International Journal of Computer Vision, 6:2, 75–104 (1991)

Uniqueness in Shape from Shading

J. OLIENSIS
Computer and Information Science, University of Massachusetts at Amherst, Amherst, MA 01003

Received November 27, 1989. Revised December 17, 1990.

Abstract
For general images of smooth objects wholly contained in the field of view, and for illumination symmetric around the viewing direction, it is proven that shape is uniquely determined by shading. Thus, shape from shading is a well-posed problem under these illumination conditions; and regularization is unnecessary for surface reconstruction and should be avoided. Generic properties of surfaces and images are established. Questions of existence are also discussed. Under the conditions above, it is argued that most images are effectively impossible, with no corresponding physically reasonable surface, and that any image can be rendered effectively impossible by a small perturbation of its intensities. This is explicitly illustrated for a synthetic image. The proofs are based on ideas of dynamical systems theory and global analysis.

1 Introduction

Shape from shading has traditionally been considered an ill-posed problem, although it is known to be well-posed in special cases (Horn & Brooks 1989). For a typical shaded image, it has been assumed that there is an infinite number of possible corresponding surfaces. On the other hand, Bruss (1982) proved that for images with exactly one singular point—that is, a single maximally bright point—and with known illumination from the camera direction, there is essentially a unique corresponding surface. Thus, shaded images of this type contain enough information to completely determine the imaged object.

The question of whether shape from shading is ill- or well-posed is important, because the traditional approach to reconstructing shape employs *regularization* techniques, implicitly assuming the problem to be ill-posed. If the problem is actually well-posed, then regularization is unnecessary, and should be avoided since it can lead to a distortion of the recovered surface (Horn 1990b). It is also important to understand what the constraints on the solutions to shape from shading are, especially if they are significant enough to render the problem well-posed. Through incorporating all available constraints in a shape reconstruction algorithm, it may be possible to improve the robustness of shape recovery.

This article presents the first uniqueness proof for shape from shading that is valid for *generic* images. As in previous work, the illumination is assumed to be from the camera direction—see Saxberg (1989a, b) and Oliensis (1990) for the case of more general illumination. More generally, our result, as well as previous ones, applies to reflectance functions symmetric around the optical axis. We also make the standard assumptions about the imaged object—that it is smooth, matte, uniform in reflectance, non self-occluding, and wholly contained in the field of view. Under these conditions, *a shaded image uniquely determines the imaged surface.* A fortiori, shape from shading is a well-posed problem. In a companion paper, our results are partially extended to the case of illumination from a general direction (Oliensis 1990). In general, therefore, shape from shading should not be assumed ill-posed, and regularization should be used with caution.

The existence of solutions is also discussed. For illumination conditions as described above, it is argued that for almost all images, that is, for almost all intensity functions $I(x, y)$, effectively no solution to shape

from shading exists.[1] Moreover, we argue that a true image can be converted into an effectively impossible one, with no object solution, or else just nongeneric and physically unacceptable solutions, by a small perturbation of its intensities. This is illustrated for an explicit image example in section 9. Thus, true images of objects are very special intensity functions. The only previous nonexistence result was limited to a restricted set of images (Horn et al. 1990).

The potential impossibility of images provides an important failure criterion for shape from shading. If the method is applied to an inappropriate image, this may be signaled by the nonexistence of any reasonable solution, showing that the assumptions made about the scene were incorrect. Another consequence is that discretized and noisy images are likely not to correspond exactly to any acceptable 3D surface. Thus, some of the difficulties encountered by Horn (1975) using the classical characteristic strip method of solution are not due merely to the discretization error of numerical integration, but inherent in the inaccuracies of the image-formation process.

Our approach is based on the properties of characteristic strips and singular points, and uses ideas of dynamical systems theory and global analysis. The techniques of dynamical system theory, well developed by mathematicians, have been applied recently to a variety of vision problems. Saxberg (1989a, b) noted that the problem of shape from shading can be usefully reinterpreted along these lines, and proposed a new method of solution which appears to have some promise. In this article, a new viewpoint on the problem is developed based on these techniques; it is simple and intuitive, and offers qualitatively new insights.

The organization of the remainder of the article is as follows. We first describe the method of characteristic strips and demonstrate that the resulting equations represent a Hamiltonian dynamical system. Next, the characteristic strips are interpreted as curves on the imaged surface, and shown to be curves of steepest ascent in depth. In section 4, singular points are introduced; their constraints on the surface and characteristic strips in their neighborhood are explored. The results of these sections can be generalized to the case of general illumination direction (Oliensis 1990). In section 5, the statement of our uniqueness theorem is given, and an intuitive overview of its proof is presented. The detailed proof is given in sections 6–8. Finally, the existence of solutions and impossible images are discussed in section 9.

2 The Characteristic Strip Method as a Hamiltonian Dynamical System

A *characteristic strip* is, roughly, a line in the image along which the surface depth and orientation can be computed, assuming that these quantities are known at the starting point of the line. Characteristic strips were used by Horn in his original algorithm for reconstructing shape from shading (Horn 1975). A consistent solution to shape from shading determines a *flow* of characteristic strips in the image, with every image point lying on exactly one characteristic strip line. Conversely, such a flow of characteristic strips uniquely determines a shape solution (see figure 1). Most of our proof is focused on uniquely specifying the flow of characteristic strips, and thus, from the above, a unique shape solution.

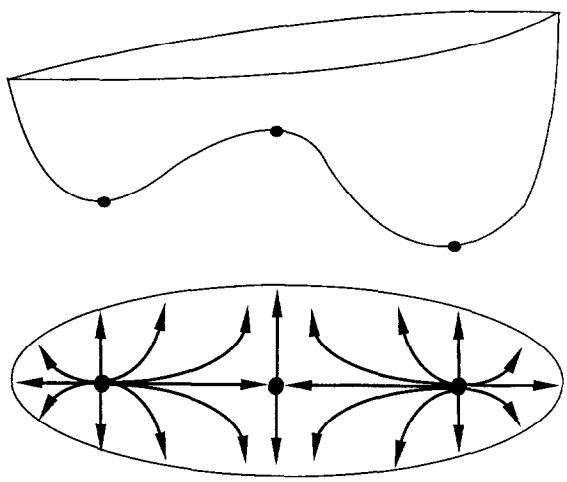


Fig. 1. The flow of characteristic strips in the image determines the surface solution, and is determined by it.

To begin with, arbitrary reflectance functions $R(p, q)$ will be considered. As usual, it is assumed that the imaged surface is uniform, so the reflectance is the same at every point, and that the image is derived via orthographic projection. The image irradiance equation can be written as

$$H \equiv I(x, y) - R(p, q) = 0 \tag{1}$$

p, q represent as usual the derivatives of the surface depth z with respect to x and y, respectively. The characteristic strip equations are:

$$\dot{x} = H_p \left(\equiv \frac{\partial H}{\partial p} \right), \quad \dot{y} = H_q, \quad \dot{p} = -H_x, \quad \dot{q} = -H_y \tag{2}$$

The dot denotes a derivative with respect to 'time,' an arbitrarily chosen variable that parameterizes the position along the characteristic strip. The subscripts denote partial differentiation. As pointed out by Saxberg, these equations constitute a *dynamical system*, that is, they can be thought of as determining the 'motion' of a particle whose 'position' is specified by the four parameters $(x(t), y(t), p(t), q(t))$. This four-dimensional 'position' space will be referred to below as *phase space*. For dynamical systems, the quantities on the right-hand side of the dynamical equations are referred to as the components of a *vector field*. However, more can be said. These equations determine a very special type of dynamical system, namely a *Hamiltonian* one.

Hamiltonian dynamical systems are familiar and well studied. They have the defining properties that (a) there exists an *energy* function which is a constant of the motion, (b) the motion parameters can be divided into equal numbers of 'coordinate' and 'momentum' parameters, and (c) the evolution of the system is governed by Hamilton's equations. Many problems of classical Newtonian mechanics, celestial mechanics, etc., fall into this category. Essentially, it comprises all systems without frictional, or dissipative, forces.

As an example that will prove to be relevant later on, consider an electrically charged particle moving on a plane, in response to an electric field. This system has a conserved energy which is the sum of a *kinetic energy* of motion, and a *potential energy*, whose gradient is equal to the negative of the electric field. The momentum vector $\vec{p}$ is equal to the mass of the particle times its velocity. The energy is explicitly:

$$H = \frac{1}{2m}\vec{p}^2 + V(\vec{x})$$

with $\vec{E}(\vec{x}) = -\nabla V(\vec{x})$, and mass m.

In general, Hamilton's equations are

$$\dot{x}_i = \frac{\partial H}{\partial p_i} \quad \text{and} \quad \dot{p}_i = -\frac{\partial H}{\partial x_i} \tag{3}$$

for each coordinate-momentum pair (x_i, p_i). H is the energy. For the example, these equations become:

$$\dot{x}_i = \frac{p_i}{m} \quad \text{and} \quad \dot{p}_i = m\frac{d^2x_i}{dt^2} = E_i \tag{4}$$

The second equation is just the familiar Newton's law, $F = ma$, while the first defines the momentum.

On comparing equation (3) with the characteristic strip equations, one finds that the latter are equivalent to a Hamiltonian system, with conserved energy $H = 0$, and generalized momenta p (corresponding to x) and q (corresponding to y). With this insight, one can think of the characteristic strip equations as describing the motion of a particle on the image plane, with momentum $\vec{P} = (p, q)$. This is now a two-dimensional problem, much easier to visualize than the apparently four-dimensional dynamical system of equation (2). Also, more is known about the special case of Hamiltonian systems than about general systems. For example, they satisfy Liouville's theorem that the flow in phase space preserves volume, with the consequence that all fixed points of the flow must be saddle points. This is useful in what follows, since the fixed (i.e., singular) points are important in characterizing the shape-from-shading solution. (To avoid later confusion, it should be remembered that the fixed points are saddle points in the full four-dimensional phase space parameterized by (x, y, p, q), but *not* necessarily saddle points of the surface $z(x, y)$.)

From now on, it is assumed that the illumination is symmetric around the viewing direction, and that the surface is matte, with known reflectance, and no self-occlusion. For simplicity, we focus on the case of a Lambertian surface with unit albedo. Then the equation for the image intensity I is

$$I = \hat{n} \cdot \hat{L} = \frac{(-p, -q, 1) \cdot (0, 0, 1)}{(1 + p^2 + q^2)^{1/2}} = \frac{1}{(1 + p^2 + q^2)^{1/2}} \tag{5}$$

where $\hat{n}$ is the unit surface normal, and $\hat{L} = \hat{z}$ is a unit vector giving the light-source direction. Note that the intensity I reaches its maximum value $I = 1$ at $p = q = 0$. After some algebra, this can be rewritten as

$$H \equiv \frac{1}{2}(p^2 + q^2) + V(\vec{x}) = 0 \tag{6}$$

with

$$2V(\vec{x}) \equiv 1 - \frac{1}{I^2} \tag{7}$$

This definition of the Hamiltonian (i.e., energy) function H differs slightly from the one used above. It is chosen because of its simplicity: the energy is expressed as a sum of a *quadratic* kinetic energy term which contains all the 'momentum' dependence, and of a potential energy term containing all the 'coordinate' dependence. The image irradiance equation can also be written in this form for more general reflectance functions

symmetric around the optical axis. This is the familiar form of the energy for elementary Newtonian mechanics. In fact, this H is exactly the same as the example above, with the mass taken to be unity. *The characteristic strips for the given reflectance function are equivalent to the motion of a particle on a plane in response to a potential.*

Explicitly, the characteristic equations are

$$\dot{x} = H_p = p, \quad \dot{y} = H_q = q$$
$$\dot{p} = -H_x = -V_x, \quad \dot{q} = -H_y = -V_y \qquad (8)$$

The first two state that the surface slopes $(p, q) \equiv \vec{P}$ are just the *velocity* of the moving point whose trajectory gives the characteristic strip.

3 Characteristic Strips as Surface Curves

The equation for the surface depth z has been neglected so far, since it is unnecessary in deriving the trajectory. It is

$$\dot{z} = p\dot{x} + q\dot{y} = \dot{x}^2 + \dot{y}^2 \geq 0 \qquad (9)$$

The direction of time has been defined so that z always increases with time along the trajectory. Actually, since

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix} = \vec{\nabla} z \qquad (10)$$

the tangent to a characteristic trajectory is parallel to the surface gradient. Thus, a characteristic trajectory $(x(t), y(t))$ is the image plane projection of a *curve of steepest ascent on the 3D object*. This result generalizes to the more general reflectance functions considered below, and to the case of general illumination direction (Oliensis 1990). The two-dimensional dynamical system specified by the above equation is an example of a *gradient* system—a dynamical system that is even simpler than a Hamiltonian one. For instance, gradient systems can not have limit cycles (which are closed periodically traversed trajectories).

A large part of the succeeding argument focuses directly on the general properties of surfaces. For each surface $z(x, y)$, the curves of steepest ascent on this surface constitute a flow: every point on the surface lies on exactly one of these curves. Also, this flow is a solution of the two-dimensional gradient dynamical system described by equation (10). Our strategy is to analyze the properties of surfaces in terms of the properties of this flow, using general theorems about gradient dynamical systems. As stated above, these curves of steepest ascent correspond exactly to the characteristic strips in the image plane. Thus, our results on the flow of surface curves can be applied directly to the flow of characteristic strips, and used to establish the properties of the solutions to shape from shading (see figure 2).

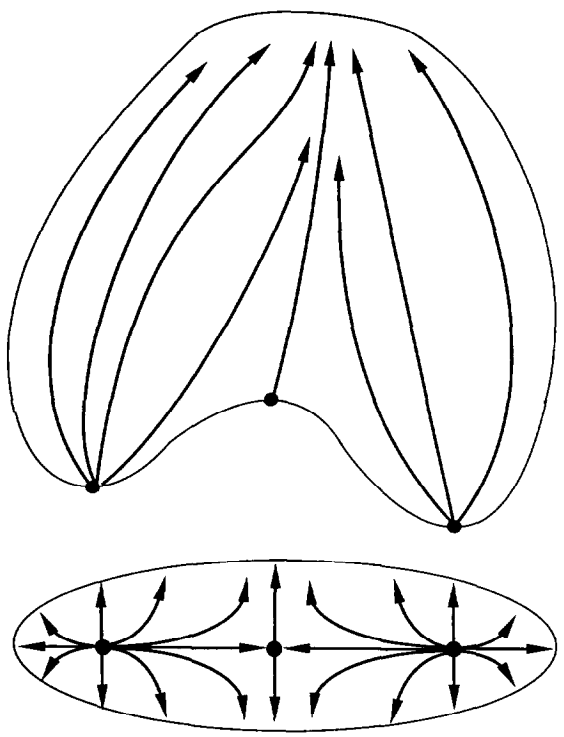


Fig. 2. Characteristic strips in the image correspond to curves of steepest ascent on the imaged surface.

The interplay between surface curves and characteristic strips—described by two different dynamical systems—is crucial to our argument. The essential difference between the two systems is that in discussing equation (10) we assume that z is a *known* function of x and y, whereas in the characteristic strip equations z is determined from the equations. This difference is reflected in the fact that equation (10) represents a two-dimensional system, while equation (2) is four-dimensional.

In the remainder of this article, the characteristic strip curves in the full four-dimensional phase space (x, y, p, q) will be referred to as *trajectories*, as a reminder that they can be considered the trajectories of particles moving on the image plane in response to forces. The projection of the characteristic trajectories onto the

image plane will be referred to as *image strips*, or as *gradient curves*, as a reminder that they can be associated with the two dimensional gradient dynamical system equation (10). The characteristic trajectories projected on an object surface will be referred to as *surface strips*.

