

(3) If the problem is not well posed as it stands, add a suitable regularization term.

(4) Find the Euler equation that must be satisfied by the surface z minimizing the functional I.

(5) Determine what boundary conditions are needed to ensure a unique solution. If there are no constraints on the function around the boundary $\partial\Omega$, determine the appropriate natural boundary conditions.

(6) Develop a discrete approximation of the associated Euler equation, using finite-difference methods.

(7) Design an iterative scheme that converges to the solution of the discrete approximation of the Euler equation.

The approach, of course, follows the same pattern if the surface is parameterized in a different way[4]. Also, similar results can be obtained by applying the finite-element method directly to the functional I.

As we shall see later, the most difficult step here is typically the discovery of an iterative scheme that enables one to recover a solution of the discrete approximation of

[4] The surface may, for example, be parametrized using the gradient (p, q) instead of the surface elevation z, for example.

the Euler equation. Such a scheme should be efficient, convergent, and preferably lend itself to parallel implementation.

Note that it is better to work with a functional that evaluates to zero for perfect solutions. In this way, one is relieved of the onus of showing that there are no unwanted surfaces that cause the functional to have a smaller value than that generated by a satisfactory solution. An additional advantage of functionals evaluating to zero is that one may use them to check how close an iterative scheme is to a solution. This is difficult with other functionals, as the minimum value is usually unknown.

2. Previous work

Only one shape-from-shading scheme (Ikeuchi & Horn, 1981), prior to this work, has been devised by explicit recourse to the calculus of variations. Two other schemes, however, (Strat 1979, Smith 1982) can be rationalized by application of the calculus of variations. We now examine these three schemes in historical sequence.

2.1. Strat's method

Strat (1979) arrived at his method by application of the standard calculus to the discrete domain. We present his analysis here as we wish to show later how it can be related to a new scheme we develop using the calculus of variations. Rationalizing Strat's scheme directly in terms of the variational calculus is complicated by the fact that it is based on an integral (rather than a differential) integrability term.

First, let the *brightness error* at a point (x, y) be

$$E(x,y) - R\big(p(x,y), q(x,y)\big).$$

This is the difference between the observed irradiance $E(x, y)$ and that predicted from the estimated gradient $\big(p(x,y), q(x,y)\big)$. In the discrete case, we might consider minimizing the total brightness error[5]

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\big(E_{ij} - R(p_{ij}, q_{ij})\big)^2$$

by suitable choice of the gradient at each picture cell in the image[6]. In this vein, then, by setting the derivative of the expression with respect to p_{kl} and q_{kl} equal to zero, we obtain, for $1 \le k \le n$ and $1 \le l \le m$, the two sets of equations

$$\big(E_{kl} - R(p_{kl}, q_{kl})\big)\, R_p\big(p_{kl}, q_{kl}\big) = 0,$$
$$\big(E_{kl} - R(p_{kl}, q_{kl})\big)\, R_q\big(p_{kl}, q_{kl}\big) = 0,$$

[5] For simplicity, we assume a rectangular image region here. There is no loss of generality, however, since the sums can be taken over whatever region is desired.

[6] By minimizing an expression containing the sum of the square of the brightness error, we are giving up strict enforcement of the image irradiance equation. This seems reasonable, given that neither E nor R are known with precision.

where R_p and R_q are the partial derivatives of R with respect to p and q respectively. These conditions can be trivially satisfied if we choose p_{ij} and q_{ij} so that

$$R(p_{ij}, q_{ij}) = E_{ij}.$$

Since this equation represents but one constraint on the two unknowns p_{ij} and q_{ij}, we expect that, in general, an infinite number of gradient values will satisfy it, for a particular i and j. Many solutions can then be constructed by combining arbitrary choices from these sets of possibilities at each picture cell.

The problem is clearly not well posed as stated. We can, however, make use of the fact that the gradients at neighboring points are related. Consider an infinitesimal segment, δC, of a curve on the surface. The change in z along the segment is given by

$$\delta z = p\,\delta x + q\,\delta y,$$

where δx and δy are the changes in x and y along the segment. The total change in z along a curve then is just the integral of $(p\,dx + q\,dy)$. In the case of a closed curve, C, this integral should be zero. Thus, if $\big(p(x,y), q(x,y)\big)$ is the gradient of a surface $z(x,y)$ then

$$\oint_C \big(p(x,y)\,dx + q(x,y)\,dy\big) = 0,$$

for all closed curves, C, in the region Ω[7].

Let ϵ denote the spacing between picture cells. Consider an elementary square path, with the picture cell (i,j) in the lower left hand corner. If we let i correspond to x and j correspond to y, then the integral counter-clockwise around this path can be estimated by

$$e_{ij} = \frac{\epsilon}{2}\left[p_{i,j} + p_{i+1,j} + q_{i+1,j} + q_{i+1,j+1} - p_{i+1,j+1} - p_{i,j+1} - q_{i,j+1} - q_{i,j}\right].$$

This expression can be obtained by approximating the slope along each of the four sides by the average of the slopes at the beginning and end of each side. The result is exactly equal to zero when z is quadratic, as can be seen using Taylor series expansion[8]. The difference between this expression and the exact loop integral is (perhaps surprisingly) of order ϵ^4.

On a discrete grid, we wish to minimize two errors: the brightness error, summed over all grid points, and the error in the loop integrals, summed over all elementary square paths constructed by connecting the centers of neighboring picture cells[9].

The total contribution of the first error term clearly depends on the number of nodes in the grid, that is, it depends inversely on ϵ^2 for a fixed image size. We show later that the second term, on the other hand, varies directly as ϵ^2. To make the relative contribution of the two terms independent of the grid spacing, we multiply the first term

[7] For a discussion of this issue, within the context of the shape-from-shading problem, see Brooks (1979).

[8] To be exact, it equals zero when z can be written as a polynomial containing only terms of the form $x^i y^j$, for $i \le 2$ and $j \le 2$, for $i = j$, for $i = 0$ with j arbitrary, and for $j = 0$ with i arbitrary.

[9] Strat actually counted each loop integral four times.

by ϵ^2 and divide the second term by ϵ^2. The quantity to be minimized then becomes[10]

$$\epsilon^2 \sum_{i=1}^{n} \sum_{j=1}^{m} \big(E_{ij} - R(p_{ij}, q_{ij})\big)^2 + \frac{\lambda}{\epsilon^2} \sum_{i=1}^{n-1} \sum_{j=1}^{m-1} e_{ij}^2.$$

Here λ is a factor that weights the relative contributions of the brightness error and the errors in the elementary loop integrals. It can be made small when the irradiance measurements are accurate, and the reflectance map is known with precision[11].

For the composite error term to be a minimum, the derivatives of the error sum with respect to p_{kl} and q_{kl} must be zero. Now, p_{kl} and q_{kl} occur in the expressions for $e_{k,l}$, $e_{k-1,l}$, $e_{k-1,l-1}$ and $e_{k,l-1}$. So performing the indicated differentiations and equating the results to zero, one obtains, for $1 < k < n$ and $1 < l < m$,

$$\epsilon^2 \big(E_{kl} - R(p_{kl}, q_{kl})\big) R_p(p_{kl}, q_{kl}) - \frac{\lambda}{2\epsilon} \left[e_{k,l} + e_{k-1,l} - e_{k-1,l-1} - e_{k,l-1}\right] = 0,$$

$$\epsilon^2 \big(E_{kl} - R(p_{kl}, q_{kl})\big) R_q(p_{kl}, q_{kl}) - \frac{\lambda}{2\epsilon} \left[e_{k-1,l} + e_{k-1,l-1} - e_{k,l-1} - e_{k,l}\right] = 0,$$

where R_p and R_q are the partial derivatives of $R(p, q)$ with respect to p and q, as before[12]. We can change dummy variables again and gather terms in a particular way to obtain

$$\epsilon^2 \big(E_{ij} - R(p_{ij}, q_{ij})\big) R_p(p_{ij}, q_{ij}) + \lambda[\bar{v}_{ij} - \tilde{q}_{ij}] = 0,$$

$$\epsilon^2 \big(E_{ij} - R(p_{ij}, q_{ij})\big) R_q(p_{ij}, q_{ij}) + \lambda[\bar{h}_{ij} - \tilde{p}_{ij}] = 0,$$

where

$$\tilde{p}_{ij} = \tfrac{1}{4} \left(p_{i+1,j+1} - p_{i-1,j+1} + p_{i-1,j-1} - p_{i+1,j-1}\right),$$

$$\tilde{q}_{ij} = \tfrac{1}{4} \left(q_{i+1,j+1} - q_{i-1,j+1} + q_{i-1,j-1} - q_{i+1,j-1}\right),$$

are discrete estimates of the second partial cross derivatives p_{xy} and q_{xy} (times ϵ^2), while

$$\begin{aligned}\bar{v}_{ij} = \tfrac{1}{4}\big[&(p_{i+1,j+1} - 2p_{i+1,j} + p_{i+1,j-1}) \\ &+ 2(p_{i,j+1} - 2p_{i,j} + p_{i,j-1}) \\ &+ (p_{i-1,j+1} - 2p_{i-1,j} + p_{i-1,j-1})\big],\end{aligned}$$

$$\begin{aligned}\bar{h}_{ij} = \tfrac{1}{4}\big[&(q_{i+1,j+1} - 2q_{i,j+1} + q_{i-1,j+1}) \\ &+ 2(q_{i+1,j} - 2q_{i,j} + q_{i-1,j}) \\ &+ (q_{i+1,j-1} - 2q_{i,j-1} + q_{i-1,j-1})\big],\end{aligned}$$

are discrete estimates of the second partial derivatives p_{yy} and q_{xx} respectively (times ϵ^2). (Note again that the subscript i in the discrete version corresponds to x in the continuous case, while the subscript j corresponds to y.) Strat wrote his result in terms of various

[10] For simplicity, we assume a rectangular image region here. There is no loss of generality, however, since the sums can be taken over whatever region is desired.

[11] Note, however, that iterative schemes derived directly from the above will become unstable if λ is made small enough.

[12] Some of the terms are omitted when a point on the boundary is being considered.

intermediate expressions, so the equivalence to discrete estimates of partial derivatives was not apparent.