4 Singular Points and the Winding Number

The importance of singular points in determining and fixing a solution to the shape-from-shading problem has been stressed by many people (Bruss 1982; Saxberg 1989a; Horn 1990b), and they are crucial in this work as well. The *singular points* of an image are conventionally defined as those at which the value of the intensity I is maximal for the given reflectance function. Their importance stems from the fact that the surface orientation is uniquely determined at these points. For the reflectance function of equation (5), singular points occur for $I = R = 1$. At a point where I attains the maximal value 1, $p = q = 0$, and the tangent plane to the surface is parallel to the image plane. Thus, for the reflectance functions we consider, singular points correspond to *critical points* of z. Also, the derivatives of the reflectance function with respect to p and q vanish at singular points. From the characteristic strip equations, equation (2), this implies that $\dot{x} = \dot{y} = 0$—a characteristic trajectory initially at a singular point never leaves it. Singular points are thus *fixed points* in the language of dynamical systems theory.

The concept of a nondegenerate singular point will also be important. For the present case, a singular point is *nondegenerate* if the matrix of second derivative of the intensity I is nonsingular at the point. At such a point, it is easy to show that the two eigenvalues of the second derivative matrix, if they are unequal, determine the principal curvature values of the surface up to sign (Oliensis 1989). We always assume below that the eigenvalues at a singular point are in fact unequal, as is generically true. Then the nonsingularity of the matrix implies that the principal curvatures are both nonzero, which is clearly the generic case for a surface, at least for an isolated singular point, as was noted by Saxberg (1989b).

As a result, the imaged surface at a nondegenerate singular point is either convex, concave, or saddle-shaped. Correspondingly, near the singular point, there are three possible types of flows of gradient curves. Consider the flow of curves of steepest ascent near the surface point corresponding to the singular point. It is clear that if the surface is locally convex, the direction of these curves is *outward* from the given point. The same is true of the corresponding characteristic strips projected as 2D curves in the image plane. The singular point in this case is referred to as a *source* (see figure 3). Similarly, if the surface is locally concave, the corresponding image strips converge toward the singular point, which is referred to as a *sink* (see figure 4). In the third case, the flow of gradient curves is more complicated: there are two gradient curves that converge to the fixed point, and two that originate at the fixed point; all other gradient curves, like comets, initially approach the saddle point, but miss it and recede into the distance (see figure 5). In this case the singular point is referred to as a saddle point. For obvious reasons, sources and sinks will be referred to collectively as *elliptical points*.

A more formal characterization of the image strip flow near a nondegenerate singular point follows from the Grobman-Hartman theorem (Saxberg 1989, a, b). Consider equation (10) in the neighborhood of a singular point. The matrix of partial derivatives of the right-hand side of this equation with respect to x, y is just the matrix of second derivatives of z:

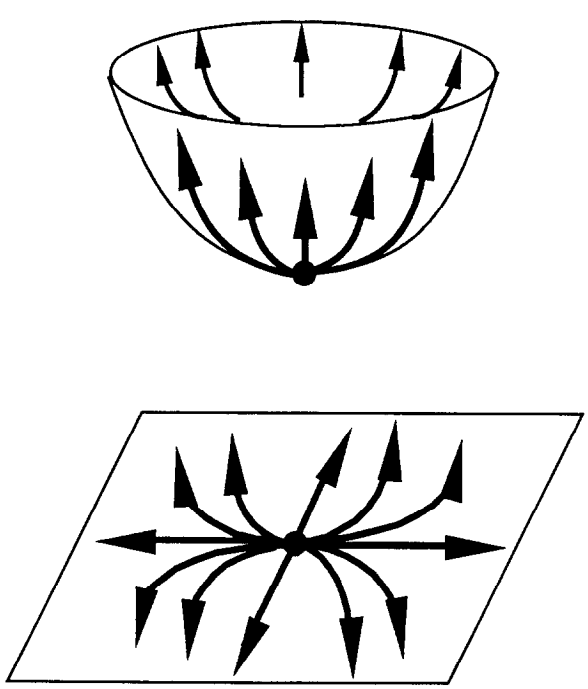
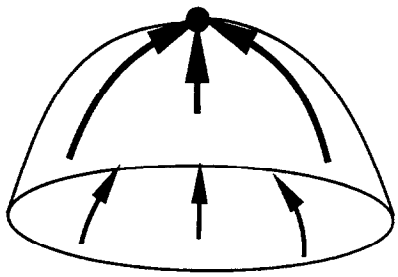


Fig. 3. A source. This image-plane flow, with all characteristic strips originating at the singular point, corresponds to a locally convex surface.



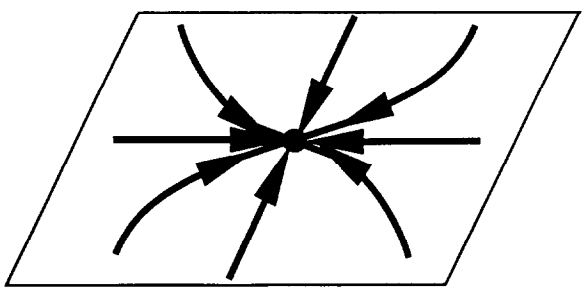


Fig. 4. A sink. This image-plane flow corresponds to a locally concave surface.

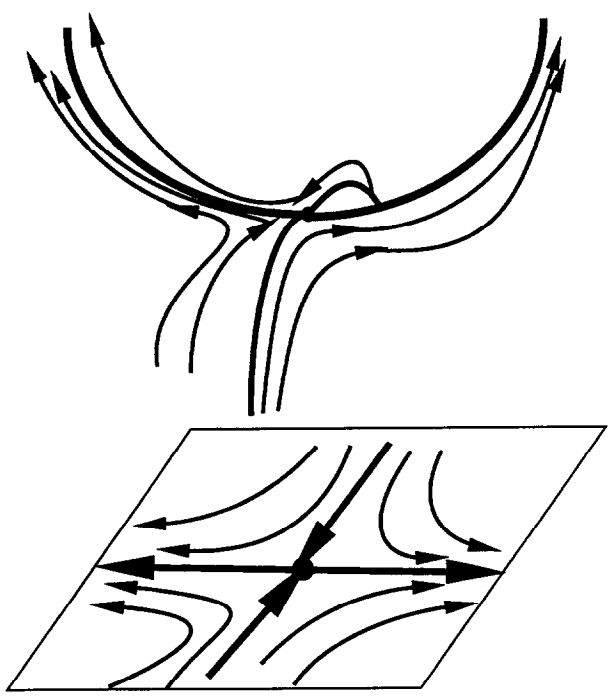


Fig. 5. A saddle. This image-plane flow corresponds to a locally saddle-shaped surface. Only four characteristics strips actually connect to the singular point.

$$\mathbf{A} \equiv \frac{\partial^2 z}{\partial x_i \partial x_i} \tag{11}$$

where $(x_1,x_2) \equiv (x, y)$. By the nondegenerateness assumption, and again assuming unequal eigenvalues of the intensity second-derivative matrix, **A** can be computed from the image up to a sign ambiguity, and is nonsingular. Then the Grobman-Hartman theorem (Palis & de Melo 1982) states that there exists a homeomorphism defined in a neighborhood of the fixed point, mapping orbits of the flow to those of the linear flow exp ($t\mathbf{A}$), preserving the sense of the orbits, and also the parameterization by time.

The meaning of the linear flow is that

$$\vec{x} \rightarrow \vec{x}(t) = e^{t\mathbf{A}}\vec{x} \tag{12}$$

In a coordinate frame in which **A** is diagonal,

$$\vec{x}(t) = (\exp(tA_{11})x, \exp(tA_{22})y) \tag{13}$$

Depending on the signs of A_{11}, A_{22} (which cannot be determined from the image), the flow will be convergent, divergent, or saddle-type, as described above. The existence of a homeomorphism between the complete flow governed by equation (10) and the linear flow means that these flows are topologically equivalent locally. Thus, the complete flow also must have one of these three forms.

The nondegenerate singular points in the image and the corresponding surface points each have an associated integer-valued *index*. To define the index, consider a small circle containing some singular point and no others. For a given flow, each point on the circle is intersected by a unique gradient curve. Define a function giving for each point on the circle the direction of the flow curve passing through that point, as a unit vector. As one goes around the circle (say in a clockwise direction), eventually returning to the starting point, this unit vector rotates, and must also return to its original direction. The number of complete rotations it makes in the process, in the clockwise direction, gives the value of the index (see figure 6). An index corresponding to a rotation in the opposite sense from the path taken around the circle is defined to be negative.

The most important fact about the index is that it is a topological invariant. It does not depend on the particular path chosen around the fixed point, which need not be a circle in general. Its value defined with respect to a path will not change as that path is distorted from

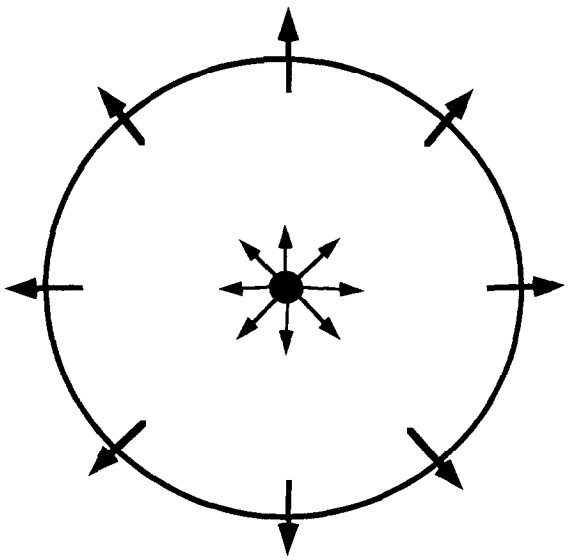


Fig. 6. The index of a singular point, here a source. The unit vector representing the characteristic strip flow direction describes a single clockwise rotation as a clockwise circuit is made around the circle containing the source. Thus, the index is +1.

its original shape, as long as the interior of the path contains only the single fixed point. In general, the index, or rather *winding number*, can be defined for an arbitrary closed curve, not necessarily containing exactly one fixed point in its interior, in the obvious way. Clearly, the winding number is always an integer. Also, the winding number of a curve is the additive sum of the indexes of all singular points contained within its interior. These general theorems can be proven fairly simply for the case at hand.

A crucial fact is that the index for sources and sinks is +1 (e.g., figure 6), while for saddles it is −1. See, for example, Arnold (1973) and Abraham & Shaw (1983) for good intuitive expositions of these results.

5 The Uniqueness Theorem

In this section the uniqueness theorem is presented and its proof is sketched.

First, we give some definitions. For a smooth, closed, object, the *occluding boundary* is defined to be the set of all object points at which the surface normal is perpendicular to the optical axis; the *limb* is the image of the occluding boundary (see figure 7). A smooth, closed object is *non-self-occluding* if all points on the limb are also on the boundary of the image region containing the projected object. This is a slight extension of the standard, common-sense meaning: in addition

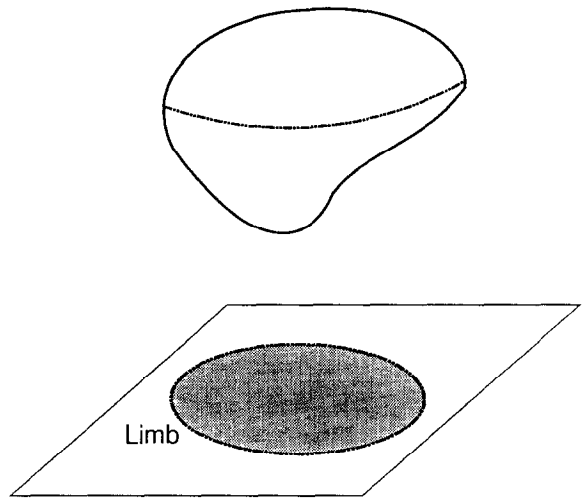


Fig. 7. The limb is defined as the image of the occluding boundary.

to the usual cases of self-occlusion, it excludes those where the object virtually self-occludes—where the surface normal in the interior of the image becomes perpendicular to the optical axis with no actual self-occlusion (see figure 8). A *genus zero* surface is a closed surface with no holes or handles, essentially a deformed sphere.

The class of reflectance functions we consider is the same as those considered by Bruss (1982). A *Bruss reflectance function* $R(p, q)$ is defined as one satisfying: (1) R is a smooth, non-negative, function depending only on $p^2 + q^2$ (thus, it is symmetrical about the optical axis $\hat{z}$). (2) R attains a unique global maximum at $p = q = 0$, with the surface normal to the optical axis. (3) The derivative of R with respect to its argument $p^2 + q^2$ is less than zero at this point. (4) $R \rightarrow 0$ as its argument goes to infinity, so that R vanishes on the occluding boundary. (5) R is monotonic (thus, the image irradiance equation can be inverted, and transformed into eikonal form). A typical Bruss reflectance function corresponds to the illumination of a matte or Lambertian surface from the camera direction.

A *generic* class is one containing essentially all instances, apart from a few special cases. If an instance is contained in this class, then all instances within some neighborhood are also. Moreover, any special case not in this class is unstable—an infinitesimal perturbation will yield an instance contained in the class. Similarly, we define a generic property to be true of essentially all instances, and false only in special and unstable cases. We show in the next section that conditions 4 and 5 of the theorems stated below hold for essentially all images—they are generic properties of images.

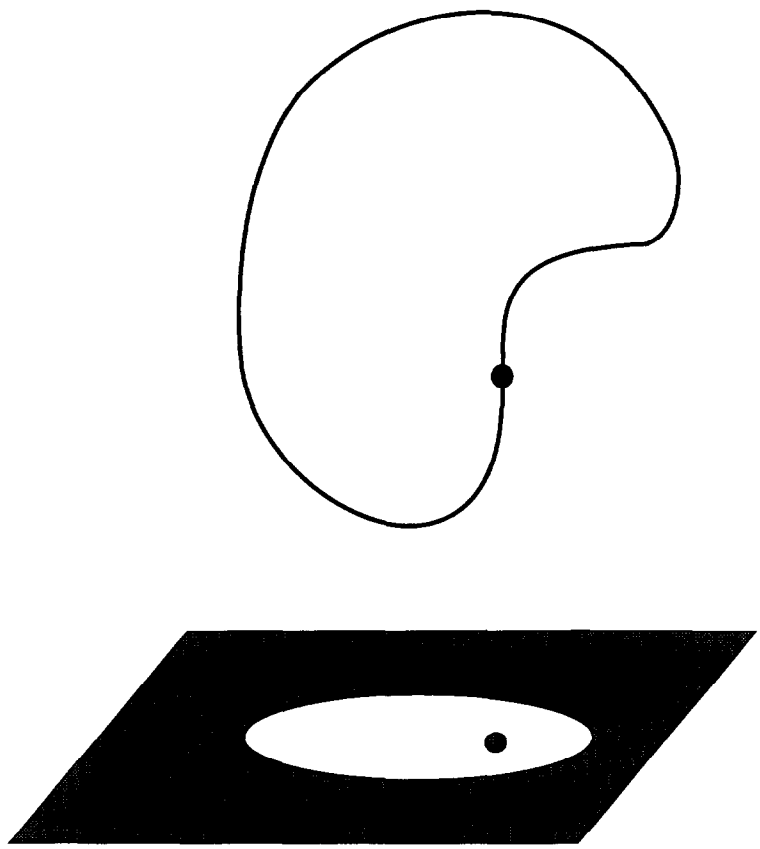


Fig. 8. An example of a self-occluding surface. At the indicated point, the surface normal is parallel to the image plane, giving rise to an intensity zero in the interior of the imaged region.

Finally, we note that for images of closed, non-self-occluding objects, the limb is generically a smooth, closed curve (Giblin & Weiss 1987). The uniqueness theorem can now be stated:

THEOREM: *Assume (1) an image of a closed, smooth, non-self-occluding, genus-zero object is produced by orthographic projection; (2) the reflectance function is a Bruss reflectance function; (3) the object is completely contained in the field of view, and the limb is a smooth, closed curve (this is generically true from the above); (4) the image contains a finite number of singular points all of which are nondegenerate; (5) at each singular point, the matrix of second derivatives of the intensity does not have two equal eigenvalues. Then the visible surface of the original object is the unique solution to shape from shading corresponding to a closed object.*

Alternatively, the conclusion of this theorem can be restated as follows. The pose of a non-self-occluding surface is defined to be *accidental* if an infinitesimal rotation will cause the surface to self-occlude.