At this point we can isolate the terms in p_{ij} from one equation, and the term in q_{ij} from the other, if we let

$$\bar{v}_{ij} = \bar{p}_{ij} - p_{ij} \qquad \text{and} \qquad \bar{h}_{ij} = \bar{q}_{ij} - q_{ij},$$

where $\bar{p}_{ij}$ and $\bar{q}_{ij}$ are given by

$$\tfrac{1}{4}\left[(p_{i+1,j+1} - 2p_{i+1,j} + p_{i+1,j-1}) + 2(p_{i,j+1} + p_{i,j-1}) + (p_{i-1,j+1} - 2p_{i-1,j} + p_{i-1,j-1})\right],$$
$$\tfrac{1}{4}\left[(q_{i+1,j+1} - 2q_{i,j+1} + q_{i-1,j+1}) + 2(q_{i+1,j} + q_{i-1,j}) + (q_{i+1,j-1} - 2q_{i,j-1} + q_{i-1,j-1})\right],$$

respectively. In this way, we obtain

$$p_{ij} = \bar{p}_{ij} - \tilde{q}_{ij} + \frac{\epsilon^2}{\lambda}\big(E_{ij} - R(p_{ij}, q_{ij})\big) R_p(p_{ij}, q_{ij}),$$
$$q_{ij} = \bar{q}_{ij} - \tilde{p}_{ij} + \frac{\epsilon^2}{\lambda}\big(E_{ij} - R(p_{ij}, q_{ij})\big) R_q(p_{ij}, q_{ij}).$$

An iterative scheme can now be developed in which the terms p_{ij} and q_{ij} on the left-hand side of the equations are considered to be new values that are to be computed by inserting the current values into the right-hand sides. Then we obtain:

$$\begin{cases} p_{ij}^{k+1} = \bar{p}_{ij}^{k} - \tilde{q}_{ij}^{k} + \dfrac{\epsilon^2}{\lambda}\left(E_{ij} - R(p_{ij}^{k}, q_{ij}^{k})\right) R_p(p_{ij}^{k}, q_{ij}^{k}), \\ q_{ij}^{k+1} = \bar{q}_{ij}^{k} - \tilde{p}_{ij}^{k} + \dfrac{\epsilon^2}{\lambda}\left(E_{ij} - R(p_{ij}^{k}, q_{ij}^{k})\right) R_q(p_{ij}^{k}, q_{ij}^{k}). \end{cases}$$

This scheme appears to work reasonably well, having good stability and convergence properties.

It is clear that one has to do something special about the boundary, since the above result applies only for $1 < k < n$ and $1 < l < m$. On the boundary, different expressions apply, which can be obtained by carefully determining which of the terms are missing from the result of the initial differentiation. Put another way, the expressions for $\tilde{p}_{ij}$, $\tilde{q}_{ij}$, $\bar{v}_{ij}$, and $\bar{h}_{ij}$ require the old values of p_{ij} and q_{ij} at picture cells bordering on the region in which one is applying the iterative scheme. That is, before the scheme can be applied, p and q must be known on a border that is one picture cell wide.

Note that one cannot incorporate occluding boundary information in this scheme because, on the occluding boundary, at least one of p and q becomes unbounded. Strat, in fact, was forced in his examples to specify the gradient along some closed curve other than the occluding boundary. This kind of information is not usually available in applications of machine vision.

2.2. The method of Ikeuchi-Horn

Ikeuchi and Horn (1981) were the first to apply the calculus of variations to the shape-from-shading problem. They effectively solved a functional minimization problem in

recovering object surface orientation. It is known that the occluding boundary provides important constraints on the solution of the shape-from-shading problem (Bruss, 1983). The difficulty with using the gradient to specify surface orientation is that, as already mentioned, at least one of p and q is unbounded on the occluding boundary.

This problem can be overcome by specifying surface orientation in another way. Consider the mapping from pq space to fg space specified by the equations

$$f = \frac{2p}{1+\sqrt{1+p^2+q^2}} \qquad \text{and} \qquad g = \frac{2q}{1+\sqrt{1+p^2+q^2}}.$$

It is easy to verify that $f^2 + g^2 \leq 4$ for all visible parts of a surface. The orientation of a point on the occluding boundary corresponds to a point on a circle of radius two in fg space. Thus occluding boundaries present no difficulties now. The correspondence between pq space and the Gaussian sphere of possible orientations can be rationalized in terms of the gnomonic projection from the center of the sphere onto a tangent plane. Likewise, the correspondence between fg space and the Gaussian sphere can be thought of in terms of the stereographic projection from a point on the sphere onto a plane tangent to the sphere at the opposite point (see Ikeuchi and Horn, 1981).

We now seek appropriate f and g values at each point in the image. This we may regard as a search for two functions, f and g, defined over Ω, that correspond to a smooth surface satisfying the image irradiance equation

$$E(x,y) = R\big(f(x,y), g(x,y)\big).$$

(Note that the reflectance map here has been parameterized on f and g.)

We now develop an appropriate functional. Noting that f and g should ideally correspond to a surface that would produce the image if illuminated the same way as the actual surface, we adopt the integral of the brightness error

$$\iint_\Omega \Big(E(x,y) - R\big(f(x,y), g(x,y)\big)\Big)^2 \, dx\, dy.$$

We could, at this point, try to add a term that depends on the loop integrals, as Strat did. A problem with the use of stereographic coordinates is that the expression for the loop integrals becomes complicated. We have

$$p = \frac{4f}{4-f^2-g^2} \qquad \text{and} \qquad q = \frac{4g}{4-f^2-g^2},$$

so that $p_y - q_x = 0$ yields

$$\frac{f_y(4+f^2-g^2) - g_x(4-f^2+g^2) + 2(g_y - f_x)fg}{(4-f^2-g^2)^2} = 0.$$

This expression, even when multiplied by $(4-f^2-g^2)^2$, is quite complex and leads to even more complicated Euler equations.

Yet without additional constraint the problem is not well posed. As we saw earlier, the minimization of the total brightness error alone does not constitute a well-posed problem. In the above case we can choose, at each point (x,y), any f and g for which

$R(f,g) = E(x,y)$. In general, there is a one-dimensional family of possibilities—contours of constant R in fg space.

We would expect, however, that neighboring points have similar orientations, so that a typical "solution" of this form would not be reasonable. Ikeuchi and Horn decided to add the measure of "lack of smoothness" given by

$$\iint_\Omega \left(f_x^2 + f_y^2 + g_x^2 + g_y^2\right)\, dx\, dy.$$

A solution that produces a small value will be one that keeps the fluctuations in f and g small. Adding this term to the brightness error, we obtain the functional

$$\iint_\Omega \Big(E(x,y) - R\big(f(x,y), g(x,y)\big)\Big)^2 + \lambda \left(f_x^2 + f_y^2 + g_x^2 + g_y^2\right)\, dx\, dy$$

that is to be minimized by choosing f and g. Here, again, λ is a scalar that assigns a relative weighting to the terms.

The additional expression can be thought of as a regularization term[13]. Such a term can be added to a functional in order to obtain a solution in the case that a minimization problem does not have a unique solution.

The Euler equations for this minimization problem can be simplified to read

$$(E - R)R_f + \lambda \nabla^2 f = 0,$$
$$(E - R)R_g + \lambda \nabla^2 g = 0,$$

where R_f and R_g are the partial derivatives of $R(f,g)$ with respect to f and g and

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

is the Laplacian operator.

These Euler equations do not have a unique solution without additional constraint. The constraints available to us here are the values of f and g on the occluding boundary. Finding the solution of the Euler equations with this particular set of boundary conditions usually constitutes a well-posed problem—although this depends on the exact nature of the reflectance map, R, and the image, E.

At this point we introduce a discrete approximation of the Laplacian. The Laplacian of a function at a given point is approximately equal to a constant times the difference between a local average of the function and its value at the point. The factor of proportionality depends on the way in which the local average is computed. So, for example, if we use the simple finite-difference approximation

$$\{\nabla^2 f\}_{ij} \approx \frac{1}{\epsilon^2}\left[(f_{i,j+1} + f_{i+1,j} + f_{i,j-1} + f_{i-1,j}) - 4f_{i,j}\right],$$

then

$$\{\nabla^2 f\}_{ij} \approx \frac{4}{\epsilon^2}(\overline{f}_{ij} - f_{ij}),$$

[13] Ikeuchi and Horn, however, did not think of the extra term as a regularization term.

where the local average, $\overline{f}_{ij}$, is given by

$$\overline{f}_{ij} = \tfrac{1}{4}\left[f_{i,j+1} + f_{i+1,j} + f_{i,j-1} + f_{i-1,j}\right].$$

The same can be done for g, of course[14]. Using these finite-difference approximations in the Euler equations derived above, we obtain

$$f_{ij} = \overline{f}_{ij} + \frac{\epsilon^2}{4\lambda}\big(E_{ij} - R(f_{ij}, g_{ij})\big)R_f(f_{ij}, g_{ij}),$$

$$g_{ij} = \overline{g}_{ij} + \frac{\epsilon^2}{4\lambda}\big(E_{ij} - R(f_{ij}, g_{ij})\big)R_g(f_{ij}, g_{ij}).$$

where we have isolated the terms in f_{ij} and g_{ij}. An iterative scheme can now be developed in which these particular terms are considered to be new values to be computed by inserting the current values into the remainder of the expression. In this fashion, we finally arrive at the scheme

$$\begin{cases} f_{ij}^{k+1} = \overline{f}_{ij}^{k} + \dfrac{\epsilon^2}{4\lambda}\big(E_{ij} - R(f_{ij}^{k}, g_{ij}^{k})\big)R_f(f_{ij}^{k}, g_{ij}^{k}), \\ g_{ij}^{k+1} = \overline{g}_{ij}^{k} + \dfrac{\epsilon^2}{4\lambda}\big(E_{ij} - R(f_{ij}^{k}, g_{ij}^{k})\big)R_g(f_{ij}^{k}, g_{ij}^{k}). \end{cases}$$

Here, as before, ϵ denotes the spacing between picture cells, while $\overline{f}$ and $\overline{g}$ are the local averages of f and g.

This scheme appears to work reasonably well, having good stability and convergence properties. We shall see later, however, that the solutions for surface orientation may not correspond to an underlying smooth surface and that solutions may be distorted by the presence of the regularizing term. The degree of distortion depends on the parameter λ.

2.3. Smith's approach

Smith's method (Smith 1982) was derived by application of the standard calculus to the discrete domain. We now rationalize his method using the variational calculus. Surface orientation can be parameterized in lm space, where

$$l = -\frac{p}{\sqrt{1+p^2+q^2}} \qquad \text{and} \qquad m = -\frac{q}{\sqrt{1+p^2+q^2}}.$$

This corresponds to an orthographic projection of the Gaussian sphere onto a plane tangent to the sphere at one of the poles. We next adopt a regularizing term and minimize the functional

$$\iint_{\Omega} \big(E(x,y) - R\big(l(x,y), m(x,y)\big)\big)^2 + \lambda\left((\nabla^2 l)^2 + (\nabla^2 m)^2\right) dx\, dy.$$

[14] Slightly better results can be obtained using the nine-point approximation of the Laplacian rather than the five-point approximation shown here.

From the associated Euler equations, we obtain

$$(E - R)R_l - \lambda\nabla^4 l = 0,$$
$$(E - R)R_m - \lambda\nabla^4 m = 0,$$

where ∇^4 is the biharmonic operator[15].

We need, once again, to impose boundary conditions to avoid ambiguity in the solution. For the biharmonic equation we need to specify l and m on the boundary, as well as the normal derivatives of l and m. The *normal derivative* is the derivative in the direction of the outward normal to the boundary curve $\partial\Omega$. Note that while the values of l and m on the occluding boundary are known, it may not be obvious what the normal derivatives of l and m ought to be. Since they are not specified they must obey the appropriate natural boundary condition.

We can now use the simple finite-difference approximation

$$\{\nabla^4 l\}_{ij} \approx \frac{20}{\epsilon^4}\left[l_{ij} - \bar{l}_{ij}\right],$$

where

$$\begin{aligned}\bar{l}_{ij} = \tfrac{1}{20}\big[8(&l_{i+1,j} + l_{i-1,j} + l_{i,j+1} + l_{i,j-1}) \\ &- 2(l_{i+1,j+1} + l_{i-1,j-1} + l_{i-1,j+1} + l_{i+1,j-1}) \\ &- (l_{i+2,j} + l_{i-2,j} + l_{i,j+2} + l_{i,j-2})\big].\end{aligned}$$

The same can be done for m, of course[16]. Isolating the terms in l_{ij} and m_{ij}, we obtain

$$l_{ij} = \bar{l}_{ij} + \frac{\epsilon^4}{20\lambda}\big(E_{ij} - R(l_{ij}, m_{ij})\big)R_l(l_{ij}, m_{ij}),$$
$$m_{ij} = \bar{m}_{ij} + \frac{\epsilon^4}{20\lambda}\big(E_{ij} - R(l_{ij}, m_{ij})\big)R_m(l_{ij}, m_{ij}),$$

which leads to the iterative scheme

$$\begin{cases} l_{ij}^{k+1} = \bar{l}_{ij}^{k} + \dfrac{\epsilon^4}{20\lambda}\big(E_{ij} - R(l_{ij}^k, m_{ij}^k)\big)R_l(l_{ij}^k, m_{ij}^k), \\ m_{ij}^{k+1} = \bar{m}_{ij}^{k} + \dfrac{\epsilon^4}{20\lambda}\big(E_{ij} - R(l_{ij}^k, m_{ij}^k)\big)R_m(l_{ij}^k, m_{ij}^k). \end{cases}$$

The biharmonic equation and its variants are known to require careful treatment. In the iterative scheme as written above, for example, the new values are based directly on old values. This is called the Jacobi method, and is appropriate for parallel implementation. But this method is unstable for computational molecules that are discrete approximations of the biharmonic operator. A stable iteration can be achieved if one uses instead the Gauss-Seidel method, in which the computation of a new value at one picture

[15] The biharmonic operator can be defined in terms of the Laplacian operator by $\nabla^4(f) = \nabla^2(\nabla^2(f))$.

[16] Slightly better results can be obtained using the twentyfive-point approximation of the biharmonic operator instead of the thirteen-point approximation shown here.

cell uses the new values of those picture cells already visited in a raster scan of the image. This method, however, does not lend itself to parallel implementation. An alternative stabilizing technique depends on the use of smoothing between steps of a Jacobi iteration.

The more complex boundary conditions mentioned above are reflected in the fact that the computational molecule used as the discrete approximation of the biharmonic operator requires values for l and m in a band two picture cells wide bordering on the region in which the iterative scheme is applied. It is not enough to know the values of l and m on the occluding boundary[17].

Smith reported difficulties with the above scheme and incorrectly concluded that smoothness constraints fail to propagate boundary conditions by more than a few pixels in the image. In fact, by a suitable application of the aforementioned stabilization techniques, the scheme can be made to work. Note that fewer problems are encountered if $(l_x^2 + l_y^2 + m_x^2 + m_y^2)$ is used in the above functional as the regularization term. This is, in part, because the Euler equations then contain the Laplacian operator, for which simple iterative schemes exist that are well behaved; but mainly because the treatment of the boundary is simpler.

2.4. Depth from gradient

A use of the variational calculus in a subsiduary problem arises in the problem of recovering depth from the surface gradient. Let us suppose that we have determined surface orientation over the region Ω. The relative depth of surface points may be determined from the gradient (p, q) by means of the equality

$$\delta z = p\,\delta x + q\,\delta y,$$

that relates infinitesimal changes in x, y and z. Integrating along a curve C from (x_o, y_o) to (x, y), we obtain

$$z(x, y) = z(x_o, y_o) + \int_C (p\,dx + q\,dy).$$

This simple method of integration performs badly when the data are noisy. A depth value obtained at some point will, in these circumstances, depend on the integration path that was taken to get there.

It is better to find a best-fit surface z to the given components of the gradient, p and q. This we can accomplish by minimizing the functional

$$\iint_\Omega (z_x - p)^2 + (z_y - q)^2\,dx\,dy,$$

whose Euler equation reduces to

$$\nabla^2 z = p_x + q_y.$$

Once again, note that this equation does not uniquely specify a solution without further constraint. In fact, we can add any harmonic function[18] to a solution to obtain a different

[17] For further details, see for example the discussion of *molecular inhibition* by Terzopoulos (1983).
[18] A harmonic function satisfies Laplaces equation, $\nabla^2 z = 0$.

solution also satisfying the given Euler equation. In the case here there are no *a priori* boundary conditions given to us. That is, the function sought is not restrained on the boundary. The calculus of variations provides us in this situation with natural boundary conditions that must be satisfied by the solution. For this particular problem, the natural boundary conditions turn out to be (see Appendix)

$$(z_x, z_y) \cdot \mathbf{n} = (p, q) \cdot \mathbf{n},$$

where

$$\mathbf{n} = \left(-\frac{dy}{ds}, \frac{dx}{ds}\right)$$

is a normal vector to the boundary curve $\partial\Omega$ and s is arc-length along the boundary. So the component of (z_x, z_y) normal to the chosen boundary curve must match the normal component of (p, q)[19].

With these boundary conditions, the solution is still not quite unique, since an arbitrary constant can be added to z without changing the functional. This reflects the fact that one cannot recover absolute depth from the gradient (and thus from shading information). To get a particular answer, one can fix one of the depth values, or fix their average.

Using the discrete approximation to the Laplacian employed earlier, we obtain the iterative scheme

$$z_{ij}^{k+1} = \bar{z}_{ij}^{k} - \frac{\epsilon}{4}(h_{ij} + v_{ij}),$$

where

$$\bar{z}_{ij} = \tfrac{1}{4}(z_{i+1,j} + z_{i-1,j} + z_{i,j+1} + z_{i,j-1}),$$

is a local average of z, while

$$h_{ij} = \tfrac{1}{2}(p_{i+1,j} - p_{i-1,j}), \qquad \text{and} \qquad v_{ij} = \tfrac{1}{2}(q_{i,j+1} - q_{i,j-1}),$$

are estimates of the partial derivatives p_x and q_y respectively. This is as derived by Horn and reported by Ikeuchi (1983). (Note again that the subscript i in the discrete version corresponds to x in the continuous case, while the subscript j corresponds to y.)

In addition to finding the discrete approximation of the Euler equation, we also must find the discrete approximation of the boundary condition. This can be done easily, provided that the boundary curve is polygonal, with horizontal and vertical segments only. This restriction does not provide a problem in our simple situation. Now $z_x = p$ on vertical segments of the boundary, while $z_y = q$ on the horizontal segments. These conditions may be translated into

$$\frac{1}{2\epsilon}(z_{i+1,j} - z_{i-1,j}) = p_{ij},$$

$$\frac{1}{2\epsilon}(z_{i,j+1} - z_{i,j-1}) = q_{ij},$$

[19] Note that here z_x and z_y are the derivatives of the purported solution $z(x, y)$, while p and q are the given surface orientation data—only with perfect data would these be the same.

respectively. These relationships can be used to modify the computation of the average, $\overline{z}_{ij}$, for points on the edge of the region in which depth is to be reconstructed. Alternatively, these equations can be used to provide phantom depth values on a border of one picture cell width around that region. In this case the computation of the average can proceed in the same fashion for all points.

3. Smoothness and integrability

Methods that attempt to recover shape information encoded in an image usually confine their attention to *smooth*, or *piece-wise smooth*, solutions. Smoothness, however, is a loose term that may be interpreted in many ways. To be specific, we here define a graph, $z(x,y)$, to be smooth over a region Ω in the xy-plane if $p_y = q_x$, that is, if

$$\frac{\partial^2 z}{\partial x \partial y} = \frac{\partial^2 z}{\partial y \partial x} \qquad \forall (x,y) \in \Omega.$$

This is a property of C^2 surfaces[20]. Because they must be twice-differentiable under this definition, surfaces that have edges (like polyhedra) are excluded, and it may be argued that this accords with our intuitions on smoothness[21].

Let us now look more closely at the "lack-of-smoothness" term used in the Ikeuchi-Horn method. Suppose that we present a shape-from-shading problem to the program by providing it with an image, a reflectance map, and the occluding boundary. The image just happens to be that of a Lambertian sphere illuminated by an overhead point source at the viewer. This is a well-posed problem with two solutions, a concave bowl and a convex ball. It turns out that the algorithm will converge to a somewhat flattened sphere, given a planar initial estimate. Interestingly, it converges to almost the same solution when given the correct shape initially. That is, the algorithm moves away from the right answer. It is interesting to consider why this should be so.

Recall the functional that is used to derive the Ikeuchi-Horn method. We are required to minimize

$$\iint_\Omega \Big(E(x,y) - R\big(f(x,y), g(x,y)\big)\Big)^2 + \lambda \left(f_x^2 + f_y^2 + g_x^2 + g_y^2\right) dx\, dy.$$

It is clear that minimizing the integral of $\big(E - R(f,g)\big)^2$ is desirable; we wish to make the brightness error as small as possible. However, it is not obvious just what is achieved by minimizing the remainder of the overall integral. Certainly, it is not smoothness as

[20] For a discussion of this issue, within the context of the shape-from-shading problem, see Woodham (1981).