THEOREM: *Assume conditions 1–5 above. Then the visible surface of the original object is the unique solution to shape from shading whose pose is nonaccidental.*

This uniqueness result may seem to conflict with the well-known, two-fold convex-concave ambiguity in reconstructing shape from shading. However, for the concave solution, the occluding boundary is seen edge on in the image, which clearly constitutes an accidental alignment of the viewing direction with the rim of the surface. Moreover, this solution is impossible for a closed object—there is no way to extend the surface to a closed object without occluding it. The second solution is therefore excluded as an acceptable one (see figure 9). These theorems may be summarized as stating that for a Bruss reflectance function, and an object wholly contained in the field of view, there is essentially always a unique solution to shape from shading.

The proof of these theorems is sketched below. The complete proof is given in the following sections. The first step is based on the local uniqueness theorem of Bruss (1982). Essentially, Bruss proved that in the

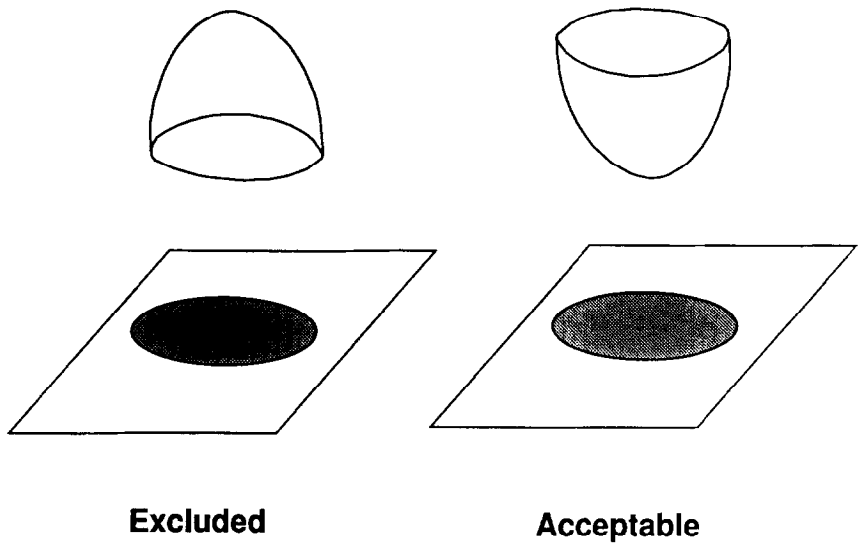


Fig. 9. Surface solutions that are concave at the occluding boundary are excluded.

neighborhood of a nondegenerate singular point, there is a unique locally convex surface solution, and, similarly, a unique locally concave solution. The singular point is a source in the first case, and a sink in the second.

The second step of the proof is the demonstration that the convex and concave solutions around source and sink singular points can be extended beyond the local neighborhoods of these points. This follows from the existence and uniqueness theorems of differential equations: a characteristic strip, as the solution of a differential equation, can always be extended until it exits the image, or else converges to some limit point—a singular point. These strips carry the surface solution with them, extending it over a large region. In fact, the solution regions associated with sources and sinks essentially cover the image. This does not quite determine the solution, however, since the relative depths of these singular points may not be known. One still needs to determine how the splicing of the different solution regions should be done.

The third step proves that all singular points in a consistent solution actually are connected together by sequences of characteristic strips. Since the surface can be computed along these strips, this determines the relative depths of the singular points and the splicing of the different solution regions. This is sufficient to show that if the nature of the surface solution is known at each singular point—that is, whether it is concave, convex, or saddle-shaped—then the shape solution is uniquely determined. The last step in the proof shows that the type of the solution at each singular point is in fact uniquely specified, and that the solution is therefore unique.

Various arguments are used to demonstrate the last step. Suppose some singular point is assumed to be a source, corresponding to a locally convex solution. This locally convex solution is uniquely determined, and therefore the characteristic strips emanating from the source are also. Then, one can show that the nature of the solution at any other singular point to which the first is connected by a strip is also determined. By a chain reaction of this reasoning, one can determine the nature of the solutions at many singular points which are connected to the original source by sequences of strips. Another important ingredient is the result that at the image boundary, all characteristic strips must be exiting the image region. Thus, all gradient curves in a consistent solution must originate at some singular point, which determines them. Also, any singular point connected to the image boundary by a characteristic strip can be identified as either a source or a saddle, but not a sink. Lastly, an argument based on the winding number can be applied to uniquely determine the number of saddle-type singular points in the image. These and other arguments determine uniquely the nature of the surface solution at each singular point, and the surface itself.

6 Proof of Uniqueness: Preliminary Results

The imaged object is assumed to be genus zero—without holes—and non-self-occluding. Also, it is contained in the field of view. The image therefore consists of a light blob in a black background, that is, the intensity is nonzero in a compact simply connected region and zero elsewhere. The image intensity is also assumed to be smooth, except at the boundary of the light region, where the intensity falls continuously to zero. This boundary corresponds to the occluding boundary of the imaged object. It will be referred to as the *image boundary* or limb.

We consider only images with a finite number of singular points, all of which are nondegenerate. This is not a real restriction: it will be shown later that such an image can always be obtained by an infinitesimal perturbation of the imaged object, and that essentially all images have these properties. In other words, these images constitute a generic class.

Some basic properties of the flow of gradient curves corresponding to a surface solution are now derived. We do this by considering the surface curves of steepest ascent—the gradient curves are just the image-plane projections of these. Let us fix a particular surface solution. Every point on the surface clearly lies on a unique curve of steepest ascent. Moreover, since the surface is finite in depth and compact, every such curve clearly must terminate—either at some critical point of the depth (corresponding to a singular point), or else at the occluding boundary. The singular points are isolated by our (generically valid) assumption. It should therefore be fairly clear that a singular point to which such a curve converges is the unique terminating point for the curve in the given time direction. Similarly, it will be shown that if a surface strip converges to the occluding boundary, then it converges to a unique point on the boundary. Thus, for a consistent surface solution, the gradient curves in the image plane fill out the plane, and, at each end, a gradient curve must terminate either at a unique singular point, or else at a unique point on the image boundary. These conclusions are now proved in more detail.

6.1 *Gradient Curves Terminate at Unique Singular Points*

Suppose that a subsequence from a gradient curve converges to an interior point w as $t \to \infty$. We show that this point must be a singular point, and that it is the unique such point. (see Palis and de Melo (1982)). An equivalent argument works for convergence at $t \to -\infty$. Suppose that ∇z is nonzero at w. The gradient curve through w itself goes from z smaller than that of w to larger z. By continuity, trajectories through all points close enough to w will also reach z greater than that of w. Thus no gradient curve can converge to w, since if it approaches too closely it will attain larger z and never be able to return to w. Thus w must have $\nabla z=0$, and is a singular point.

Next, suppose a gradient curve converges to more than one interior point as $t \to \infty$. Then, it must travel back and forth between them an infinite number of times. Different branches of the curve pass through points infinitesimally close to one of the fixed points, and thus, by continuity, these branches converge to an infinite set of points. But in the present case the image contains only a finite number of singular points, so this is impossible. Thus, as claimed, a gradient curve converging to an interior point as $t \to \infty$ converges to a unique singular point. When a point lies on a gradient curve that converges to some singular point s, it will be called *connected* to s. Also, if there exists a gradient curve converging to two different singular points, one at either end, then we will say that these singular points are connected to each other by this curve.

6.2 *Gradient Curves Point Outward on the Image Boundary*

We next consider gradient curves near the image boundary. Along this boundary, there are essentially two possibilities for the shape of the object: either the surface protrudes toward the camera (the object is concave at the boundary), or else the surface recedes from the camera (it is convex at the boundary). The former case violates general position for nontransparent objects—the viewing direction accidentally coincides with a head-on view of the object rim. In this concave case, an infinitesimal rotation would cause the object to self-occlude. Also this case is impossible for a closed surface, since a smooth continuation of the object beyond the rim must again produce self-occlusion. This is shown in more detail below. In the convex case, in contrast, a small rotation would presumably just reveal a part of the continuation of the visible surface of the object. Therefore, in accordance with the statements of the uniqueness theorem in section 5, we will assume in the following sections that the surface is *convex everywhere along its occluding boundary.* Note,

however, that the proof contained in this article also demonstrates that there is a unique solution with the surface concave everywhere along the boundary.

As will be shown, convexity implies that the surface curves of steepest ascent are heading *outward* from the visible region near the occluding boundary. Thus, the corresponding flow of gradient curves near the image boundary is also outward. (see figure 10). This fact, which follows from our stipulation that the recovered surface is either closed, or else in general position, will be very important in the uniqueness proof.

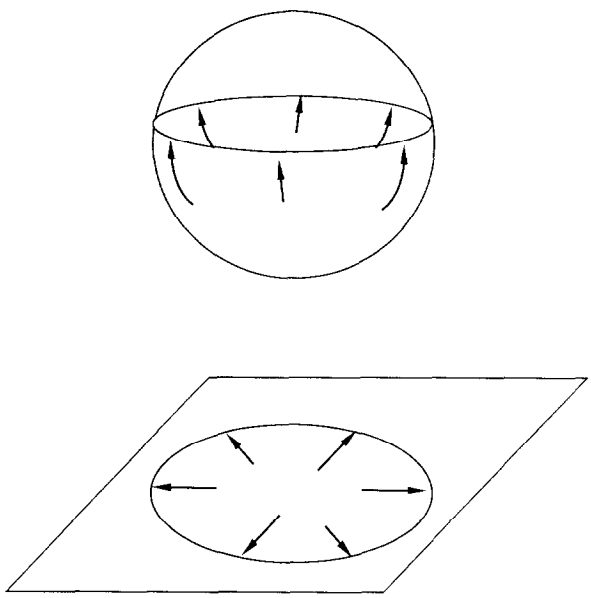


Fig. 10. For a closed surface, the direction of the steepest ascent curves at the occluding boundary is outward from the visible region. Correspondingly, the characteristic strip direction at the limb is outward.

The foregoing statements will now be proved in more detail. It is shown that for each image boundary point there exists a gradient curve that smoothly converges to it, that the direction of this curve is outward, and that it intersects the image boundary perpendicularly. Consider a particular surface solution. We assume that it is smooth everywhere, including at its occluding boundary; this implies that the surface can be smoothly extended beyond the occluding boundary. The surface strips are determined on the 3D object by the standard equations:

$$\frac{dx}{dt} = p, \qquad \frac{dy}{dt} = q, \qquad \frac{dz}{dt} = p^2 + q^2$$

Let $\alpha(x)$ be a real infinitely differentiable function such that $1 \geq \alpha(x) \geq 0$, and $\alpha(x) = 1$ if $|x| \leq 0.2$, $\alpha(x) = 0$ if $|x| \geq 0.4$. Such a function can be shown to exist. Consider the modified equations for the gradient curves:

$$\begin{aligned} \frac{dx}{dt} &= pR\left(\frac{1}{p^2 + q^2}\right) \\ \frac{dy}{dt} &= qR\left(\frac{1}{p^2 + q^2}\right) \\ \frac{dz}{dt} &= (p^2 + q^2)R\left(\frac{1}{p^2 + q^2}\right) \end{aligned} \tag{14}$$

with

$$R(x) \equiv 1 + \alpha(x)(x - 1) \tag{15}$$

These equations determine exactly the same gradient curves as the previous ones, since at every point the tangent direction to the curve is the same as before. Moreover, the new vector field (i.e., the right-hand sides of these equations) is as smooth as the previous one. Lastly, it is perfectly finite and well defined at the occluding boundary. Thus the flow of surface curves described by these equations can be defined on the object surface *in a region containing the occluding boundary* as well as on the visible surface. These curves are just the surface curves of steepest ascent, as illustrated in figure 2. Note that we are explicitly defining the flow on the object rather than in the image plane—the corresponding flow projected in the image plane is *not* smooth at the occluding boundary.

Consider a point on the limb b and a corresponding point B on the occluding boundary of the object. (Every point on the limb corresponds to at least one point on the occluding boundary of the object—just consider a line in the image that converges to b, and its projection on the object surface.) For smooth surfaces with no self-occlusion, the limb is differentiable (Giblin & Weiss 1987). For convenience, we switch to a new set of image-plane coordinates $(\bar{x}, \bar{y})$ with the origin at b, and rotated so that the tangent to the limb at b is in the $\bar{x}$ direction. We also take the image region to be on the left of the boundary line, that is, at negative $\bar{y}$ values (see figure 11). Then the tangent plane to the surface at B is given by the $\bar{x}$ and z directions. Because the surface is smooth, it can be parameterized locally by $\bar{y}(\bar{x}, z)$. Also,

$$p = \frac{\partial z}{\partial \bar{x}}\bigg|_{\bar{y}} \equiv z_{\bar{x}}|_{\bar{y}} = -\frac{\bar{y}_{\bar{x}}|_z}{\bar{y}_z|_{\bar{x}}}, \qquad q = z_{\bar{y}}|_x = \frac{1}{\bar{y}_z|_{\bar{x}}} \tag{16}$$

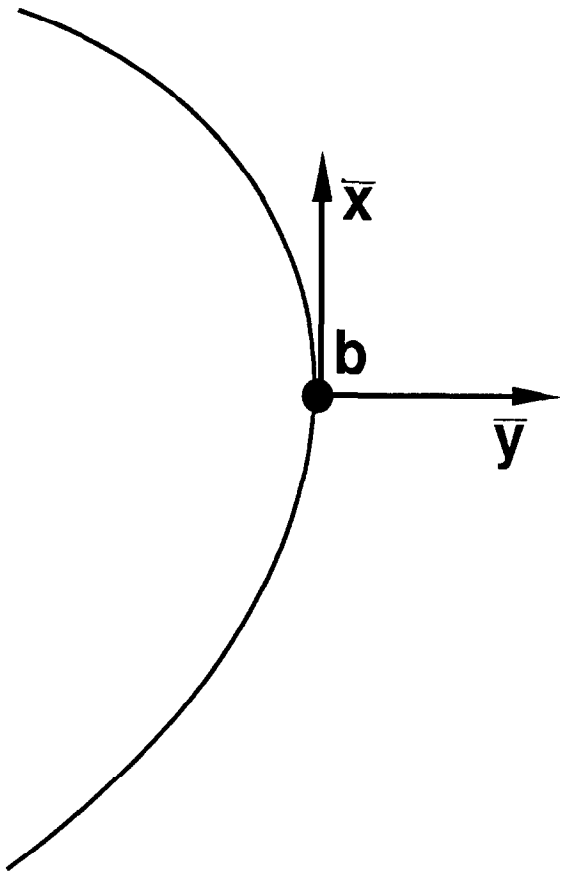


Fig. 11. An image plane coordinate system defined so that the origin is a point on the limb b, and the tangent to the limb is in the $\bar{x}$-direction. The z direction is into the page.

and

$$\frac{p}{q} = \frac{z_{\bar{x}}|_{\bar{y}}}{z_{\bar{y}}|_{\bar{x}}} = -\bar{y}_{\bar{x}}|_z \tag{17}$$

The notation $\bar{y}_z|_{\bar{x}}$ signifies a partial derivative with respect to z keeping $\bar{x}$ fixed. From the above,

$$\frac{1}{p^2 + q^2} = \frac{q^{-2}}{1 + p^2/q^2} = \frac{(\bar{y}_z|_{\bar{x}})^2}{1 + (\bar{y}_{\bar{x}}|_z)^2}$$

The surface normal can be written

$$\hat{n} = \frac{(-\bar{y}_{\bar{x}},\ 1,\ -\bar{y}_z)}{(1 + \bar{y}_x^2 + \bar{y}_z^2)^{1/2}} \text{ sign } (\bar{y}_z) \tag{18}$$

The occluding boundary is characterized by $\hat{n}_z = 0 = \bar{y}_z$. Elsewhere, $\bar{y}_z$ must be nonzero, since the surface is non-self-occluding, and therefore has the same sign over the whole visible surface near B. Our stipulation that the surface be convex at the occluding boundary amounts to choosing $\bar{y}_z > 0$. It is shown below that this is the only possible choice when the surface is closed.