[21] Actually, one may claim that our intuition of smoothness accords better with a less restrictive definition. Consider, for example, adjoining a planar surface and a portion of a cylinder along a generator of the cylinder. This can be done in such a way that there is no discontinuity in surface orientation, and the surface may be considered to be "smooth" (Ikeuchi & Horn, 1981). The second derivatives, however, are discontinuous at the join. This suggest that we ought to identify our intuition of smoothness not with C^2, but PC^2, the set of surfaces that have piece-wise continuous second derivatives. In fact, some may even argue that surfaces in C^1, that is, those with continuous first derivatives, are "smooth." We ignore this subtle point in what follows, and restrict attention to surfaces in C^2.

defined above. In fact, if f and g are solutions to the Euler equations for this problem, it will in general be the case that there exists no physical surface corresponding exactly to the surface orientation specified by f and g (Brooks, 1982).

The expression $\left(f_x^2 + f_y^2 + g_x^2 + g_y^2\right)$ is instead best regarded as a regularizing term that is primarily intended as a means of finding a particularly smooth shape that is close to a solution of the original problem (Poggio & Torre, 1984). Different surfaces will give rise to different values for

$$\iint_\Omega \left(f_x^2 + f_y^2 + g_x^2 + g_y^2\right)\, dx\, dy.$$

Those that fluctuate in depth the least will likely give rise to small values. When a shape-from-shading problem is highly ambiguous, in that there is an infinite number of possible solutions, a regularizing term is precisely what is needed to get close to one of them. If, however, the problem is unambiguous, regularization will usually result in loss of accuracy, as the correct solution is unlikely to minimize the integral of the regularizing term.

The distortion due to regularization depends on the parameter λ. A large value of λ, appropriate when the image data is very noisy, leads to large errors, since the emphasis will be on producing as smooth a surface as possible, while permitting considerable error in brightness. Conversely, a small value for λ causes brightness errors to be weighted more. In this case, a more undulating surface is acceptable since the contribution of the regularizing term to the overall functional is relatively small[22].

4. Imposing integrability as a constraint

In any case, it is desirable to have a shape-from-shading method that neither moves away from correct solutions, nor converges to surfaces that are not solutions. To derive such a method, we need to impose the smoothness condition defined earlier, instead of using regularization. We first consider forcing the solution to satisfy the condition exactly.

Let us suppose that a shape-from-shading method recovers smooth functions $p(x, y)$ and $q(x, y)$ defined over the image, thereby specifying the gradient. In general, there will be no smooth surface that corresponds to this gradient. This is because the functions p and q must be related in a special way if they are to correspond to a smooth surface (Brooks, 1982). Noting once again that

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

it follows that, for our earlier definition of smoothness to be satisfied, we must have

$$\frac{\partial p}{\partial y}(x, y) = \frac{\partial q}{\partial x}(x, y),$$

or $z_{xy} = z_{yx}$. This is known as the constraint of *integrability*. If the gradient does not possess this property, there exists no C^2 surface that could give rise to it. Thus we shall now attempt to ensure that solutions are integrable.

[22] The iterative scheme becomes unstable, however, when the value of λ is reduced too much.

4.1. Direct recovery of relative depth

With the exception of the method of characteristic strips (Horn, 1975), all shape-from-shading programs have recovered surface orientation in a separate step, prior to recovering relative depth. We saw earlier an iterative scheme that determines depth values from the surface gradient. Of interest here is the direct recovery of depth information, achieved without the explicit manipulation of surface orientation.

Following the guidelines listed earlier, our initial task is to formulate an appropriate functional. The brightness error is readily expressed as

$$\iint_{\Omega} \Big(E(x,y) - R\big(z_x(x,y), z_y(x,y)\big)\Big)^2 \, dx\, dy.$$

Now we are to ensure the satisfaction of the constraint $z_{xy} = z_{yx}$. We might therefore consider adding the functional

$$\iint_{\Omega} (z_{xy} - z_{yx})^2 \, dx\, dy.$$

It is easy to veryfy that such a term makes no contribution to the subsequent Euler equation. This is because the integrand is a *divergence expression* (Courant & Hilbert, 1953)[23]. In our terms, by definition, we seek a smooth surface satisfying the partial differential equation. The integrability constraint is redundant. Put yet another way, we cannot avoid imposing the integrability constraint if we look for a scheme that gives us $z(x,y)$ directly. This was not the case when we used p and q as parameters. The functions p and q had to be related in a special way to satisfy integrability.

After simplification and reordering of terms, the Euler equation for the brightness-error functional alone is

$$\begin{aligned}(R_p^2 z_{xx} + 2R_p R_q z_{xy} + R_q^2 z_{yy}) - (E_x R_p + E_y R_q) \\ = (E - R)(R_{pp} z_{xx} + 2R_{pq} z_{xy} + R_{qq} z_{yy}),\end{aligned}$$

where we have used the condition $z_{yx} = z_{xy}$. Note that p and q replace z_x and z_y as subscripts of R to improve readability. A solution to this equation will give the functional an extremal value. By converting the Euler equation to discrete form, employing discrete approximations of the derivatives of z, and isolating terms in z_{ij}, we obtain the complex scheme

$$\begin{aligned}z_{ij}^{k+1} \big(R_p^2 + R_q^2 - (E - R)(R_{pp} + R_{qq})\big) \\ = \bar{h}_{ij}^k \big(R_p^2 - (E - R)R_{pp}\big) + \tilde{z}_{ij}^k \big(R_p R_q - (E - R)R_{pq}\big) + \bar{v}_{ij}^k \big(R_q^2 - (E - R)R_{qq}\big) \\ - \frac{\epsilon^2}{2}\big(E_x R_p + E_y R_q\big).\end{aligned}$$

where

$$\tilde{z}_{ij} = \tfrac{1}{4}\big(z_{i+1,j+1} + z_{i-1,j-1} - z_{i-1,j+1} - z_{i+1,j-1}\big),$$

[23] A divergence expression in the functional does, however, affect the natural boundary conditions.

is a discrete estimate of the cross derivative of z (times ϵ^2), and

$$\overline{h}_{ij} = \tfrac{1}{2}(z_{i+1,j} + z_{i-1,j}), \qquad \text{and} \qquad \overline{v}_{ij} = \tfrac{1}{2}(z_{i,j+1} + z_{i,j-1}),$$

are horizontal and vertical averages of z respectively. This scheme, unfortunately, is not convergent. Other schemes tried also failed. We found little in the literature about how one might discover successful iterative schemes for complicated non-linear equations such as the one above. Certainly, as far as the variational approach is concerned, the above Euler equation must be regarded as fundamental to the problem: the original functional is not easily formulated in a more basic way.

4.2. An alternative approach

Not surprisingly, if we parameterize the surface on p and q, and impose the integrability condition $p_y = q_x$, we obtain an Euler equation identical to the one obtained above. The functional to be minimized is in this case

$$\iint_\Omega \Big(E(x,y) - R\big(p(x,y), q(x,y)\big)\Big)^2 + \mu(x,y)\,(p_y - q_x)\,dx\,dy,$$

where μ is a Lagrangian multiplier used to enforce the constraint $p_y = q_x$. The associated Euler equations lead to

$$(E - R)R_p + \tfrac{1}{2}\mu_y = 0, \qquad \text{and} \qquad (E - R)R_q - \tfrac{1}{2}\mu_x = 0.$$

In order to eliminate μ, we take the (total) derivative of the first equation with respect to x and the (total) derivative of the second with respect to y. Adding the results we obtain

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

Taken together with the constraint $p_y = q_x$, this is the same result as that obtained in the previous section.

5. An integrability penalty term

It appears to be difficult to extract convergent iterative schemes from Euler equations obtained through the imposition of integrability. Consequently, we now assess the usefulness of the penalty term, $(p_y - q_x)^2$, appearing in the functional

$$\iint_\Omega \big(E(x,y) - R(p,q)\big)^2 + \lambda\big(p_y - q_x\big)^2\,dx\,dy.$$

This has the desirable property that if smooth functions $p(x,y)$ and $q(x,y)$ are found that cause this integral to evaluate to zero, we will, by definition, have solved our problem, for the surface will generate the image, and will be smooth everywhere[24].

[24] Note, however, that we now admit the possibility that p_y and q_x may be only approximately equal over the region Ω.

The Euler equations for this problem yield

$$(E - R)R_p + \lambda(p_{yy} - q_{xy}) = 0,$$
$$(E - R)R_q + \lambda(q_{xx} - p_{yx}) = 0.$$

Upon isolation of the center term in the discrete approximation of the highest-order, even partial derivatives, we arrive at the iterative scheme

$$\begin{cases} p_{ij}^{k+1} = \bar{p}_{ij}^{k} - \frac{1}{2}\widetilde{q}_{ij}^{k} + \dfrac{\epsilon^2}{2\lambda}\big(E_{ij} - R(p_{ij}^{k}, q_{ij}^{k})\big)R_q(p_{ij}^{k}, q_{ij}^{k}) \\ q_{ij}^{k+1} = \bar{q}_{ij}^{k} - \frac{1}{2}\widetilde{p}_{ij}^{k} + \dfrac{\epsilon^2}{2\lambda}\big(E_{ij} - R(p_{ij}^{k}, q_{ij}^{k})\big)R_p(p_{ij}^{k}, q_{ij}^{k}), \end{cases}$$

where

$$\bar{p}_{ij} = \tfrac{1}{2}(p_{i,j+1} + p_{i,j-1}) \qquad \text{and} \qquad \bar{q}_{ij} = \tfrac{1}{2}(q_{i+1,j} + q_{i-1,j})$$

are the vertical average of p and the horizontal average of q, respectively, while $\widetilde{p}_{ij}$ and $\widetilde{q}_{ij}$ are estimates of the cross derivatives (times ϵ^2) obtained using the approximations

$$\widetilde{p}_{ij} = \tfrac{1}{4}(p_{i+1,j+1} + p_{i-1,j-1} - p_{i-1,j+1} - p_{i+1,j-1}),$$
$$\widetilde{q}_{ij} = \tfrac{1}{4}(q_{i+1,j+1} + q_{i-1,j-1} - q_{i-1,j+1} - q_{i+1,j-1}),$$

respectively.

This iterative scheme appears to work well. Only very small departures from the correct initial solutions have been observed, these being due to the fact that the finite-difference expressions are approximations to derivatives. The scheme does not converge to a flattened surface as is the case with the Ikeuchi-Horn method. Rather, we obtain asymptotic convergence to the correct solution. Note once again, however, that this method requires that the gradient (p, q) be supplied on some closed curve other than the occluding boundary.

This iterative scheme produced very accurate results in tests conducted on synthetic images, although, like most shape-from-shading methods, it typically takes many iterations to converge. The observed slow convergence could be alleviated by the recently popularized multi-grid technique of processing images and gradient fields at various resolutions (Terzopoulos 1984).

It appears that the use of a penalty term based on a constraint leads to iterative schemes that adjust the present estimates in the direction that reduces the penalty term. This is in distinction to the behaviour of the schemes that result from attempts to strictly enforce the constraint itself. The use of the penalty term gives a scheme some directionality or "push" towards the desired solution. This may be why we were unsuccessful in discovering convergent iterative schemes based on the Euler equation derived in the previous section.

5.1. Relationship to Strat's scheme

It is interesting to observe how similar the iterative method we derived here is to that obtained by Strat. We can see in retrospect why this should be so, by applying Gauss's

integral formula to Strat's elementary loop integrals. We have

$$\oint_{\partial R} (p(x,y)\,dx + q(x,y)\,dy) = \iint_R \left(\frac{\partial q}{\partial x} - \frac{\partial p}{\partial y}\right) dx\,dy,$$

for a simply connected region R, where the boundary ∂R is traversed in a counter-clockwise direction.

Now, if c is constant in the region R, then

$$\left(\iint_R c\,dx\,dy\right)^2 = \iint_R dx\,dy \iint_R c^2\,dx\,dy = A(R) \iint_R c^2\,dx\,dy,$$

where $A(R)$ is the area of the region R. For a smooth surface, p_y and q_x are continuous, so that, for a small enough region R, we can consider them to be nearly constant. That is,

$$\left(\oint_{\partial R} (p(x,y)\,dx + q(x,y)\,dy)\right)^2 = \left(\iint_R (p_y - q_x)\,dx\,dy\right)^2 \approx A(R) \iint_R (p_y - q_x)^2\,dx\,dy.$$

So

$$e_{ij}^2 \approx \epsilon^2 \iint_{\delta R} (p_y - q_x)^2\,dx\,dy,$$

where δR is a square region with sides of length ϵ. Consequently, we can consider the sum of the error terms squared,

$$\left(\frac{\epsilon}{2}\right)^2 \sum_{i=1}^{n-1} \sum_{j=1}^{m-1} \left[p_{i,j} + p_{i+1,j} + q_{i+1,j} + q_{i+1,j+1} - p_{i+1,j+1} - p_{i,j+1} - q_{i,j+1} - q_{i,j}\right]^2,$$

or, written more suggestively,

$$\left(\frac{\epsilon}{2}\right)^2 \sum_{i=1}^{n-1} \sum_{j=1}^{m-1} \left[(p_{i,j+1} - p_{i,j}) + (p_{i+1,j+1} - p_{i+1,j}) - (q_{i+1,j} - q_{i,j}) - (q_{i+1,j+1} - q_{i,j+1})\right]^2,$$

to be a discrete approximation of

$$\epsilon^2 \iint_\Omega (p_y - q_x)^2\,dx\,dy.$$

Our final result in the previous section looks a little different from that of Strat, in part because we end up with simpler estimates for the second partial derivatives p_{yy} and q_{xx}.

5.2. Constraints and penalty terms

We have two equalities: the image irradiance equation, $E = R$, and the integrability condition, $p_y = q_x$. If we enforce both strictly, we obtain Horn's original characteristic strip equations. We have seen that a convergent iterative scheme can be obtained if we instead build a functional based on the penalty terms, $(E - R)^2$ and $(p_y - q_x)^2$. We also described our lack of success in deriving schemes for minimizing the integral of $(E - R)^2$

while enforcing the constraint $p_y = q_x$. We have not yet explored the fourth alternative of minimizing the integral of $(p_y - q_x)^2$ while enforcing the constraint $E = R$. That is, minimizing

$$\iint_\Omega \big(p_y(x,y) - q_x(x,y)\big)^2 + \mu(x,y)\ \big(E(x,y) - R\big(p(x,y), q(x,y)\big)\big)\ dx\,dy.$$

The resulting Euler equations are

$$(p_{yy} - q_{xy}) + \mu(E - R)R_p = 0,$$
$$(q_{xx} - p_{yx}) + \mu(E - R)R_q = 0,$$

which, upon elimination of μ lead to

$$(p_{yy} - q_{xy})R_p = (q_{xx} - p_{yx})R_q.$$

This equation is to be solved subject to the constraint $E = R$, of course. We were unable to convince ourselves of the utility of pursuit of this particular approach, since we know that brightness measurements will be corrupted by noise in practice.

6. Incorporating occluding boundary information

One problem not easily coped with is that of dealing with the occluding boundary. Recall that the Ikeuchi-Horn method placed considerable emphasis on the ability to be able to handle the occluding boundary. So, although we have taken a step forward in the above by incorporating integrability, we have also taken a step backwards in that we are no longer able to use the occluding boundary. Note, however, that the integrability constraint can be expressed using parameterizations that permit incorporation of the occluding boundary information.

Suppose that instead of seeking surface orientation parameterized on $p(x,y)$ and $q(x,y)$, we attempt to recover directly a field of unit normal vectors $\mathbf{n}(x,y)$. We need to express the integrability constraint in terms of the unit normal and its derivatives. Let $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ denote unit vectors in the x, y and z directions, respectively. We have that

$$p = -\frac{\mathbf{n}\cdot\hat{\mathbf{x}}}{\mathbf{n}\cdot\hat{\mathbf{z}}} \qquad \text{and} \qquad q = -\frac{\mathbf{n}\cdot\hat{\mathbf{y}}}{\mathbf{n}\cdot\hat{\mathbf{z}}},$$

so it follows that

$$p_y = \frac{(\mathbf{n}\cdot\hat{\mathbf{x}})(\mathbf{n}_y\cdot\hat{\mathbf{z}}) - (\mathbf{n}\cdot\hat{\mathbf{z}})(\mathbf{n}_y\cdot\hat{\mathbf{x}})}{(\mathbf{n}\cdot\hat{\mathbf{z}})^2} = \frac{\mathbf{n}\cdot\big((\mathbf{n}_y\cdot\hat{\mathbf{z}})\hat{\mathbf{x}} - (\mathbf{n}_y\cdot\hat{\mathbf{x}})\hat{\mathbf{z}}\big)}{(\mathbf{n}\cdot\hat{\mathbf{z}})^2} = \frac{\mathbf{n}\cdot(\mathbf{n}_y\times(\hat{\mathbf{x}}\times\hat{\mathbf{z}}))}{(\mathbf{n}\cdot\hat{\mathbf{z}})^2},$$

using the identity $(\mathbf{c}\cdot\mathbf{a})\mathbf{b} - (\mathbf{a}\cdot\mathbf{b})\mathbf{c} = \mathbf{a}\times(\mathbf{b}\times\mathbf{c})$. Noting that $\hat{\mathbf{x}}\times\hat{\mathbf{z}} = -\hat{\mathbf{y}}$ we obtain[25]

$$p_y = -[\mathbf{n}\,\mathbf{n}_y\,\hat{\mathbf{y}}]/(\mathbf{n}\cdot\hat{\mathbf{z}})^2.$$

Similarly,

$$q_x = +[\mathbf{n}\,\mathbf{n}_x\,\hat{\mathbf{x}}]/(\mathbf{n}\cdot\hat{\mathbf{z}})^2.$$

[25] Here $[\mathbf{a}\,\mathbf{b}\,\mathbf{c}]$ denotes the vector triple product $\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})$.

We conclude that the constraint $(p_y - q_x) = 0$ can be written in the form

$$\frac{1}{(\mathbf{n}\cdot\hat{\mathbf{z}})^2}\left([\mathbf{n}\,\mathbf{n}_x\,\hat{\mathbf{x}}] + [\mathbf{n}\,\mathbf{n}_y\,\hat{\mathbf{y}}]\right) = 0.$$

As it stands, this form of the constraint will lead to numerical problems in the implementation of an iterative scheme, since $(\mathbf{n}\cdot\hat{\mathbf{z}})$ becomes very small near the occluding boundary. It makes sense then to use instead a constraint obtained by multiplying the one above by $(\mathbf{n}\cdot\hat{\mathbf{z}})^2$, giving

$$I' = [\mathbf{n}\,\mathbf{n}_x\,\hat{\mathbf{x}}] + [\mathbf{n}\,\mathbf{n}_y\,\hat{\mathbf{y}}].$$

One could, of course, tackle this problem using other parametrizations for surface orientation, such as f and g. We saw earlier that the integrability constraint expressed in terms of f and g is quite complex, and the derivation of the corresponding Euler equations somewhat tedious. We felt that the compactness of vector notation provided sufficient incentive to tackle the problem the way we did. There is an advantage to using f and g, however: one can avoid the redundancy inherent in the use of a vector to represent surface orientation, a quantity that has only two degrees of freedom. It is this redundancy that leads us to consideration of the pseudo-inverse of a matrix later on.

6.1. Using a penalty term based on I'

We are to minimize a functional of the form

$$\iint_\Omega \Big(E(x,y) - R\big(\mathbf{n}(x,y)\big)\Big)^2 + \lambda I'^2 + \mu(x,y)\,(\mathbf{n}^2 - 1)\,dx\,dy.$$

Here we use the Lagrangian multiplier, μ, to enforce the constraint $\mathbf{n}^2 = 1$. The corresponding Euler equation can be simplified to read

$$-(E-R)R_{\mathbf{n}} + 2\lambda I' I'_{\mathbf{n}} + \mu\mathbf{n} + \lambda\big(I'_x(\mathbf{n}\times\hat{\mathbf{x}}) + I'_y(\mathbf{n}\times\hat{\mathbf{y}})\big) = 0,$$

where

$$I'_{\mathbf{n}} = \mathbf{n}_x\times\hat{\mathbf{x}} + \mathbf{n}_y\times\hat{\mathbf{y}}$$

is the derivative of I' with respect to $\mathbf{n}$, while

$$I'_x = [\mathbf{n}\,\mathbf{n}_{xx}\,\hat{\mathbf{x}}] + [\mathbf{n}_x\,\mathbf{n}_y\,\hat{\mathbf{y}}] + [\mathbf{n}\,\mathbf{n}_{yx}\,\hat{\mathbf{y}}],$$
$$I'_y = [\mathbf{n}\,\mathbf{n}_{yy}\,\hat{\mathbf{y}}] + [\mathbf{n}_y\,\mathbf{n}_x\,\hat{\mathbf{x}}] + [\mathbf{n}\,\mathbf{n}_{xy}\,\hat{\mathbf{x}}],$$

are the derivatives of I' with respect to x and y respectively. We can find the Lagrangian multiplier μ by taking the dot product of the Euler equation with $\mathbf{n}$, to give

$$\mu = (E-R)R_{\mathbf{n}}\cdot\mathbf{n} - 2\lambda I'^2,$$

where we use the fact that $I'_{\mathbf{n}}\cdot\mathbf{n} = I'$. We can now eliminate μ by substituting back into the Euler equation. The result is

$$-(E-R)R_{\mathbf{n}}^{\perp} + 2\lambda I'\mathbf{j}' + \lambda\big(I'_x(\mathbf{n}\times\hat{\mathbf{x}}) + I'_y(\mathbf{n}\times\hat{\mathbf{y}})\big) = 0,$$

where

$$R_{\mathbf{n}}^{\perp} = R_{\mathbf{n}} - (R_{\mathbf{n}} \cdot \mathbf{n})\mathbf{n} = \mathbf{n} \times (R_{\mathbf{n}} \times \mathbf{n})$$

is the component of $R_{\mathbf{n}}$ perpendicular to $\mathbf{n}$ and

$$\mathbf{j}' = I'_{\mathbf{n}} - I'\mathbf{n}.$$

Note that $\mathbf{j}' \cdot \mathbf{n} = 0$, since $I'_{\mathbf{n}} \cdot \mathbf{n} = I'$. In fact, each term in the above equation is orthogonal to $\mathbf{n}$. This vector equation thus provides only two constraints on $\mathbf{n}$. The necessary third constraint is given by $\mathbf{n}^2 = 1$.