Near the boundary point, the equation for the 3D gradient curve becomes:

$$\frac{d\vec{r}}{dt} = \begin{pmatrix} \dfrac{-\bar{y}_x|_z \bar{y}_z|_{\bar{x}}}{(1 + (\bar{y}_{\bar{x}}|_z)^2} \\ \dfrac{\bar{y}_z|_{\bar{x}}}{(1 + (\bar{y}_{\bar{x}}|_z)^2} \\ 1 \end{pmatrix} \tag{19}$$

where $\vec{r} \equiv (\bar{x}, \bar{y}, z)$. Since $\bar{y}$ has an extremum at B, by our coordinate choice, $\bar{y}_z|_{\bar{x}}$ and $\bar{y}_{\bar{x}}|_z$ go smoothly to zero near B. Consider the gradient curve whose corresponding surface strip passes through B. The above equation indicates that its tangent approaches the $\bar{y}$ direction as the curve converges to b. Thus, as claimed, the direction of the gradient curve is outward as it crosses the limb, and it crosses this boundary perpendicularly. It is also clear that the point at which the gradient curve reaches the limb is the unique limit point of the curve in the positive time direction. It has therefore been demonstrated that every gradient curve converges to a unique point at either end.

Let us now reconsider the sign choice for $\bar{y}_z$. Whatever the choice, equation (19) implies that the direction of all gradient curves is the same—either all outward or all inward—along a portion of the limb near b. Since b was an arbitrary point on the limb, it follows that the gradient curve direction is the same on the entire limb. Now suppose that $\bar{y}_z \leq 0$ near b, so that the gradient curve direction is inward on the limb. Consider the corresponding surface curves of steepest ascent. These can be defined over the invisible portion of the object, and therefore the surface strips that emerge into the visible region from the occluding boundary can be extended backward into the invisible region. These backward extensions must project onto the image region in the xy plane. However, since z decreases in the backward direction, points on these extensions are closer to the camera than points on the later, visible portions of the curves. Therefore, these backward extensions, which were presumed invisible, will in fact occlude the later portions of the steepest ascent curves following their (presumed) entry into the visible region. This contradiction shows that only the sign choice $\bar{y}_z \geq 0$ is valid for a closed surface, and that consequently the gradient curve direction is outward everywhere on the limb.

The fact that the gradient curves point outward at the limb has an extremely important consequence. Since

these curves cannot originate on the limb, it must be true for a consistent solution that *all gradient curves originate at singular points*. This is important because it will be shown later that singular points essentially determine the characteristic trajectories connecting to them. Since we have shown that all image points connect to singular points, this implies that the surface solution may be determined over the whole image.

Another consequence is as follows: consider the winding number of the image boundary curve. Because the gradient curves intersect the boundary perpendicularly, and because the boundary does not self-intersect, it has winding number +1. (For instance, compare figure 10 with figure 6.) It follows from the additivity property of the winding number that

$$N_i + N_f - N_s = 1 \tag{20}$$

where N_i is the number of sources, that is, singular points where the object is convex; N_f is the number of sinks (concave singular points on the object); and N_s counts the number of saddle singular points. Since the total number of singular points is known from the image, *this equation fixes uniquely the number of saddle points in the image*. Also, it implies that the image contains at least one source or sink.

Note as an aside that equation (20) implies also that the total number of singular points in the image must be odd. If there is an even number of nondegenerate singular points, the image cannot correspond to a smooth object that does not fold upon itself. This result excludes surface solutions that are concave at the occluding boundary, as well as convex ones. It has also been shown using a different method in (Horn et al. 1990). The only surface solutions possible are physically unreasonable ones in which the surface changes from convex to concave at some point along the occluding boundary. More 'impossibility' results will be presented later.

6.3 Properties of Generic Images

Next, we justify our restriction to images containing finite numbers of singular points, all of which are nondegenerate. Based on some ideas of dynamical systems theory, it is proven that images with this property are generic among all smooth images. In fact, we prove slightly more than this.

We define a closed surface to be *structurally stable* if its image satisfies the following three properties: (1) There are a finite number of singular points. (2) These are all nondegenerate. (3) No two saddle singular points

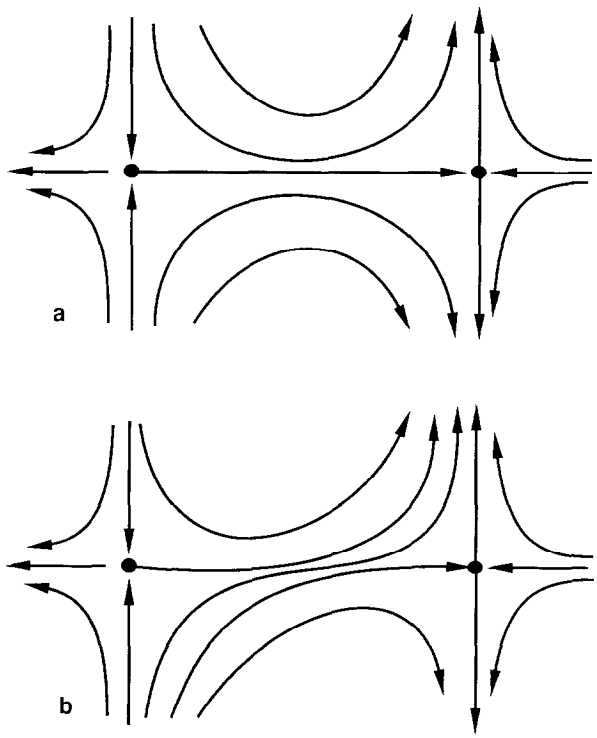


Fig. 12. An example of a nonstructurally stable flow, with two saddle points connected by a characteristics strip, is shown in 12a. An infinitesimal perturbation of the imaged surface will result in the situation shown in 12b, in which the two saddles are no longer connected.

are connected together by a gradient curve (figure 12). These surfaces can be shown to be stable in the sense that the image flow of gradient curves does not change drastically when the surface is perturbed—its topology remains the same. The Palis–Smale theorem (Palis & Smale 1968; Hirsch & Smale 1974; Palis & de Melo 1982) states in part the following:

> Let v be a C^r-smooth gradient vector field (with a continuous rth derivative), defined on an open region W of the plane, such that v points outward on the boundary δD of an open set D contained in W. Then, the set of all structurally stable v is open and dense in the set of all C^r gradient vector fields that point outward on δD (see figure 13).

A gradient vector field ∇z defined on the xy plane is equivalent to a surface $z(x, y)$. Thus, this theorem essentially states that any surface can be approximated by a structurally stable one, and that a perturbation of a structurally stable surface is also structurally stable. For a given image, W can be identified with the interior image region where the intensity is nonvanishing. For the surfaces we are considering, we showed above that

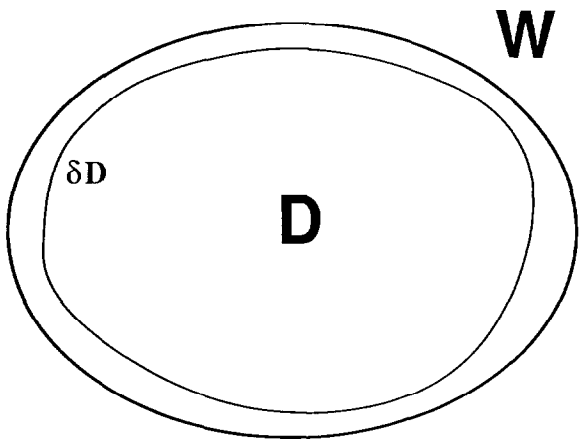


Fig. 13. An open region W of the image plane containing a closed curve δD, which is defined so that the characteristic strip flow is outward on this curve.

the vector field ∇z points outward near the limb. Also, the magnitude of this vector field becomes infinite on the occluding boundary. Thus, it is clear (and demonstrated below) that a contour δD in the image interior can be found such that the vector field also points outward on this line. δD must simply be chosen close enough to the limb. The Palis-Smale theorem therefore applies to the surfaces we consider.

If a given surface is not structurally stable, the theorem states that an infinitesimal perturbation will render it structurally stable. Since δD can clearly be chosen so that there are no singular points outside it, and since the flow is outward on δD , it is clear that this perturbation can be chosen (or modified) so that it vanishes outside δD, and does not alter the occluding boundary. Also, for a small enough perturbation of a surface, the corresponding vector field will clearly continue to point outward on δD since it is very large there—that is, the surface will continue to be convex on this curve. The Palis-Smale theorem therefore implies that a small enough perturbation of a structurally stable surface is also structurally stable. Structurally stable surfaces are therefore generic, and comprise essentially all surfaces. This justifies our restriction to images with a finite number of nondegenerate singular points.

It can be proved that a contour δD with the right properties can be found: First, at each point b on the limb, there is an open region R_b centered at this point such that throughout R_b the direction of the vector field does not deviate from the normal to the limb at b by more than an angle $d\theta << 1$. This is true because of the smoothness of the vector field direction p/q near the boundary, which was shown in section 6.2. Similarly, there is an open interval of the limb such that the limb tangent direction does not vary by more than $d\theta$ over this interval. This follows from the fact that the limb is smooth (Giblin & Weiss 1987). Therefore there is clearly an open interval I_b of the limb around b, and some distance ϵ_b, such that there is a smooth contour transversal to the vector field joining any two points whose distances from the boundary points of I_b are less than ϵ_b. Since the limb is compact, there is a finite subset of such open intervals that covers the boundary. Let ϵ_{min} be the least of the ϵ_b associated with these intervals. Then one can clearly patch together contours with endpoints less than ϵ_{min} from the limb, corresponding to the different intervals, to form a smooth closed contour that is everywhere transversal to the vector field. Q.E.D.

A direct proof that structurally stable surfaces are generic among all closed surfaces, which dispenses with the necessity for finding a contour δD, can be found in Oliensis (1990).

Condition 3 of the definition of structural stability is useful because it implies that if a gradient curve connects two singular points, then one must be elliptical; elliptical singular points uniquely determine the surface solution along strips connected to them (Bruss 1982; Saxberg 1989a). Although this property is not needed here (its assumption would simplify the proof, however), it is useful in the general illumination direction case (Oliensis 1990).

Finally, it can easily be shown that, at a singular point, if the two principal curvatures of the imaged surface are unequal in magnitude, then condition 5 of the uniqueness theorem is satisfied. Thus, condition 5 clearly holds generically—any structurally stable surface violating this condition can be perturbed infinitesimally at the offending singular point, so that the two principal curvature magnitudes, and, consequently, the two eigenvalues of the second-derivative intensity matrix, are no longer exactly equal.

7 Proof of Uniqueness: I

In the previous section, it was shown that for surface solutions convex at the occluding boundary, the gradient curves of the corresponding flow all connect to singular points in the image interior. From the Grobman-Hartman theorem, only four gradient curves connect to each saddle singular point, while an infinite number connect to each source and sink. By our

generically valid assumption, the image has just a finite number of singular points. Thus, for a consistent solution, *all points in the image apart from those lying on a finite number of gradient curves connect to sources and sinks.* We will argue that the solution is fixed at all points connected to sources and sinks. Then z is also determined on the isolated lines of points connected to saddle singular points, by continuity.

In this section, we assume that the type of each singular point is known—each is identified as either a source, sink, or saddle. We prove, based on this assumption, that the imaged surface is uniquely determined. In section 8, it will be proven that the type of each singular point is also uniquely fixed.

Bruss (1982) proved that the surface solution in the neighborhood of a source or sink is unique up to an additive constant in z, that is, up to an overall translation in depth. If the intensity function is r times differentiable, that is, C^r, the local solutions are C^{r-1}, from the stable manifold theorem (Palis & de Melo 1982). Assume $r \geq 2$. Since p and q—the first derivatives of z—are uniquely determined, the existence and uniqueness theorems for differential equations applied to equation (2) state that the characteristic trajectories can be uniquely extended from starting points in the neighborhood of the elliptical point. This determines z uniquely, up to an overall constant, over the region of the image plane connected to this singular point. The unknown constant can be thought of as the depth of the elliptical point itself.

From the above, the depths of essentially all image points are known relative to the depths of the elliptical points. To show that the surface is uniquely determined, all that remains is to show that the relative depths of the elliptical points are uniquely fixed. We prove this based on the following:

LEMMA: *For a consistent surface solution, every singular point is connected to every other singular point by a sequence of gradient curves.*

Proof: Note that this proof concerns the properties of surfaces, and assumes that a particular surface is given—it deals with gradient curves as solutions of the two-dimensional dynamical system equation (10).

Assume there is more than one singular point. We prove first that, for a consistent solution, every singular point is connected to at least one other singular point by a gradient curve. This is clearly true for any saddle point or sink. These singular points are terminal points of gradient curves, which must originate at singular points, since they cannot originate at the boundary. This immediately implies that a sink must be connected to some other singular point. Since gradient curves cannot be closed curves, one cannot begin and end at the same saddle point. Thus, saddle points also must be connected to some other singular point.

The only remaining case is that of sources. Suppose there exists a source O that only connects to the boundary. The region U_o consisting of all points in the image interior that lie on a gradient curve originating at O is open. This is so because the unstable manifold of the source point, for the dynamical system equation (10) defined over any open set contained in the image interior, is open (Palis & de Melo 1982). By assumption, there are points in the image interior not contained in U_o, for example all the other singular points. Thus the boundary of U_o must contain points in the image interior. Consider a gradient curve passing through one of these points. Every point on this curve must also be on the boundary by continuity of the flow. One end of this curve must connect to a singular point s, which is also on the boundary. First, s cannot be a source. If it were, then all points in some neighborhood must originate from s. But since s is on the boundary of U_o, there are points infinitesimally close to s that originate from O (see figure 14). Similarly, s cannot be a sink, since then all points in some neighborhood would terminate at s. But there are points infinitesimally close to s that originate at O, implying that O is connected to s contrary to assumption.

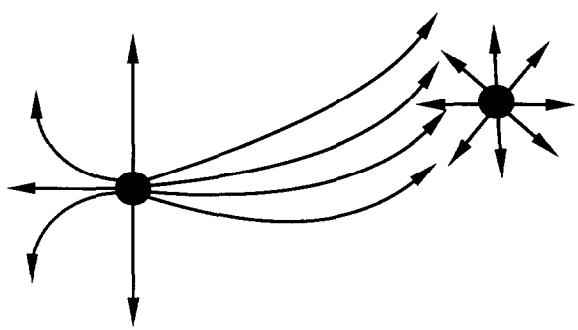


Fig. 14. A source cannot lie on the boundary of the region U_o of charateristic strips connected to another source O.

Suppose then that s is a saddle. It cannot be an isolated boundary point. For, suppose there is a neighborhood of s that excluding s itself contains only points of U_o. Then since there exist gradient curves converging to s, s is connected to O contrary to assumption.

On the other hand, if every neighborhood of s contains both points in U_o and points not in U_o, then there are boundary points of U_o infinitesimally close to s. Consider the gradient curves through a sequence of these boundary points approaching s. Since s is a nondegenerate and therefore an isolated singular point, one can assume that none of these points is singular. Suppose first that they constitute an infinite number of different gradient curves. All of them must connect to singular points. But only a finite number can connect to saddle points, so some must connect to either a source or a sink, which by the previous arguments is impossible. But if there is a finite number of gradient curves, then some curve must contain an infinite subsequence of these boundary points converging to s. This gradient curve L therefore connects to s, and is in the boundary of U_o. It is one of the four special curves that connect to s. (These curves constitute the stable and unstable manifolds associated with s for the two-dimensional gradient dynamical system in the image plane.)

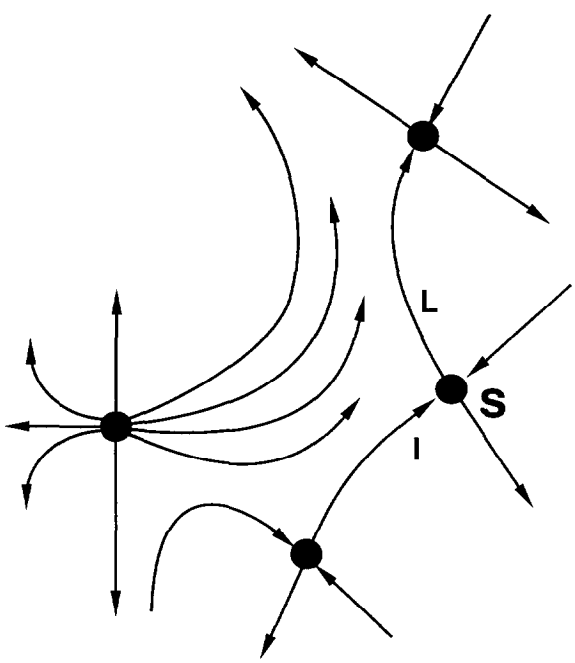


Fig. 15. A sequence of saddle points connected by characteristic strips. These are assumed to form part of the boundary of a region U_o connected to a source O. If the saddle point s is not connected to O, then there are two characteristic strips I and L on the boundary of U_o, which terminate and originate at s, respectively.