Now let

$$J_x = (\mathbf{n}_x \times \mathbf{n}_y + \mathbf{n} \times \mathbf{n}_{yx}) \cdot \hat{\mathbf{y}} \qquad \text{and} \qquad J_y = (\mathbf{n}_y \times \mathbf{n}_x + \mathbf{n} \times \mathbf{n}_{xy}) \cdot \hat{\mathbf{x}}.$$

Then

$$I'_x = [\mathbf{n}\, \mathbf{n}_{xx}\, \hat{\mathbf{x}}] + J_x \qquad \text{and} \qquad I'_y = [\mathbf{n}\, \mathbf{n}_{yy}\, \hat{\mathbf{y}}] + J_y,$$

and the Euler equation can be rewritten in the form

$$-(E - R)R_{\mathbf{n}}^{\perp} + 2\lambda I' \mathbf{j}' + \lambda \mathbf{l} = \lambda \left[(\mathbf{n} \times \hat{\mathbf{x}})(\mathbf{n} \times \hat{\mathbf{x}})^T \mathbf{n}_{xx} + (\mathbf{n} \times \hat{\mathbf{y}})(\mathbf{n} \times \hat{\mathbf{y}})^T \mathbf{n}_{yy} \right],$$

where

$$\mathbf{l} = J_x(\mathbf{n} \times \hat{\mathbf{x}}) + J_y(\mathbf{n} \times \hat{\mathbf{y}}).$$

So we can write

$$\lambda \left(\mathbf{M}_x \mathbf{n}_{xx} + \mathbf{M}_y \mathbf{n}_{yy} \right) = \lambda \mathbf{l} + 2\lambda I' \mathbf{j}' - (E - R)R_{\mathbf{n}}^{\perp},$$

where

$$\mathbf{M}_x = (\mathbf{n} \times \hat{\mathbf{x}})(\mathbf{n} \times \hat{\mathbf{x}})^T \qquad \text{and} \qquad \mathbf{M}_y = (\mathbf{n} \times \hat{\mathbf{y}})(\mathbf{n} \times \hat{\mathbf{y}})^T$$

are the so-called *dyadic products* of the vectors $(\mathbf{n} \times \hat{\mathbf{x}})$ and $(\mathbf{n} \times \hat{\mathbf{y}})$ with themselves[26]. We now have a non-linear second order partial differential equation for the normal $\mathbf{n}(x, y)$.

We can use the following finite-difference approximations for the derivatives that appear:

$$\mathbf{n}_x \approx \frac{1}{2\epsilon} \left(\mathbf{n}_{i+1,j} - \mathbf{n}_{i-1,j} \right) \quad \text{and} \quad \mathbf{n}_y \approx \frac{1}{2\epsilon} \left(\mathbf{n}_{i,j+1} - \mathbf{n}_{i,j-1} \right),$$

$$\mathbf{n}_{xy} \approx \frac{1}{4\epsilon^2} \left(\mathbf{n}_{i+1,j+1} + \mathbf{n}_{i-1,j-1} - \mathbf{n}_{i-1,j+1} - \mathbf{n}_{i+1,j-1} \right),$$

as well as

$$\mathbf{n}_{xx} \approx \frac{1}{\epsilon^2} \left(\mathbf{n}_{i+1,j} - 2\mathbf{n}_{ij} + \mathbf{n}_{i-1,j} \right) \quad \text{and} \quad \mathbf{n}_{yy} \approx \frac{1}{\epsilon^2} \left(\mathbf{n}_{i,j+1} - 2\mathbf{n}_{ij} + \mathbf{n}_{i,j-1} \right),$$

or

$$\mathbf{n}_{xx} \approx \frac{2}{\epsilon^2} \left(\bar{\mathbf{h}}_{ij} - \mathbf{n}_{ij} \right) \qquad \text{and} \qquad \mathbf{n}_{yy} \approx \frac{2}{\epsilon^2} \left(\bar{\mathbf{v}}_{ij} - \mathbf{n}_{ij} \right),$$

where

$$\bar{\mathbf{h}}_{ij} = \tfrac{1}{2} \left(\mathbf{n}_{i+1,j} + \mathbf{n}_{i-1,j} \right) \qquad \text{and} \qquad \bar{\mathbf{v}}_{ij} = \tfrac{1}{2} \left(\mathbf{n}_{i,j+1} + \mathbf{n}_{i,j-1} \right)$$

[26] These dyadic products are matrices of rank one.

are horizontal and vertical averages of $\mathbf{n}$ respectively.

We now develop an iterative scheme based on the isolation of the center term in the discrete approximations of the highest-order, even partial derivatives. For convenience, let $\mathbf{m}_{ij}$, say, be the new value of the normal to be calculated in the iterative step. Then

$$\mathbf{n}_{xx} \approx \frac{2}{\epsilon^2}\left(\bar{\mathbf{h}}_{ij} - \mathbf{m}_{ij}\right) \qquad \text{and} \qquad \mathbf{n}_{yy} \approx \frac{2}{\epsilon^2}\left(\bar{\mathbf{v}}_{ij} - \mathbf{m}_{ij}\right).$$

If we let $\mathbf{M} = \mathbf{M}_x + \mathbf{M}_y$, then the new value is obtained using the equation

$$\mathbf{M}\,\mathbf{m} = \left(\mathbf{M}_x\bar{\mathbf{h}} + \mathbf{M}_y\bar{\mathbf{v}}\right) - \frac{\epsilon^2}{2}\mathbf{l} - \epsilon^2 I'\,\mathbf{j}' + \frac{\epsilon^2}{2\lambda}(E - R)R_{\mathbf{n}}^{\perp},$$

and the constraint $\mathbf{m}^2 = 1$. Here we omit subscripts in order to simplify the notation. Let $\mathbf{r}$ denote the right hand side of the equation above. All terms in $\mathbf{r}$ can be easily computed using the old estimate of the normal, $\mathbf{n}$, in the expressions for $\mathbf{M}_x$, $\mathbf{M}_y$, $\bar{\mathbf{h}}$, $\bar{\mathbf{v}}$, $\mathbf{l}$, $\mathbf{j}'$, I', R and $R_{\mathbf{n}}^{\perp}$. The remaining problem is the solution of the equation for the new estimate of the normal, $\mathbf{m}$.

6.2. Solving the equations $\mathbf{M}\,\mathbf{m} = \mathbf{r}$ and $\mathbf{m}\cdot\mathbf{m} = 1$

The equation $\mathbf{M}\,\mathbf{m} = \mathbf{r}$ is underdetermined, since $\mathbf{M}$ here only has rank two. The matrix is singular and so does not have an inverse in the usual sense. There are an infinite number of solutions that can be written in terms of the *pseudo-inverse*, $\mathbf{M}^+$. They are

$$\mathbf{m} = \mathbf{M}^+\mathbf{r} + \left(\mathbf{I} - \mathbf{M}^+\mathbf{M}\right)\mathbf{x},$$

for arbitrary $\mathbf{x}$ (Albert, 1982), where $\mathbf{I}$ is the 3×3 identity matrix. Of these solutions, we seek the one with unit norm, $\mathbf{m}^2 = 1$.

The pseudo-inverse of a matrix $\mathbf{M}$ can be defined as the limit

$$\mathbf{M}^+ = \lim_{\delta \to 0}\left(\mathbf{M}^T\mathbf{M} + \delta^2\mathbf{I}\right)^{-1}\mathbf{M}^T.$$

Alternatively, it can be defined using the conditions of Penrose (Albert, 1982), which state that the matrix $\mathbf{M}^+$ is the pseudo-inverse of the matrix $\mathbf{M}$ if, and only if,

- $\mathbf{M}\mathbf{M}^+$ and $\mathbf{M}^+\mathbf{M}$ are symmetric, and
- $\mathbf{M}^+\mathbf{M}\mathbf{M}^+ = \mathbf{M}$, as well as,
- $\mathbf{M}\mathbf{M}^+\mathbf{M} = \mathbf{M}^+$.

The pseudo-inverse may also be found using spectral decomposition. The eigenvectors of the pseudo-inverse are the same as those of the original matrix, while the corresponding non-zero eigenvalues are the inverses of the non-zero eigenvalues of the original matrix.

Now, in our case,

$$\mathbf{M} = (\mathbf{n} \times \hat{\mathbf{x}})(\mathbf{n} \times \hat{\mathbf{x}})^T + (\mathbf{n} \times \hat{\mathbf{y}})(\mathbf{n} \times \hat{\mathbf{y}})^T,$$

so one can show that

$$\mathbf{M}^+ = \mathbf{I} - \mathbf{n}\mathbf{n}^T + \frac{1}{(\mathbf{n}\cdot\hat{\mathbf{z}})^2}(\mathbf{n} \times \hat{\mathbf{z}})(\mathbf{n} \times \hat{\mathbf{z}})^T.$$

It is also possible to verify that

$$\mathbf{M}^{+}\mathbf{M} = \mathbf{I} - \mathbf{n}\mathbf{n}^{T},$$

from which it follows that

$$\mathbf{I} - \mathbf{M}^{+}\mathbf{M} = \mathbf{n}\mathbf{n}^{T},$$

and so

$$\mathbf{m} = \mathbf{M}^{+}\mathbf{r} + \nu\mathbf{n},$$

for some ν chosen to make $\mathbf{m}^2 = 1$. In our case $\mathbf{r} \perp \mathbf{n}$, so we can further simplify matters by noting that

$$\mathbf{M}^{+}\mathbf{r} = \mathbf{r} + \frac{1}{(\mathbf{n}\cdot\hat{\mathbf{z}})^2}\,[\mathbf{r}\,\mathbf{n}\,\hat{\mathbf{z}}]\,(\mathbf{n}\times\hat{\mathbf{z}}).$$

Let this be called $\mathbf{p}$. Since $\mathbf{p} \perp \mathbf{n}$, we have that $\mathbf{m}^2 = \mathbf{p}^2 + \nu^2$. This completes the calculation of the new estimate of the normal, $\mathbf{m}$. The only potential problem occurs when $|\mathbf{M}^{+}\mathbf{r}| > 1$, as may happen when the current estimate of the solution is far from the correct one. In this case it is advisable to limit the adjustment of the local normal away from its previous value, $\mathbf{n}$[27].