We will show that a second gradient curve I connecting to s must also be on the boundary of U_o, as shown in figure 15. Moreover, assuming that s is not connected to O, then I and L converge in the opposite sense to s—that is, assuming for convenience that L originates at s, I must terminate at s. In other words, the depth z along the composite curve $I + s + L$ is increasing. By the argument above, L and I cannot connect at their other ends to elliptical points, since these would also be on the boundary. They can only connect to other saddle points, as shown in figure 15, or else the limb. There is therefore a sequence of gradient curves on the boundary of U_o joining saddle points, along which the depth is continuously increasing. Since there are only a finite number of saddle points, this sequence must eventually end either at an elliptical point, which is ruled out, or the limb. But it cannot originate at the limb either, since all gradient curves exit at the limb. The only possibility is that one of the saddle points does connect to O. This shows that every singular point is connected to some other singular point as claimed.

It remains to show that there exists a gradient curve I as above. The Grobman-Hartman result that there exists a ball V around s such that the flow is homeomorphic to a linear flow on V is used. This homeomorphism can be taken to be a finite distance from the identity map. See Palis and de Melo (1982). V is divided into four *quadrants* by the four special curves mentioned above, one of which is L. Let Q be one of these quadrants bordering L, which contains points of U_o arbitrarily close to L. Let I be the other gradient curve bounding quadrant Q. I must converge to s (see figure 16). It will be shown that I is also contained in the boundary of U_o. All that is necessary is to show that some point i from I is on the boundary.

Let l be a point from L. l is chosen so that there are points of U_o arbitrarily close to l. Choose i from I, with i not on the boundary of U_o. Then there is a neighborhood of i which contains no point of U_o. The Grobman-Hartman theorem implies that one can choose new coordinates u, v on the region V around s such that: (a) the curve I converging to s is on the u-axis; (b) L is on the v-axis; (c) other gradient curves, characterized by constants c_1 and c_2, are given by

$$(u(t),\ v(t)) = (c_1 e^{-at},\ c_2 e^{bt})$$

where a, b are universal positive constants, valid for every gradient curve; (d) these coordinates are valid for all u, v with $u^2 + v^2 \leq D$ for some distance D; and (e) u, v are continuous bounded functions of x, y. For concreteness, Q is taken to correspond to the upper left quadrant in the u, v plane, as in figure 16.

Let i and l be identified respectively with the points $(i, 0)$ and $(0, l)$. These points are chosen to lie in the

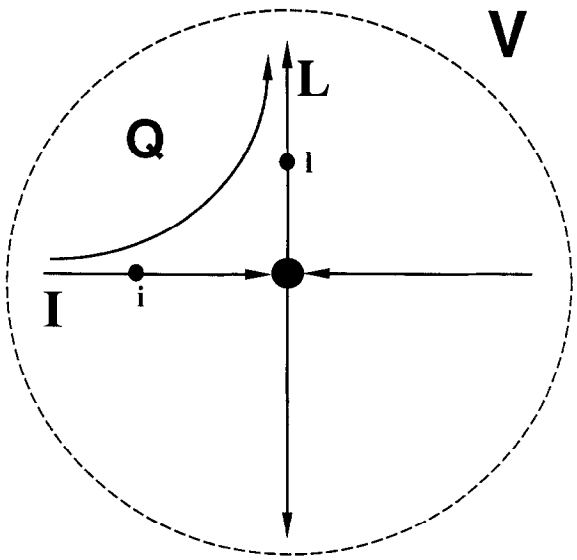


Fig. 16. An image-plane flow of characteristic strips around a saddle point, in linearized coordinates. V is the region over which the linearization is valid. Q is a quadrant bounded by the gradient curves L and I. Gradient curves that pass infinitesimally close to a point l on L also pass infinitesimally close to a point i on I.

region V where the linearizing coordinates are valid. It is easy to see from the above equation that for any ϵ, there exists a δ such that all points within a distance δ from l lie on gradient curves that pass within a distance ϵ of i. Since the change of coordinates is continuous, there does exist an ϵ such that no point in Q within ϵ of i is contained in U_o. All points on the same curve as a point not in U_o are also not in U_o. Therefore, there is a Q neighborhood of l that contains no point of U_o, contrary to assumption. I is accordingly contained in the boundary of U_o. Since this result holds for every saddle point on the boundary, the argument given previously shows that one of the saddle points on the boundary must connect to O.

It is therefore true as claimed that every singular point is connected to at least one of the other singular points. Next, it is shown that there is a path along gradient curves between any two singular points. Consider the set C of all singular points connected to a particular singular point O along some sequence of paths. Clearly, all the points in C are connected to each other through O. Suppose that this set is not the complete set of all singular points. Then the remaining singular points form a set N none of which is connected to any point of C. Define the set $\hat{C}$ to consist of all points in the image connected to the singular points C. Similarly, define $\hat{N}$ to be the set of all points in the image connected to the singular points N. Every point in the image is either in $\hat{C}$ or $\hat{N}$, since all points connect to some singular point. $\hat{C}$ contains the union U of a number of open sets consisting of those points in the image interior lying on gradient curves connecting to the sources and sinks in C. Exactly the same argument as above shows that the boundary of U can contain no point connected to a singular point in N. (If not, then some point n in N is on the boundary of U. As before, n must be a saddle, since otherwise it connects to an elliptical point in U, and there is a sequence on the boundary of connected saddles in N which eventually connects to the limb. This gives a contradiction, since gradient curves at the limb exit only.)

The closure $\bar{U}$ of U therefore contains, besides points on the image boundary, only the saddle points in C and gradient curves connecting to these saddle points. Moreover, it must contain all the curves connecting to saddle points in C. For, if it does not contain one of these curves, then there is a neighborhood of a point s on the curve entirely disconnected from the sources and sinks of C. Thus, there are points infinitesimally close to s connected to the sources and sinks of N, that is, s is a boundary point for the (un)stable manifold of one of these points. Again, by the arguments above, this is ruled out. Therefore $\bar{U} = \hat{C}$.

A similar argument applies to $\hat{N}$. Define W to be the union of all points connected to the sources and sinks of N. Then $\bar{W} = \hat{N}$. The compact connected image region has thus been shown to be the union of two disjoint closures of open sets. But this is impossible, and our original assumption is contradicted. This proves the lemma.

Finally, it is shown that the relative depths of all the singular points are determined. When an elliptical point is connected to a second singular point, the relative depth of the two points is uniquely determined, since, from Bruss's theorem, z is uniquely determined along the gradient curve connecting them.

Consider the remaining case of a gradient curve G connecting two saddle points, as shown in figure 17. Let T be a line of constant z intersecting this curve at g. In a neighborhood of g, every point on T lies on a different gradient curve. Since there are only a finite number of gradient curves connecting to saddles, there is a neighborhood of g which just contains (besides G) gradient curves connecting to elliptical points. By continuity of the flow, there are thus gradient curves connecting to elliptical points arbitrarily close to G, and to the saddle points at either end of G. The relative depths of any two points on a gradient curve connecting

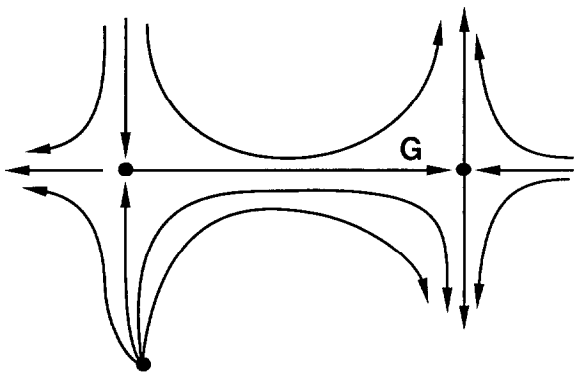


Fig. 17. Two saddle points connected by a gradient curve G. Their relative depth is determined.

to an elliptical point are determined. Thus, by continuity, the relative depth of the two saddle points connected by G is determined.

It has been shown that the relative depth of any two singular points connected by a gradient curve is determined. Since by the lemma all singular points are connected by sequences of gradient curves, the relative depths of all singular points are determined, and the surface is uniquely determined.

8 Proof of Uniqueness: II

The previous section demonstrated that if the nature of the surface at the singular points is known (whether concave, convex, or saddle-shaped), then the surface solution is uniquely determined up to an overall additive constant, assuming it exists. In this section it is shown that the nature of the surface at the singular points is also uniquely determined. This implies that the solution, if it exists, is determined completely uniquely.

8.1 The Saddle Points Are Uniquely Determined

First, it is shown that the singular saddle points are uniquely identifiable as such. Suppose that there are two solutions in which some number m of the saddle points in solution A become elliptical in solution B. Because of the topological formula,

$$N_i + N_f - N_s = 1 \tag{21}$$

the total number of saddle points is fixed. Thus m elliptical points in solution A must also become saddles in solution B.

Divide the singular points in the image into two classes: Y, containing the points that change from elliptical to saddle, or vice versa, between A and B, and N, containing those that do not. By assumption, Y is nonempty and, from the above, contains an even number of singular points. Since the total number is odd, N is also nonempty. Moreover, N must contain at least one elliptical point, since the contribution of Y to the sum in the topological formula above is zero. The sets Y and N are connected by gradient curves in both solutions A and B, from the arguments in the previous section.

The plan of the following proof is to first characterize the region of all points connected to the singular points of Y and, more specifically, the boundary of this region. (This can be done for either of the possible solutions. For convenience, we focus on one of them, e.g., A.) Then, essentially by counting the different types of singular points, we demonstrate that regions with the established properties are not consistent. Thus, Y must be empty, and the saddle points are uniquely determined.

8.2 Characterization of the Region Connected to Y

Let U be the open region of all points in the image interior lying on gradient curves connecting to the elliptical points of Y. As discussed in section 7, the boundary of U consists of gradient curves joining singular points (and possibly segments of the limb). These singular points cannot be elliptical. If they were they would be in N, and also connected to the elliptical points in Y by the arguments of section 7. However, we claim that elliptical points in N do not connect to the singular points of Y. Therefore, *all singular points in the boundary of U are saddles.*

Proof of Claim: This result is just the one quoted in section 5: an elliptical point that is connected to a second singular point determines the character of that point.

Suppose that an elliptical point n in N connects to a singular point y in Y. n will connect to y in both solutions A and B, since the assumption that it is elliptical determines the gradient curves connected to n uniquely. (It is easy to see from equation (2) that the gradient curves connecting to an elliptical point are exactly the same image curves independent of whether the point is a source or a sink—however, the *direction* of the curves alters between the two possibilities.) Now y is elliptical in one of the two solutions (say A), and in that case there is an open region around y such that every point contained in it connects to y. Consider the intersection of this open region with the open set of all points connected to n. The intersection is nonempty

by assumption, and must be open. Thus there are infinitely many gradient curves connecting to n, which also connect to y; and these gradient curves are determined purely by the image intensity and the stipulation that n is elliptical. In B, y switches to a saddle, but the character of n and the image intensity are unchanged; and therefore there will continue to be an infinite number of gradient curves connecting y and n, which violates the assumption that y is now a saddle. Q.E.D.

This argument is illustrated in figure 18.

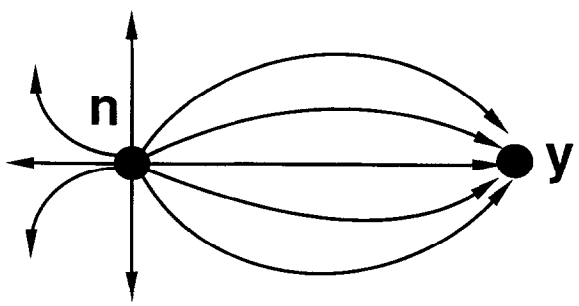


Fig. 18. An elliptical point connected by an infinite number of characteristic strips to a second singular point. The second point is determined to be elliptical as well.

The following lemma is also needed:

LEMMA. *For any quadrant of a saddle point s, there is a neighborhood of the saddle such that its intersection with the quadrant consists of gradient curves all connecting to the same elliptical point.*

A quadrant as defined here refers to the region bounded between neighboring arms of s—that is, between two of the gradient curves connected to s. It does not include the bounding gradient curves themselves.

Proof: The boundary of a region connected to some elliptical point consists of gradient curves joining singular points (apart from limb segments). We claim that all the gradient curves in this boundary must connect at one end to a saddle.

Proof of Claim: For a source, for instance, the boundary can contain no sources, by the arguments of section 7 (figure 14). Therefore, a bounding gradient curve terminating at a sink at one end must originate at a saddle—it cannot originate on the limb since the direction of all gradient curves there is outward. For a sink, the region connected to the sink is isolated from the limb, so again gradient curves on the boundary originating at a source must terminate at a saddle. Q.E.D.

There are an infinite number of different gradient curves passing arbitrarily close to s without connecting to it—consider for instance the curves intersecting some line of constant z beginning at s. Since there are only a finite number of curves connected to saddles, there is a neighborhood of s containing no gradient curves connected to saddles (apart from the four gradient curves connected to s). Consider the intersection of a quadrant with this neighborhood. This entire region must connect to the same elliptical point, since otherwise it would have to contain a boundary between the curves connecting to different elliptical points, which we have just shown is not the case. Q.E.D.

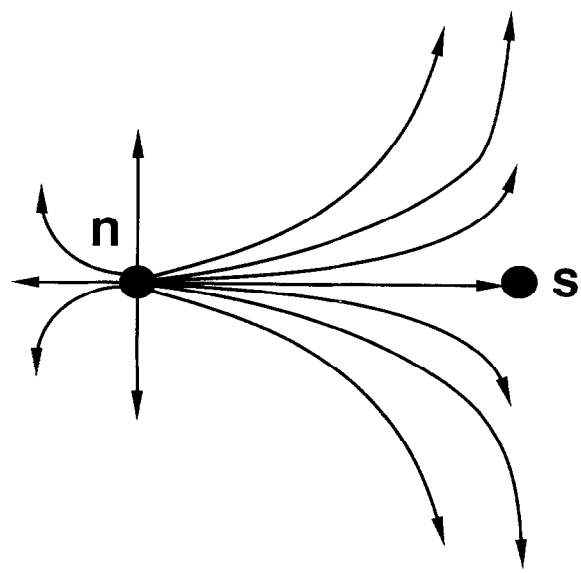


Fig. 19. An elliptical point connected by a single characteristic strip to a second singular point. The second point is determined to be a saddle point.

Now let s be a saddle point from Y. We claim that there is a neighborhood of s which contains no gradient curves connecting to elliptical points of N. Otherwise there would be an infinite number of such curves passing arbitrarily close to s without connecting to it, in both solutions A and B. But this contradicts the assumption that s becomes an elliptical point in the alternate solution (see figure 19). Therefore, from the above, there is a neighborhood of s containing just points from U, apart from the gradient curves connecting to s. By continuity of the flow, s, and the gradient curves connecting to s, are contained in the boundary

of U. Moreover, *s and its connecting curves are contained in the interior of the closure of U* (see figure 20).