6.3. Using a penalty term based on I

Implementations of the above iterative scheme work well except for minor problems near the occluding boundary. What happens is that the components of $\mathbf{n}_x$ and $\mathbf{n}_y$ become unbounded on the occluding boundary, so that the individual terms in

$$I' = [\mathbf{n}\,\mathbf{n}_x\,\hat{\mathbf{x}}] + [\mathbf{n}\,\mathbf{n}_y\,\hat{\mathbf{y}}]$$

tend to become very large[28]. It may be better to use the slightly more complicated expression

$$I = (\mathbf{n}\cdot\hat{\mathbf{z}})I' = (\mathbf{n}\cdot\hat{\mathbf{z}})\left([\mathbf{n}\,\mathbf{n}_x\,\hat{\mathbf{x}}] + [\mathbf{n}\,\mathbf{n}_y\,\hat{\mathbf{y}}]\right).$$

This can be viewed as the difference of two quantities that remain bounded, provided that the curvature of the surface is bounded.

We now are to minimize a functional of the form

$$\iint_{\Omega}\Big(E(x,y) - R\big(\mathbf{n}(x,y)\big)\Big)^2 + \lambda I^2 + \mu(x,y)\,(\mathbf{n}^2 - 1)\,dx\,dy.$$

The corresponding Euler equation can be simplified to read

$$-(E-R)R_{\mathbf{n}} + \lambda I\big(I'\hat{\mathbf{z}} + 2(\mathbf{n}\cdot\hat{\mathbf{z}})I'_{\mathbf{n}}\big) + \mu\mathbf{n} + 2\lambda I\mathbf{k} + \lambda(\mathbf{n}\cdot\hat{\mathbf{z}})^2\big(I'_x(\mathbf{n}\times\hat{\mathbf{x}}) + I'_y(\mathbf{n}\times\hat{\mathbf{y}})\big) = 0,$$

[27] Also, note that there are theoretically two solutions for ν, one positive and one negative. The positive value leads to a new estimate close to the previous one, while the negative value gives rise to one almost opposite to the old one. It is clear that one should use the positive root.

[28] The problem is different, of course, from the one we encountered earlier when using the gradient to parameterize surface orientation. The components of the gradient, p and q become infinite on the boundary, while $\mathbf{n}$ remains finite.

where

$$\mathbf{k} = (\mathbf{n}_x \cdot \hat{\mathbf{z}})(\mathbf{n} \times \hat{\mathbf{x}}) + (\mathbf{n}_y \cdot \hat{\mathbf{z}})(\mathbf{n} \times \hat{\mathbf{y}}),$$

and $I'_{\mathbf{n}}$, I'_x and I'_y are as defined before.

We can find the Lagrangian multiplier μ by taking the dot product of the Euler equation with $\mathbf{n}$. Thus we have

$$\mu = (E - R)R_{\mathbf{n}} \cdot \mathbf{n} - 3\lambda I^2.$$

Now we eliminate μ by substituting back into the Euler equation. The result is

$$-(E - R)R_{\mathbf{n}}^{\perp} + \lambda I\mathbf{j} + 2\lambda I\mathbf{k} + \lambda(\mathbf{n} \cdot \hat{\mathbf{z}})^2\big(I'_x(\mathbf{n} \times \hat{\mathbf{x}}) + I'_y(\mathbf{n} \times \hat{\mathbf{y}})\big) = 0,$$

where

$$\mathbf{j} = I'\hat{\mathbf{z}} + 2(\mathbf{n} \cdot \hat{\mathbf{z}})I'_{\mathbf{n}} - 3I\mathbf{n},$$

and $R_{\mathbf{n}}^{\perp}$ is the component of $R_{\mathbf{n}}$ perpendicular to $\mathbf{n}$ as before. Now $I'_{\mathbf{n}} \cdot \mathbf{n} = I'$, so $\mathbf{j} \cdot \mathbf{n} = 0$. In fact, each term in the above equation is orthogonal to $\mathbf{n}$. This vector equation thus provides only two constraints on $\mathbf{n}$. The necessary third constraint is again given by $\mathbf{n}^2 = 1$.

Now let J_x and J_y be defined as before. Then the Euler equation can be rewritten in the form

$$\begin{aligned} -(E - R)R_{\mathbf{n}}^{\perp} + \lambda I\mathbf{j} + 2\lambda I\mathbf{k} + \lambda(\mathbf{n} \cdot \hat{\mathbf{z}})^2\mathbf{l} \\ = \lambda(\mathbf{n} \cdot \hat{\mathbf{z}})^2 \left[(\mathbf{n} \times \hat{\mathbf{x}})(\mathbf{n} \times \hat{\mathbf{x}})^T\mathbf{n}_{xx} + (\mathbf{n} \times \hat{\mathbf{y}})(\mathbf{n} \times \hat{\mathbf{y}})^T\mathbf{n}_{yy}\right], \end{aligned}$$

where $\mathbf{l} = J_x(\mathbf{n} \times \hat{\mathbf{x}}) + J_y(\mathbf{n} \times \hat{\mathbf{y}})$. So we obtain

$$\lambda(\mathbf{n} \cdot \hat{\mathbf{z}})^2 (\mathbf{M}_x\mathbf{n}_{xx} + \mathbf{M}_y\mathbf{n}_{yy}) = \lambda(\mathbf{n} \cdot \hat{\mathbf{z}})^2\mathbf{l} + 2\lambda I\mathbf{k} + \lambda I\mathbf{j} - (E - R)R_{\mathbf{n}}^{\perp},$$

where $\mathbf{M}_x$ and $\mathbf{M}_y$ are defined as before. We now have a non-linear second order partial differential equation for the normal $\mathbf{n}(x, y)$.

Note that both sides of this equation are orthogonal to $\mathbf{n}$, since $\mathbf{l} \cdot \mathbf{n} = 0$, $\mathbf{k} \cdot \mathbf{n} = 0$, $\mathbf{j} \cdot \mathbf{n} = 0$, and $R_{\mathbf{n}}^{\perp} \cdot \mathbf{n} = 0$. So the equation provides two constraints only, with the third coming from $\mathbf{n}^2 = 1$.

If we use the same discrete approximations as before, and isolate the central value in the finite-difference approximations of the highest order even partial derivatives, we obtain

$$\mathbf{M}\,\mathbf{m} = \big(\mathbf{M}_x\bar{\mathbf{h}} + \mathbf{M}_y\bar{\mathbf{v}}\big) - \frac{\epsilon^2}{2}\mathbf{l} - \frac{\epsilon^2}{(\mathbf{n} \cdot \hat{\mathbf{z}})}I'\mathbf{k} - \frac{\epsilon^2}{2(\mathbf{n} \cdot \hat{\mathbf{z}})}I'\mathbf{j} + \frac{\epsilon^2}{2\lambda(\mathbf{n} \cdot \hat{\mathbf{z}})^2}(E - R)R_{\mathbf{n}}^{\perp}.$$

We once again obtain an underdetermined equation, of the form $\mathbf{M}\,\mathbf{m} = \mathbf{r}$, together with a constraint $\mathbf{m}^2 = 1$. We can solve for the new estimate of the surface normal using the pseudo-inverse of the matrix $\mathbf{M}$, as before.

It is curious that several of the terms involve division by $(\mathbf{n} \cdot \hat{\mathbf{z}})$, a term that becomes large near the occluding boundary. We multiplied the penalty term by this expression in the first place in order to avoid problems near the occluding boundary. Apparently,

however, the terms so affected are all small near the occluding boundary anyway. In fact, we determined experimentally that several of the terms on the right hand side are very small compared to the others, particularly as one approaches the correct solution. We found that one can leave them out without noticably affecting convergence, or the surface arrived at ultimately. Preliminary testing of the scheme on synthetic images yielded promising results. Comprehensive assessment of the performance of the two schemes has, however, been left for future work.

7. Summary

The shape-from-shading problem was regarded here as one of finding a surface that minimizes an integral expression involving the brightness error. The expression we used has the form of a functional measuring the departure of a hypothesized surface from a solution surface. Iterative schemes for solving the shape-from-shading problem were based on the appropriate Euler equation.

We reviewed the use of a regularization term in an existing iterative scheme. Regularization techniques allow one to obtain results when faced with ill-posed problems. We argued, however, that the addition of a regularization term is not appropriate when one is dealing with a well-posed problem. The additional term tends to flatten and distort the solution.

We next discussed the fact that surface orientation must satisfy an integrability constraint if it is to correspond to an underlying smooth surface. The method using the regularization term does not guarantee this. We attempted to use the integrability constraint instead of a regularization term, but failed to find convergent iterative schemes for solving the resulting Euler equations.

A convergent iterative scheme was obtained, however, when, instead of enforcing integrability, we introduced a penalty term derived from the integrability constraint. It seems that the penalty term provides the iterative process with a "sense of direction" that helps it head towards the solution. This approach allows one to recover surface gradients that are approximately integrable. The scheme so derived was shown to be similar to that obtained in the discrete domain by Strat. A drawback of his scheme is its inability to incorporate occluding boundary information.

We overcame this difficulty by employing a different parametrization for surface orientation. The integrability penalty term can be expressed in terms of the unit surface normal and its derivatives. Subsequent application of the variational calculus proved to be somewhat involved, but two usable iterative schemes were finally obtained. Initial tests indicate that they perform well. Our new schemes are the first to make use of the integrability constraint while allowing incorporation of the occluding boundary normals. Future work will assess the relative performance of the new schemes in detail.

Acknowledgements

We thank Alan Yuille for many useful discussions. Comments on a draft by Steve Bagley, Eric Grimson, Mike Brady and, especially, Demetri Terzopoulos were much appreciated. Mike Brooks wishes to express gratitude to the MIT AI laboratory for the opportunity to

spend some months there as a visiting scientist, and to Flinders University for granting leave.

References

Albert, A., *Regression and the Moore-Penrose Pseudoinverse*, Academic Press, New York, 1972.