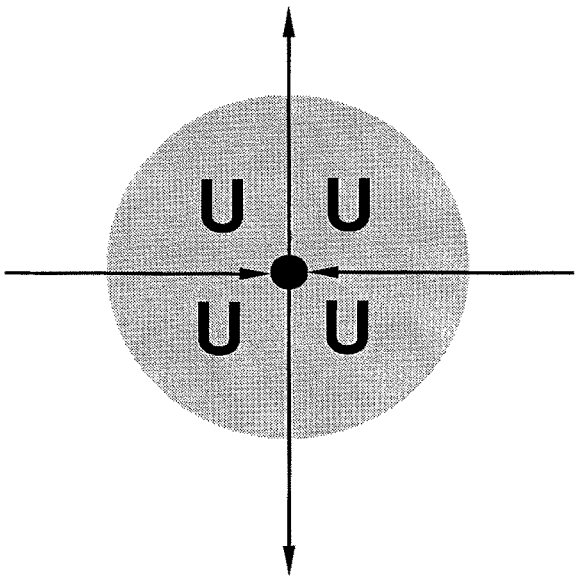


Fig. 20. A neighborhood of a saddle point with all quadrants contained in the set U. The saddle is therefore contained in the interior of the closure of U.

Next consider a saddle point s from N on the boundary of U. It is possible that, as before, there exists a neighborhood of s containing only points from U (figure 20) in addition to s itself and its connecting gradient curves. In this case, s and these curves are again contained in the interior of the closure of U. The other possibility is that at least one of the quadrants of s is filled up by gradient curves connecting to an elliptical point of N (in some neighborhood of s). In this case, s is on the boundary of the closure of U. There are various possibilities depending on which quadrants connect to N, and which to Y; these are illustrated in figure 21.

Denote the closure of U by $\hat{Y}$, and let δY denote its boundary. Since δY is contained in the boundary of U, it consists of gradient curves joining saddles. It was shown above that all saddles in Y are contained in the interior of $\hat{Y}$. Thus, *all the saddles on the boundary of $\hat{Y}$ are in N.*

The remainder of the proof shows that configurations with $\hat{Y}$ surrounded by saddles from N are impossible. Essentially, we use the winding number to count singular points in $\hat{Y}$, and show that the number of saddles and elliptical points in Y cannot be equal—a contradiction.

8.3 The Configurations $\hat{Y}$ Are not Consistent

We consider first the simplest case: a connected component $\hat{Y}_C$ of $\hat{Y}$ that is isolated from the limb, and simply connected. Since $\hat{Y}_C$ is simply connected, its boundary is a single closed circuit of gradient curves connecting saddle points from N. Inside this boundary there are no holes—no regions connected to elliptical points of N, which would require internal boundaries. We must compute the winding number of this region.

Consider the circuit C around $\hat{Y}_C$ illustrated in figure 22. The segments of this curve fall into two categories: either (1) they lie on gradient curves, or else (2) they lie on curves of constant depth (level contours) which run perpendicular to the gradient curves. The sharp corners between segments should be thought of as representing infinitesimal smooth joining curves. C runs along gradient curves paralleling the bounding curves connecting saddle points. In the vicinity of the saddle points themselves, C crosses the external strips connecting to these saddles by means of the level contour segments.

The winding number will be computed on C by keeping track of the relative orientation of the vector field ∇z to the curve tangent. The curve tangent makes exactly one revolution as C is traversed, equivalent to a winding number $+1$. We measure orientation of vectors using an angle that increases in the counterclockwise direction, that is, $\theta = \tan^{-1}(V_y/V_x)$ for a vector $\vec{V}$. We also take C to be traversed in the left-hand or clockwise sense, so that the vector normal to the curve and pointing outward is rotated by $+90°$ with respect to the tangent vector on C, which points in the direction of motion along C.

Along the segments in category 1, the vector field ∇z and the tangent to C are parallel or antiparallel. Therefore, along these segments the vector field and tangent rotate through the same angle. Also, along the level contour segments of category 2, the tangent vector and vector field are perpendicular, and again rotate through the same angle. Therefore, the tangent direction and the vector field rotate by different amounts only at the joins (corners) between the gradient curve and level contour segments. We assume that the joining curves are sufficiently small so that the vector field is essentially constant along these curves. Since the gradient curves are perpendicular to the level contours, the tangent direction changes along these joins by approximately $\pm 90°$. The total relative rotation along C of the tangent direction with respect to the vector field is

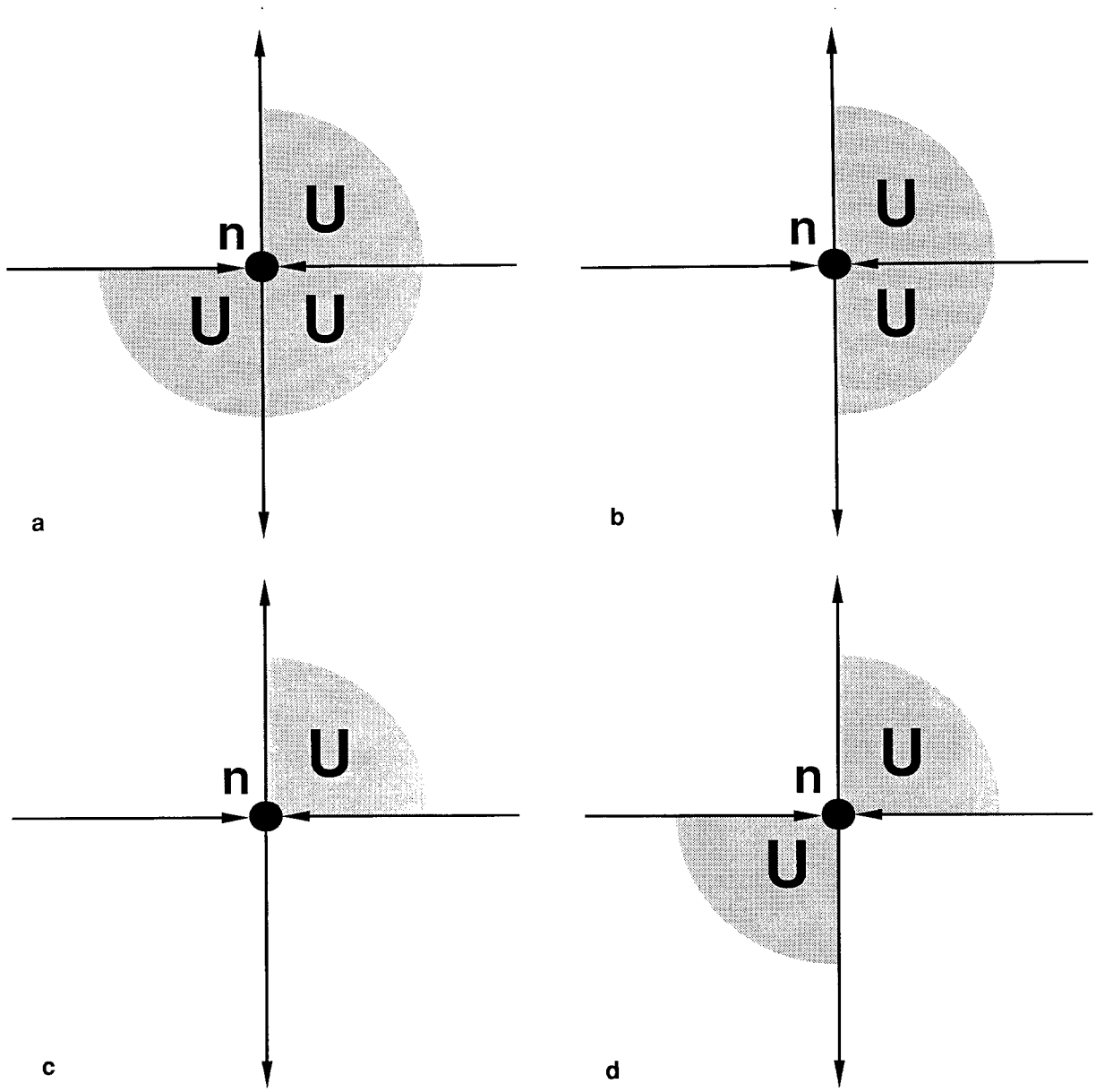


Fig. 21. The various possibilities for a saddle point s on the boundary of the closure of the set U. See text.

therefore given by summing the 90° relative rotations over all the joins. Essentially, we must compute the relative rotation induced by the passage around each saddle point on the boundary.

First, the case of a saddle with a single external connecting gradient curve is considered, such as one of the saddles labeled s_1 in figure 22. The saddle, and part of its connecting gradient curve, are inside C. Thus, the first 90° turn is toward the inside direction, since C immediately intersects this curve. The tangent direction rotates by −90° with respect to the vector field, which remains approximately constant in direction during the turn. After the next turn, the tangent direction has clearly rotated by −180° with respect to the vector field.

Next we consider a saddle with two external lines, for example, s_2 in figure 22. Clearly, the rotation here is just twice as much as for the previous case—two cuts to the inside are necessary. Thus, the relative rotation is −360°.

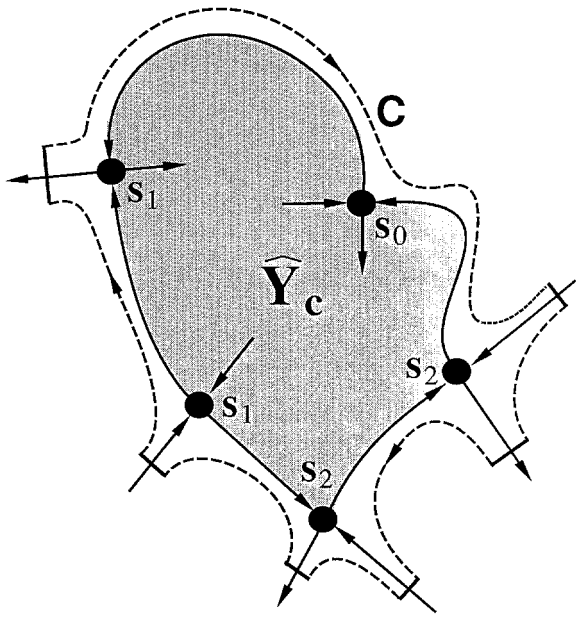


Fig. 22. A connected region $\hat{Y}_C$ containing image strips connected to elliptical singular points in Y. Its boundary consists of saddle points from N connected by characteristic strips. The winding number of this region is computed on a contour C. The dashed segments of C lie on characteristic strips, and the thick solid lines represent level contour segments running perpendicular to the characteristic strip flow.

Finally, it is clear that when the saddle has no external connecting curves, like s_0 in figure 22, one can choose C to lie on the gradient curve bypassing the saddle, and there is no relative rotation.

Let the number of saddle points on the boundary of $\hat{Y}_C$ with one external line be denoted S_1, and let S_2 be the number with two. For figure 22, clearly $S_1 = 2$, and $S_2 = S_0 = 1$. A saddle point with one external connection represents an extremum of z along the boundary, while for the other two possibilities the depth increases monotonically through the saddle point in the flow along the boundary. Since the depth on the boundary has a maximum and a minimum, there are at least two saddle points with one external connection, that is, $S_1 \geq 2$. Also, between any two maxima there must be a minima, and vice versa, so that S_1 is even.

From the above:

$$\bar{\theta}_{\text{tan}} = -360° = \bar{\theta}_{\text{vec}} - 180° S_1 - 360° S_2 \qquad (22)$$

where $\bar{\theta}_{\text{tan}}$ is the total rotation of the tangent direction along C, and $\bar{\theta}_{\text{vcc}}$ is similarly the total rotation of the vector field. Expressing this relation in terms of the winding number W of $\hat{Y}_C$, with $W \equiv \bar{\theta}_{\text{tan}}/(-360°)$, gives

$$W = 1 - \frac{S_1}{2} - S_2 \leq 0 \qquad (23)$$

Note that we can choose C to lie along gradient curves connecting to elliptical points of N, and also along level contour segments perpendicular to these curves. Thus the value of W is the same for both solutions A and B.

From section 4, W is also equal to the sum of the indexes of all singular points contained within C. Thus

$$1 - \frac{S_1}{2} - S_2 = -S_1 - S_2 - S_0 - S_{\text{int}} - S' + E' \qquad (24)$$

where S_0 is the number of saddles from N on the boundary with zero external connections; S_{int} is the number of saddles from N in the interior of $\hat{Y}_C$; S' is the number of saddles from Y in $\hat{Y}_C$, and E' is the number of elliptical points from Y in $\hat{Y}_C$. Therefore

$$E' - S' = 1 + \frac{S_1}{2} + S_0 + S_{\text{int}} > 0 \qquad (25)$$

However, in the alternate solution, we have:

$$S' - E' = 1 + \frac{S_1}{2} + S_0 + S_{\text{int}} \qquad (26)$$

since the saddles in Y become elliptical points, and vice versa, while the character of the singular points in N is unchanged. But this implies $S' - E' = E' - S'$, which could only be satisfied if the difference were zero, which we have shown is not the case. Therefore, the configuration $\hat{Y}_C$ is impossible.

It remains to show that the other possibilities for $\hat{Y}$, for instance with $\hat{Y}$ bounded partly by the limb or with islands of N-connected points contained in $\hat{Y}$, can also be ruled out. The additional complications will be dealt with one by one. First, we consider an isolated region $\hat{Y}_B$ which is similar to $\hat{Y}_C$ except that it is bounded in part by the limb. This is illustrated in figure 23. The winding number of $\hat{Y}_B$ will be computed on a circuit C_{ext} defined similarly to the circuit C above. Since $\hat{Y}_B$ is bordered by the limb, however, we must consider how to define the segments of C_{ext} that run along the limb.

In both of the solutions A and B, the flow of gradient curves is outward on the limb. By a slight extension of an argument in section 6.3, it is possible to find a curve segment δD just inside the limb such that in both solutions the flow is outward on this segment. At every point along δD, therefore,

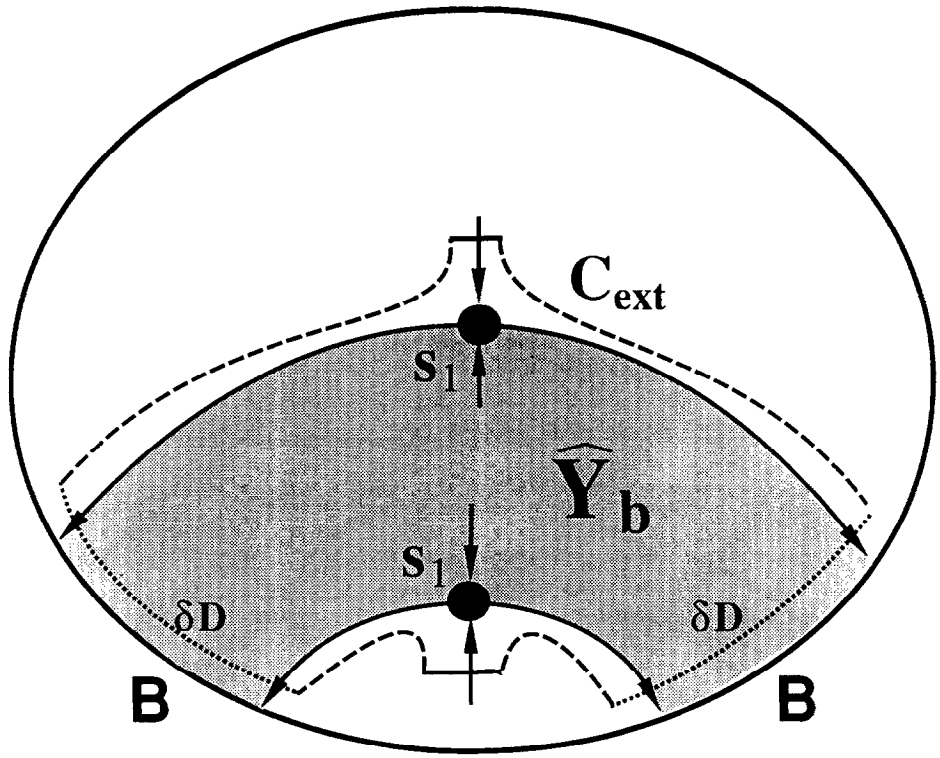


Fig. 23. A connected region $\hat{Y}_B$ of $\hat{Y}$ bounded in part by the limb. The winding number is computed on the indicated contour.

$$\theta_{tan} < \theta_{vec} < \theta_{tan} + 180° \tag{27}$$

where, for example, θ_{tan} gives the orientation of the tangent direction.

Near a point where it joins onto a gradient curve outside $\hat{Y}_B$, δD clearly can be chosen to lie on a level contour. Around such a point of juncture, one can assume that all gradient curves connect to a single elliptical point of N, and therefore the level contours near this point are the same in both solutions A and B. Thus the turns from the initial gradient curve segment to δD, and then from δD to the succeeding gradient curve, are both turns of $-90°$ to the inside. Also, at the beginning and end of δD, the tangent direction is at the same angle ($-90°$) with respect to the vector field direction. The tangent direction must therefore have rotated through a total angle $360°n$ with respect to the vector field, for some integer n. However, by the inequality in equation (27), the tangent direction is bounded to be within a range of the vector field, and it therefore could not have achieved a complete rotation relative to this field. Consequently, n is zero. In traversing through δD, the total rotation of the tangent direction with respect to the vector field is therefore $-180°$.

The generalization of equation (23) can now be written as

$$W = 1 - \frac{S_1}{2} - S_2 - \frac{B}{2} \tag{28}$$

where B counts the number of segments running close to the limb. Using the same reasoning as before,

$$E' - S' = 1 + \frac{S_1}{2} + S_0 + S_{int} - \frac{B}{2} \tag{29}$$

We claim that $S_1 \geq B$. The reason for this is that the direction of gradient curves is outward at the limb, so that the direction of flow along C_{ext} must reverse at least once between limb segments. This reversal occurs only at a saddle point with a single exterior connection, giving the inequality. As a result, it is again true that $E' - S' > 0$, which as before is impossible, so that the configuration $\hat{Y}_B$ is ruled out.

Next, the stipulation that the region of $\hat{Y}$ be without holes is relaxed. We consider a region $\hat{Y}_H$ which contains within its external boundary elliptical points from N. The regions spanned by gradient curves connecting to these elliptical points constitute the holes within $\hat{Y}_H$. This is illustrated in figure 24.

We focus on one of the holes $\hat{H}_i$. Its boundary consists of saddles from N connected by gradient curves, since it is also part of the boundary of $\hat{Y}$. We can assume that it is isolated from the limb, since otherwise it is not contained within $\hat{Y}_B$, and a contour surrounding $\hat{Y}_B$ can be found which excludes this region, such as the one shown in figure 23. We can also assume that there are no regions from $\hat{Y}$ contained in $\hat{H}_i$—otherwise we can simply apply our argument to these smaller regions of $\hat{Y}$.

The singular points in $\hat{H}_i$ can be counted as before by computing the winding number as shown in figure 24:

$$W_i = 1 - \frac{S_{1i}}{2} - S_{0i} \leq 0 \tag{30}$$

Note that gradient curves that are exterior to $\hat{H}_i$ are *interior* to $\hat{Y}$—this is why S_0 appears in the above equation rather than S_2. (S_0, as before, counts the saddles such that none of their connecting gradient curves is exterior to $\hat{Y}$.)

W_i can also be computed in a different way. The winding number can be computed on a curve just inside the boundary of $\hat{H}_i$, along the gradient curves which are the same for both solutions A and B. W_i is then given by the sum of this winding number, minus the number of saddle points on the boundary of $\hat{H}_i$. This method makes it clear that W_i is the same in both solutions. This must be the case, since W_i counts the indexes of points contained in N; the indexes of these points are the same in both solutions.

Calculating as before,

$$1 - \frac{S_1}{2} - S_2 - \frac{B}{2} = -S_1 - S_2 - S_0 - S_{\text{int}} + \sum_i W_i - S' + E' \tag{31}$$

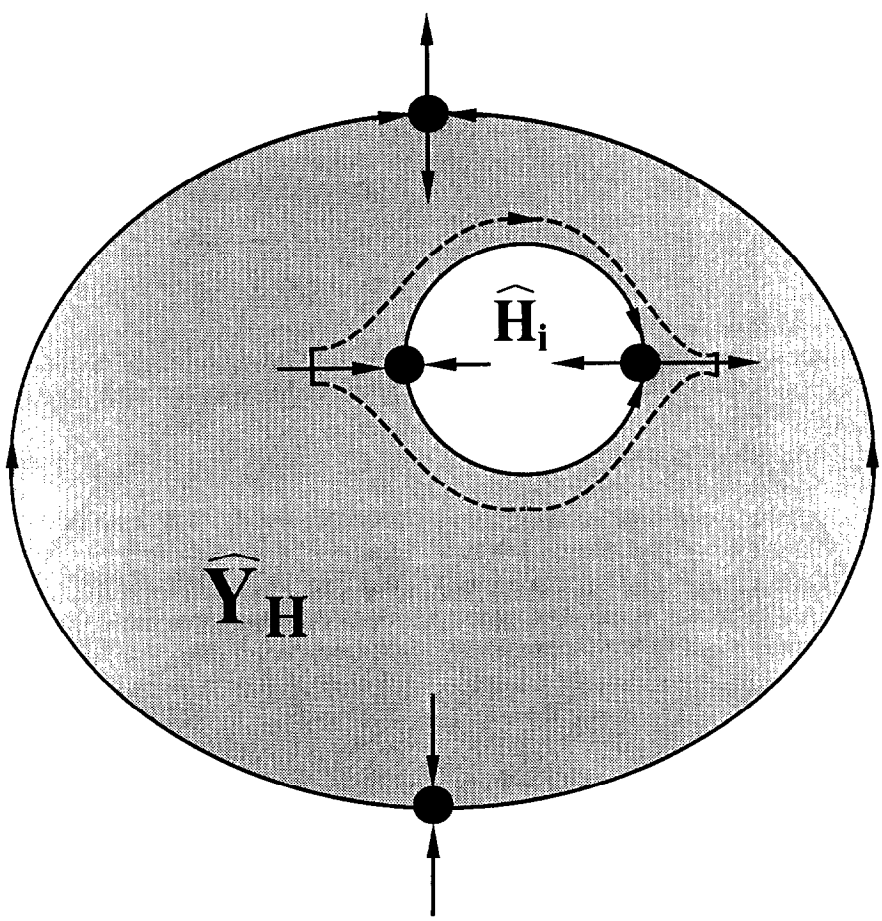


Fig. 24. A non-simply connected region $\hat{Y}_H$ of $\hat{Y}$ containing a region $\hat{H}_i$ connected to elliptical points in N. The winding number of the hole $\hat{H}_i$ is computed using the indicated contour.

Since the $W_i \leq 0$, an argument identical to that following equation (29) shows that the configuration $\hat{Y}_H$ is also ruled out.

Lastly, the effects of saddle points with four connecting gradient curves on the boundary of $\hat{Y}$, as in figure 21d, are considered. We claim that saddle points of this type introduce nothing qualitatively new. At such a saddle point, two regions of $\hat{Y}$ touch "accidentally." We will imagine separating these two regions as shown in figure 25, which simultaneously joins together the two regions in the complement of $\hat{Y}$, consisting of gradient curves connecting to N. The saddle point itself will be included as part of this complementary region, as illustrated in the figure. If this separation can be implemented consistently, then the counting of singular points is essentially the same as for the previous cases.

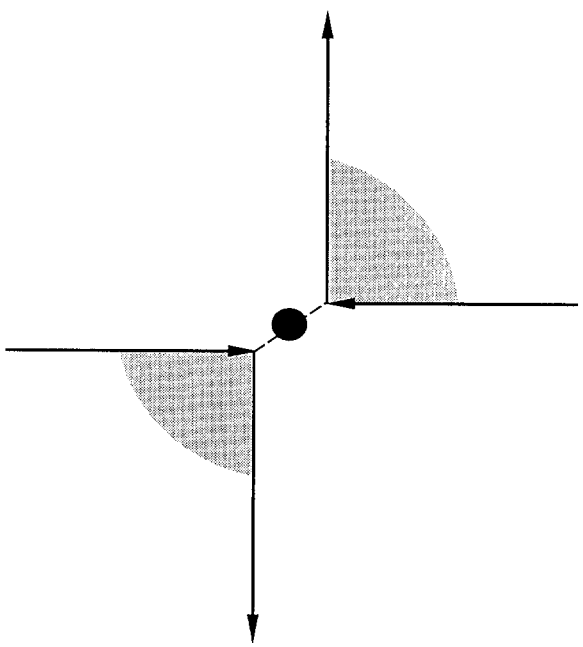


Fig. 25. Imaginary separation of two regions of $\hat{Y}$ accidentally touching at a saddle point of the type shown in figure 21d.

In figure 26, it is shown how to draw external contours C_{ext} so that a saddle point of this type is avoided. Clearly, there is no relative rotation of the tangent direction with respect to the vector field along the portion of C_{ext} indented away from the saddle. Thus, for contours so defined, the winding number is the same as if there were no saddle point and the two regions in figure 26 were separated.

If both of the non-$\hat{Y}$ regions are holes—that is, contained in a larger region of $\hat{Y}$—then it is necessary to

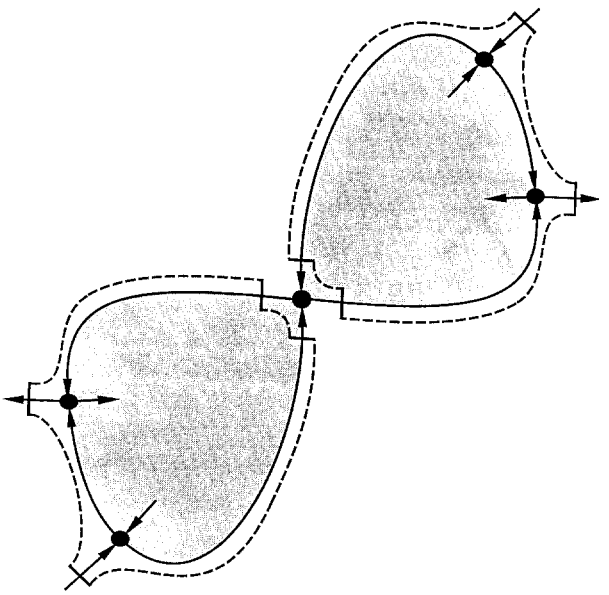


Fig. 26. A region of $\hat{Y}$ containing a saddle point of the type shown in figure 21d. The winding numbers of the separate regions are computed on the contours shown, avoiding the saddle point.

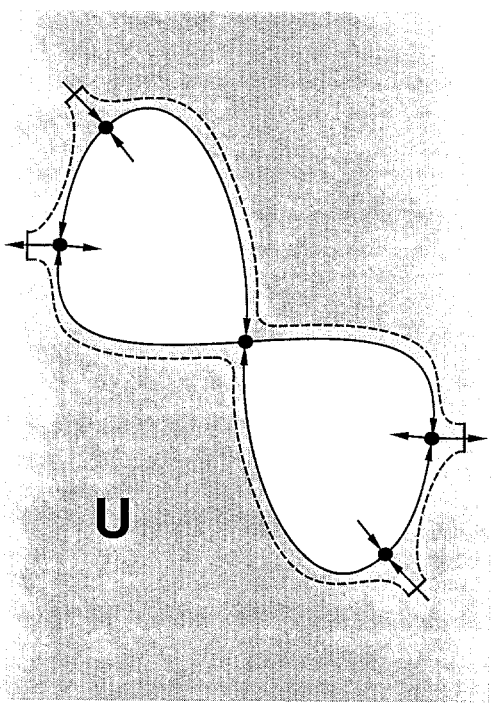


Fig. 27. A hole—containing image strips connected to elliptical points in N—in a region of $\hat{Y}$ containing a saddle point as shown in figure 21d. The winding number of the hole region is computed on the indicated contour. The saddle is counted as if contained in the hole region.

compute the sum of the indexes of the points contained in these. One can compute the winding number on a contour containing both holes as indicated in figure 27. Near the saddle, this contour falls on gradient curves that simply bypass the saddle point. Since the saddle has no external connecting curves, it induces no rotation of the tangent direction relative to the vector field. The W_i for the combined hole region is therefore unmodified:

$$W_i = 1 - \frac{S_{1i}}{2} - S_{0i} \tag{32}$$

Actually, we now have $S_{1i} \geq 4$ since there are at least two separate closed loops on the boundary of the combined hole, and for each loop $S_{1i} \geq 2$. Thus W_i is strictly less than zero for combination holes of this type. These arguments justify our assertion that saddles as in figure 21d can be taken into account as in the previous cases.

The final, generally valid, version of equations (23) and (29) is therefore

$$E' - S' = 1 + \frac{S_1}{2} + S_0 + S_{\text{int}} - \frac{B}{2} - \sum_i W_i \tag{33}$$

The same argument as before applies. Therefore $\hat{Y}$ cannot consistently exist, and the saddle points of the solution are uniquely determined. Recall that the gradient curves connected to an elliptical point are the same image curves, apart from the direction of flow, regardless of whether the point is a source or a sink. The result derived in this section therefore implies that the gradient curves are uniquely determined as curves in the image.

8.4 Sources and Sinks Are Uniquely Determined

It remains to demonstrate that the source and sink singular points are also uniquely assigned. From the above, the gradient curves are uniquely determined as image curves. Also, it has been shown that every singular point is connected by a sequence of gradient curves to every other singular point. Finally, there are clearly singular points connected to the limb, and the flow direction of the connecting curves is determined to be outward. Thus, every singular point is connected to the limb by a sequence of gradient curves.

At an elliptical point, the flow direction of all gradient curves connected to the point is the same. At a saddle point, the relative directions of the gradient curves connecting to the saddle are also determined—adjacent strips flow in opposite directions. Thus the flow direction is determined along any sequence of gradient curves, assuming the direction is determined on one of them. Since all singular points are connected by such a sequence to the limb, and since the flow direction is determined there, it is determined at every singular point. Sources and sinks therefore can be uniquely assigned. This concludes the proof.

The results obtained here can be partially extended to the case of more general light source direction (Oliensis 1990).

9 Impossibility of Solutions

By definition, an *impossible* image is one for which there exists no corresponding smooth, non-self-occluding surface of which it can be the image, at least for the given reflectance function. Recently, it has been shown that such images do exist (Horn et al. 1990); however, the image examples proved to be impossible by Horn et al. (1990a) are rather special. See also the argument at the end of section 6.2

An argument is presented here that indicates that images are effectively impossible *generically* (but see Note 1). An image is defined to be *effectively impossible* if the only possible corresponding surfaces are physically unreasonable and nongeneric (in a sense made precise below). For the class of images considered here, it is shown that any image can be modified by a small perturbation of its intensities so that the perturbed image is impossible, or at best has a nongeneric surface solution. Also, it is suggested that essentially every image (i.e., intensity function $I(x, y)$ of this class is effectively impossible. Our argument applies explicitly to reflectance functions that are approximately Bruss functions, with the illumination assumed symmetric around the viewing direction. However, an analogous result is very probably true for arbitrary reflectance functions. We also present a concrete example illustrating that a small intensity perturbation can render an image effectively impossible. Our arguments in this section are not at the level of proofs; a rigorous discussion will be presented in future work.

As stated in section 6, every point on a smooth surface clearly lies on a unique curve of steepest ascent. Correspondingly, for a consistent smooth surface solution corresponding to an image, every image point lies on a unique characteristic strip. *Thus image strips cannot intersect in the image plane.* (This argument is also valid for illumination from a general direction (Oliensis

1990).) Another way to show this is as follows: a characteristic strip, since it is the solution of the first-order differential equation (8), is uniquely determined through a point if p and q are known at the point. In fact, (p, q) is just the tangent to the image strip at the point, as noted in section 2. This implies that two different image strips can intersect at an image point only if they have different tangents. But this cannot happen either. The surface normal at a point is also specified by the values of p, q. Two tangents at one point means that the surface normal simultaneously has two different directions at the point, which is impossible for a smooth, non-self-occluding surface. Below, it is argued that an image can be perturbed in such a way that characteristic strips effectively *must* intersect in the image plane. The perturbed image is then an effectively impossible one.

As stated in section 2, the ensemble of image strips filling out the image plane determines, and is equivalent to, a surface solution (figure 1). Thus, in a consistent solution, all image strips have neighboring, infinitesimally close strips with which they are never allowed to intersect. However, for general solutions of the characteristic strip equations (3), there is no reason why different characteristic trajectories should not intersect when projected into the image plane. These statements are valid for arbitrary illumination, and, for this reason, our impossibility results probably extend to general illumination. Below, for Bruss reflectance functions, we show that it is in fact easy to perturb an image so that neighboring trajectories do intersect.