Brady, J.M., & Horn, B.K.P., "Rotationally Symmetric Operators for Surface Interpolation," *Computer Vision, Graphics, and Image Processing,* Vol. 22, pp. 70–95, 1983.

Brooks, M.J., "Shape from Shading Discretely," Ph.D thesis, Essex University, 1982.

Brooks, M.J., "Surface Normals from Closed Paths," *Proceedings of the International Joint Conference on Artificial Intelligence,* Tokyo, 1979.

Bruss, A.R., "Is What You See What You Get?", *Proceedings of the International Joint Conference on Artificial Intelligence*, Karlsruhe, August, 1983.

Courant, R. & Hilbert, D., *Methods of Mathematical Physics*, Vol. 1, Interscience Publishers, Inc., New York, 1953.

Horn, B.K.P., "Shape-from-Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from one View," MAC-TR-79 and also AI-TR-232, Artificial Intelligence Laboratory, M.I.T., November 1970.

Horn, B.K.P., "Obtaining Shape from Shading Information," in *The Psychology of Computer Vision,* P.H.Winston (Ed.), McGraw-Hill, 1975.

Horn, B.K.P., "Understanding Image Intensities," *Artificial Intelligence*, 1977, Vol. 8, No. 2, pp. 201–231.

Horn, B.K.P. & Sjoberg, R.W., "Calculating the Reflectance Map," *Applied Optics*, Vol. 18, No. 11, June 1979, pp. 1770–1779.

Ikeuchi, K., "Constructing a Depth Map from Images," A.I.M. 744, Artificial Intelligence Laboratory, M.I.T., August, 1983.

Ikeuchi, K. & Horn, B.K.P., "Numerical Shape from Shading and Occluding Boundaries," *Artificial Intelligence*, Vol. 17, 1981, pp. 141–185.

Poggio, T. & Torre, V., "Ill-posed Problems and Regularization Analysis in Early Vision," A.I.M. 773, Artificial Intelligence Laboratory, M.I.T., 1984.

Smith, G., "The Recovery of Surface Orientation from Image Irradiance," *Proceedings of the DARPA Image Understanding Workshop*, Palo Alto, September 1982.

Strat, T.M., "A Numerical Method for Shape from Shading from a Single Image," M.S. thesis, Department of E.E. & C.S., M.I.T., 1979.

Terzopoulos, D., "The Role of Constraints and Discontinuities in Visible-Surface Reconstruction," *Proceedings of the International Joint Conference on Artificial Intelligence*, Karlsruhe, August, 1983.

Terzopoulos, D., "Efficient Multiresolution Algorithms for Computing Lightness, Shape from Shading, and Optical Flow," *Proc. of the Fourth National Conference on Artificial Intelligence*, University of Texas, Austin, Texas, 1984.

Woodham, R.J., "Analysing Images of Curved Surfaces," in *Computer Vision*, J.M. Brady (Ed.), North-Holland, 1981, pp. 117–140.

Appendix

A. Minimizing Functionals

Let $F(x, y, z, z_x, z_y)$ measure the distance of a surface, z, from a satisfactory solution at a point (x, y). For now, assume that F is dependent not only on z, but also on the first partial derivatives z_x and z_y. Given that we seek a surface defined over some region Ω in the plane, and that F is everywhere non-negative, we may regard

$$I_1(z) = \iint_\Omega F(x, y, z, z_x, z_y)\, dx\, dy$$

as an overall measure of error whose value is to be minimized. This is not a conventional minimization problem since we search over a space of functions and not a region of coordinate space. The value of I_1 depends on the choice of the function z, and for this reason I_1 is termed a *functional*. Minimizing I_1 is a problem in the calculus of variations.

A fundamental result of the calculus of variations is that the extrema of functionals must satisfy an associated *Euler equation* over the domain of interest. For the above form of the functional, the equation is

$$F_z - \frac{\partial}{\partial x} F_{z_x} - \frac{\partial}{\partial y} F_{z_y} = 0.$$

This is a *necessary* condition for the existence of an extremum, z (p. 185, Courant & Hilbert, 1953). It is not a *sufficient* condition. Note that local minima, global minima, local maxima, global maxima, and inflexion points are all examples of extrema.

It will prove useful to note two other Euler equations corresponding to other forms for F. In the event that F is dependent also on the second partial derivatives as in

$$I_2(z) = \iint_\Omega F(x, y, z, z_x, z_y, z_{xx}, z_{xy}, z_{yy})\, dx\, dy,$$

the Euler equation expands to

$$F_z - \frac{\partial}{\partial x} F_{z_x} - \frac{\partial}{\partial y} F_{z_y} + \frac{\partial^2}{\partial x^2} F_{z_{xx}} + \frac{\partial^2}{\partial x \partial y} F_{z_{xy}} + \frac{\partial^2}{\partial y^2} F_{z_{yy}} = 0.$$

Sometimes we seek a surface that is parameterized not in terms of relative depth, but in terms of surface normals. Two parameters are needed in this case. If the functions p

and q are used to describe surface orientation and if the associated functional incorporates their first partial derivatives in x and y, the expression to be minimized then takes the form

$$I_3(p,q) = \iint_\Omega F(x,y,p,q,p_x,p_y,q_x,q_y)\,dx\,dy,$$

which has two corresponding Euler equations given by

$$F_p - \frac{\partial}{\partial x}F_{p_x} - \frac{\partial}{\partial y}F_{p_y} = 0,$$

$$F_q - \frac{\partial}{\partial x}F_{q_x} - \frac{\partial}{\partial y}F_{q_y} = 0.$$

In general, these constitute a pair of coupled partial differential equations in p and q. A pair of functions satisfying these equations will generate an extremum of I_3. As before, the extremum may be a minimum, maximum or inflexion point.

Note that if the functional involves higher derivatives of p and q, the Euler equations generalize in a straightforward way. Courant and Hilbert (1953) provide an excellent chapter on the variational calculus.

B. Boundary Conditions

In general, a problem involving partial differential equations is ill posed in the absence of suitable boundary conditions because the solution is not unique without additional constraint. The type of boundary condition that ensures a given problem is well posed depends on the particular type of partial differential equation. There may be more than one way of adding boundary conditions to a partial differential equation in order to force a unique solution.

In our case, boundary conditions may be given as part of the basic minimization problem. That is, the solution sought must minimize the functional subject to additional constraints, such as prescribed values on the boundary of the region of integration. In the case that the function is not constrained on the boundary, however, the calculus of variations provides so called *natural boundary conditions* that the solution must satisfy. For example, for the functional I_1 given above, a suitable boundary condition is the value of z along the boundary $\partial\Omega$ of the region Ω. The natural boundary condition in this case is just

$$\left(F_{z_x}, F_{z_y}\right) \cdot \mathbf{n} = 0,$$

where the normal to the boundary, $\mathbf{n}$, is given by

$$\mathbf{n} = \left(-\frac{dy}{ds}, \frac{dx}{ds}\right),$$

and s is arc-length measured along the boundary $\partial\Omega$. So the component of the vector (F_{z_x}, F_{z_y}) normal to the boundary should be everywhere zero.

In the case of the functional I_3 given above, the values of p and q along the boundary will usually be suitable. The natural boundary conditions in this case happen to be

$$\left(F_{p_x}, F_{p_y}\right) \cdot \mathbf{n} = 0 \qquad \text{and} \qquad \left(F_{q_x}, F_{q_y}\right) \cdot \mathbf{n} = 0.$$

C. Regularizing Terms

At times, the problem of minimizing a functional is not well posed as there is an infinite number of solutions, even with constraints on the boundary. One can then find a surface that is close to a solution, while minimizing some measure of departure from smoothness, by *regularization* (see Poggio & Torre, 1984). The regularization method of interest to us here involves the addition of a regularizing term to the functional. If we deal with the problem of recovering a surface $z(x, y)$ from shading, we may wish to include a regularizing term such as the square Laplacian appearing in

$$\iint_\Omega \left(\nabla^2 z\right)^2 \, dx \, dy,$$

or the quadratic variation in

$$\iint_\Omega (z_{xx}^2 + 2z_{xy}^2 + z_{yy}^2) \, dx \, dy.$$

Each of these has the desirable property of rotational invariance (Brady and Horn, 1983). The lower order rotationally-symmetric regularizing term

$$\iint_\Omega (z_x^2 + z_y^2) \, dx \, dy,$$

leads to excessive flattening of solutions. The latter form may well be appropriate when applied to surface orientation parameters, such as p and q, or f and g, but this is not the case with a depth parameter, such as z.

D. Enforcing Constraints

Sometimes we seek a minimum of a functional subject to some independent constraint. Suppose, for example, that we are required to minimize the previously defined $I_1(z)$, subject to the constraint that

$$g(x, y, z, z_x, z_y) = 0.$$

In this case we may use the *Lagrangian multiplier* method in which we minimize not I_1 but the augmented functional

$$I_4(z) = \iint_\Omega F(x, y, z, z_x, z_y) + \mu(x, y) \, g(x, y, z, z_x, z_y) \, dx \, dy.$$

We now seek solutions to the associated Euler equation. Note that the Lagrangian multiplier μ is a function of x and y and must be treated as such when deriving the Euler equation.

If we differentiate the functional with respect to $\mu(x, y)$, for a particular x and y, and set the result equal to zero, we get back the original constraint equation. This equation is required to help solve for μ, something we typically have to do in order to eliminate it from the Euler equation. At times, this may take some skill. More importantly, however, the equations that result often do not suggest convergent iterative schemes. In any case, a solution of the resulting Euler equation will have the property that $g(x, y, z, z_x, z_y) = 0$, with I_4 having an extremal value on the manifold $g(x, y, z, z_x, z_y) = 0$.

E. Penalty terms based on constraints

In view of the difficulties experienced when attempting to impose constraints exactly, we often consider an alternative method. In this approach, a penalty term derived from the contraint is employed. Thus we might rely on the Euler equation corresponding to the functional

$$I_5(z) = \iint_\Omega F(x, y, z, z_x, z_y) + \lambda \left[g(x, y, z, z_x, z_y)\right]^2 dx\, dy.$$

Here, λ is a scalar that aligns the arbitrary scales of F and g. Alternatively, it may be regarded as a weighting of the relative importance of the components of the functional. It is, of course, not necessary to square g if it is already guaranteed to be non-negative over Ω for all functions z.

Solutions to the Euler equation for I_5 now specify surfaces that generate an extremal value of I_5. However, these surfaces will not, in general, satisfy the constraint $g(x, y, z, z_x, z_y) = 0$ exactly. Rather, it will be the case that the value of g is small, along with the values of the other expressions being minimized. This is usually an acceptable compromise. More often than not, this approach proves more tractable than the Lagrangian method as there is no multiplier to be eliminated.