The Hamiltonian viewpoint, as discussed earlier in section 2, equates the trajectory of a characteristic strip with the motion of a point 'particle' in a potential, that is, a 'particle' moving in the image plane acted on by position-dependent but not time-dependent forces. The potential V represents nothing more than an encoding of the space-dependent forces acting on the 'particles' tracing out these trajectories. To cause two neighboring trajectories to cross, one simply adds forces over some region along the trajectories pushing them toward each other. This can be done by introducing a local *valley* in the potential in a region of the image that lies in the path of the two neighboring trajectories. The valley should be oriented so that its long axis is parallel to the trajectories, with the valley floor in between the two trajectories. From equation (8), the perturbation introduces an acceleration in the negative gradient direction, that is, in the downhill direction, of the perturbation. Effectively, like marbles rolling along the opposite slopes of a valley, the trajectories will feel the force of gravity, move downhill toward the valley floor, and therefore toward each other. If the valley walls are steep enough, both trajectories will reach the valley floor, and cross.

If the original trajectories are chosen very close together in the image plane, the valley that induces them to cross can be very narrow. Also, the valley need not be deep, nor long, as long as it is steep over the actual path of the trajectories. This is so since the forces on the 'particles' tracing out the trajectories are given by the *derivatives* of the potential. Also, if the original trajectories are very close, not much of a perturbation is necessary to cause them to cross. Thus, the perturbation of the intensities necessary to induce crossing can be quite small, and restricted to a small image region. Probably, crossing can be induced by an arbitrarily small perturbation.

The uniqueness results can be used to show that one can choose perturbations such that characteristic strips *must* cross in the image plane (at least in the generic case). Suppose that one of the singular points in the image is taken to be elliptical. As discussed in section 7, the image strips connected to the point are uniquely determined as curves in the image plane, by Bruss's theorem. (Only their direction alters depending on whether the singular point is a source or a sink.) Consider integrating the characteristic strips outward from the singular point, in the positive time direction if the point is a source, and in the negative time direction if it is a sink. This integration is uniquely determined. By introducing an intensity perturbation in the path of some of these strips, as described above, the strips can be induced to cross in the image plane. The intensity perturbation can be confined to a small region at some distance from the singular point, and the same perturbation will induce crossing regardless of whether the elliptical point is a source or a sink. Since the characteristic strips connected to an elliptical point are uniquely determined, this crossing of trajectories cannot be avoided: the perturbed elliptical solution is not consistent.

This procedure can be repeated for every singular point in the image. (We assume as before that the singular points are finite in number, and nondegenerate, as is generically the case.) For each singular point, the intensities can be perturbed in a different, small image region, such that for the elliptical solution around this point, the characteristic trajectories connected to it intersect in the image plane. Thus, for the perturbed image, since image strips are not allowed to cross in a consistent solution, no singular point can be consistently

interpreted as elliptical. In other words, *any* image containing a finite number of nondegenerate singular points can be modified by a small perturbation so that the surface solution can contain no elliptical singular points.

As before, let us focus on images generated from objects completely within the field of view. Then a surface solution which is closed, or in general position, is convex at the occluding boundary, and the surface point closest to the camera projects to a singular point in the interior of the image. This singular point must be a source. If the surface is taken to be concave all along the occluding boundary, so that its pose is accidental, there must similarly be a sink in the image interior. Then, from the above, *the perturbed image cannot correspond to any surface solution which is concave at the occluding boundary, or is in general position, or corresponds to a closed surface.* However, it does remain possible that a nongeneric solution exists, with all singular points saddles, and for which the surface must change from concave to convex at points along the occluding boundary. Whether these solutions can also be ruled out is not investigated further here. A surface that changes from convex to concave at the occluding boundary is a physically unreasonable solution. The perturbed image is effectively impossible in the sense that these unreasonable solutions are the only ones possible.

Characteristic trajectories depend continuously on the image intensities. The elliptical solutions around a singular point do as well (Palis & de Melo 1982). Thus, it is clear that if two image strips connected to an elliptical point cross, then slightly perturbing the image intensities will not affect this fact. Thus, a sufficiently small perturbation of an effectively impossible image will yield one that is also effectively impossible. Also, the characteristic trajectories depend continuously on the reflectance function. This is also true for the elliptical solutions around a singular point (Palis & de Melo 1982). Again, sufficiently small modifications of the reflectance function will not uncross intersecting trajectories. Thus an effectively impossible image remains so under perturbations of the assumed reflectance function.

Suppose that our conjecture above is valid, so that an infinitesimal perturbation is sufficient to induce trajectory crossing in any image flow of characteristic strips. Then, from the above, the class of effectively impossible images with crossed trajectories is stable under perturbation, while the images not contained in this class—the 'possible' images—can be perturbed infinitesimally so that the result is contained in this class. This implies that almost all images containing the complete limb are effectively impossible: the effectively impossible images are a *generic class* among these images.

These arguments are now illustrated with an explicit example. We consider the simple surface,

$$z = x^2 + 1.5y^2 \tag{34}$$

The viewing and illumination directions are along the *z-axis*, and the surface is taken to be Lambertian. The point closest to the camera occurs at $z = x = y = 0$, and gives rise to a singular point in the image. This singular point is a source for the given solution.

We will show that, assuming the singular point is elliptical, the characteristic trajectories connected to it, and uniquely determined by it, can be induced to cross in the image plane by a small, local perturbation of the intensity. The perturbation is chosen arbitrarily to be centered at $(x, y) = (2.6, 2.6)$, at some distance from the singular point. Various sizes of the perturbation were tried; all induced crossing of the image strips. At the chosen center point, the image-strip tangent for the exact solution is parallel to (2, 3). Thus, the perturbation is chosen to be a valley whose long axis is parallel to this tangent direction.

The unperturbed potential is

$$V = \frac{1}{2}\left(1 - \frac{1}{I^2}\right) \tag{35}$$

where I is the intensity. At the chosen center point, the unperturbed potential is approximately $V = -44$. The perturbation is of the form

$$\delta V \equiv -S \exp\left(\frac{1}{\left(\frac{x_{\text{par}}}{r_{\text{par}}}\right)^2 + \left(\frac{x_{\text{perp}}}{r_{\text{perp}}}\right)^2 - 1}\right) \tag{36}$$

Here x_{par} measures the displacement from the center point (2.6, 2.6) projected onto the long axis of the perturbation, namely the (2, 3) direction. x_{perp} measures the displacement projected onto the perpendicular direction. Explicitly,

$$x_{\text{par}} \equiv \frac{2(x - 2.6) + 3(y - 2.6)}{13^{1/2}}$$

$$x_{\text{perp}} \equiv \frac{-3(x - 2.6) + 2(y - 2.6)}{13^{1/2}} \tag{37}$$

Also, S gives the scale of the perturbation, and r_{par} and r_{perp} determine the size of the perturbation in the long axis and perpendicular directions, respectively. Note that the perturbation, and all of its derivatives, go to zero on the ellipse (from the inside):

$$\left(\frac{x_{\text{par}}}{r_{\text{par}}}\right)^2 + \left(\frac{x_{\text{perp}}}{r_{\text{perp}}}\right)^2 = 1 \qquad (38)$$

The perturbation is assumed to be zero outside this ellipse as well. Therefore, it is a local perturbation. This form of the perturbation was chosen since it is a C^∞ function, with all derivatives vanishing on a bounding ellipse. As a result, the perturbed potential $V + \delta V$ can be differentiated as many times as the original potential V.

If the singular point at $x = y = 0$ is assumed to be elliptical, this uniquely determines the surface solution in a local neighborhood of the point, by Bruss's theorem. (*Note:* elliptical singular points should not be confused with the arbitrarily chosen elliptical region inside which the perturbation is nonzero.) Then the characteristic strips can be uniquely extended from starting points near the elliptical point. The extension of a strip will differ from the exact solution only if it enters the perturbation region. Our strategy is therefore as follows: we begin integrating the characteristic strip at a point prior to its entry into the perturbation region. More precisely, the starting point is chosen so that the characteristic strip connecting it to the singular point never enters the perturbation region. Therefore, the surface and its slope are uniquely determined at the starting point. They are the same as for the exact solution if the singular point is a source, and the negative of the exact solution if it is a sink. This provides the initial conditions necessary to begin integrating at this point.

We find starting points as above such that the extended strip passes through the perturbation region. The influence of the perturbation is as expected: the characteristic trajectories move toward the floor of the valley, which is located along the line passing through (2.6, 2.6) in the direction (2, 3). Integrating several strips from different starting points, we verify that the perturbation does in fact cause these strips to cross. Again, since all these strips originated at the elliptical point, they are uniquely determined, and their crossing can be avoided only if the singular point is assumed to be a saddle.

Our results are illustrated in figures 28 and 29. In figure 28, characteristic strips integrated from a variety of starting points are shown for the exact, unperturbed solution. The integration accuracy is better than one part in 10^{11}. In figure 29, the strips integrated from the same starting points are shown. The perturbation region is indicated by the darker lines, which represent the long and short axes of the elliptical perturbation region. Before their entry into this region, the strips in figure 29 agree with those in figure 28. However, the effect of the perturbation is clearly to attract the trajectories toward the center of the perturbation region. As a result, the trajectories cross shortly after exiting this region. In this figure, $S = 5$, so that the maximum size of the perturbation is 1.8, much less than the unperturbed potential magnitude of 44. The dimensions of the perturbation are $r_{\text{par}} = 0.4$ and $r_{\text{perp}} = 0.1$. The smallest perturbation tried had $S = 0.5$, $r_{\text{par}} = 0.04$, and $r_{\text{perp}} = 0.01$. This also produced trajectory crossing, although on a finer scale. These experimental results support our conjecture than trajectory crossing can be achieved using an infinitesimal perturbation. Note also that a differently shaped perturbation, tailored so that the gradients are steepest precisely along the paths of some characteristic trajectories, could be more efficient in inducing trajectory crossing.

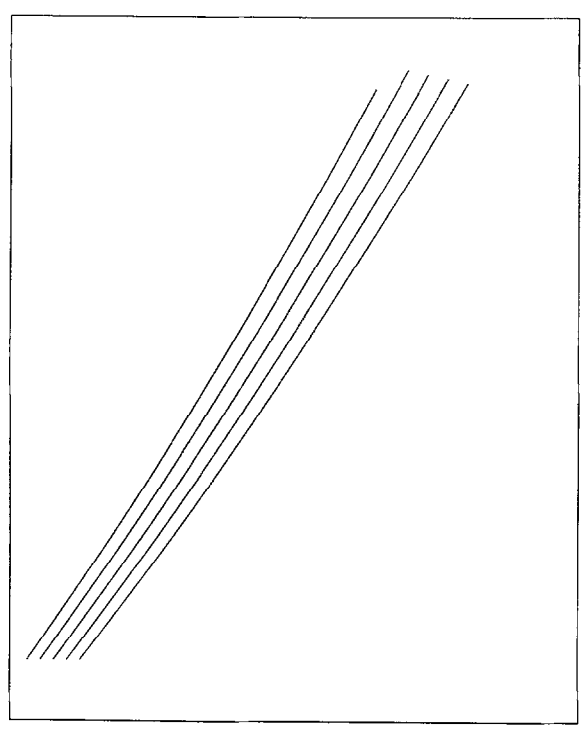


Fig. 28. Characteristic strips in the image plane derived by numerical integration, for the unperturbed image described in the text. The starting points are at the lower-left, at $y = 1.84$, and $x = 1.95, 2.0, 2.5, 2.1, 2.15$. The integration is halted when $y > 4$.

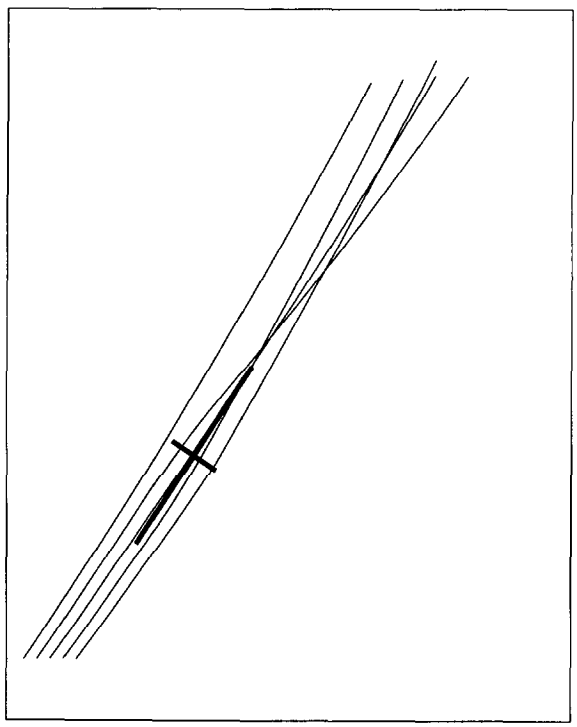


Fig. 29. Characteristic strips in the image plane derived by numerical integration, for the *perturbed* image described in the text. The enhanced crossed lines specify the position and extent of the intensity perturbation. Starting points for the strips and the scale of the display are the same as for the previous figure. Because of the crossing of the image strips, the singular point in the perturbed image cannot be interpreted consistently as elliptical.

Acknowledgements

I would like to acknowledge useful conversations with Don O'Shea, B.K.P. Horn, who corrected an error in an early version of this work, and Richard Weiss. I also thank R. Szeliski for sending me a copy of (Horn et al. 1990) prior to its publication. This work was supported by the Defense Advanced Research Projects Agency under grants F30602-87-C-0140 and DACA76-89-C-0017, and by the National Science Foundation under grants DCR-8500332 and IRI-9014698.

Note

1. That is, no twice-differentiable solution exists. A generalized surface solution—the *viscosity* solution—always exists, but is not necessarily differentiable everywhere (Elliot 1987).

References

Abraham, R., and Shaw, C. 1983. *Dynamics-The Geometry of Behavior, pt. 3: Global Behavior.* Aerial Press: Santa Cruz, CA.

Arnold, V.I. 1973. *Ordinary Differential Equations.* MIT Press: Cambridge, MA.

Bruss, A.R. 1982. The eikonal equation: some results applicable to computer vision. *J. Math. Phys.* 23(5): 890–896.

Elliot, R.J. 1987. *Viscosity Solutions and Optimal Control*, Longman Scientific and Technical: New York.

Giblin, P., and Weiss, R. 1987. Reconstruction of surfaces from profiles. *Proc. Intern. Conf. Comput. Vision*, London, June 1987, pp. 136–144.

Hirsch, M., and Smale, S. 1974. *Differential Equations, Dynamical Systems, and Linear Algebra.* Academic Press: NY.

Horn, B.K.P. 1975. Obtaining shape from shading information. In *The Psychology of Computer Vision.* P.H. Winston (ed.). McGraw Hill: New York, pp. 115–155.

Horn, B.K.P. 1990. Height and gradient from shading. *Intern. J. Comput. Vision* 5(1); 37–75.

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

Horn, B.K.P., Szeliski, R., and Yuille, A. 1990. Impossible shaded images. Submitted to *IEEE Trans. Patt. Anal. Mach. Intell.*

Oliensis, J. 1989. Existence and uniqueness in shape from shading. University of Massachusetts TR 89–109, October.

Oliensis, J. 1990. Shape from shading as a partially well-constrained problem. *Comput. Vision, Graphics, Image Process: Image Understanding*, to appear 1991.

Palis, J., and de Melo, W. 1982. *Geometric Theory of Dynamical Systems.* Springer-Verlag: New York.

Palis, J., and Smale, S. 1968. Structural stability theorems. *Proc. of Symp. Pure Math.* vol. 15, Berkeley, CA, July 1968, pp. 223–231.

Saxberg, B.V.H. 1989a. A modern differential geometric approach to shape from shading. MIT Artificial Intelligence Laboratory, TR-1117.

Saxberg, B.V.H. 1989b. An application of dynamical systems theory to shape from shading. *Proc. DARPA Image Understanding Workshop*, Palo Alto, CA, May 1989, pp. 1089–1104.