PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

A Computational Theory of Visual Surface Interpolation

W. E. L. Grimson

Phil. Trans. R. Soc. Lond. B 1982 **298**, 395-427
doi: 10.1098/rstb.1982.0088

Phil. Trans. R. Lond. B **298**, 395–427 (1982)
Printed in Great Britain

A COMPUTATIONAL THEORY OF VISUAL SURFACE INTERPOLATION

By W. E. L. GRIMSON

Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.

(*Communicated by S. Brenner, F.R.S. – Received* 28 *October* 1981)

CONTENTS

Computational theories of structure-from-motion and stereo vision only specify the computation of three-dimensional surface information at special points in the image. Yet the visual perception is clearly of complete surfaces. To account for this a computational theory of the interpolation of surfaces from visual information is presented.

The problem is constrained by the fact that the surface must agree with the information from stereo or motion correspondence, and *not* vary radically between these points. Using the image irradiance equation, an explicit form of this *surface consistency constraint* can be derived.

To determine which of two possible surfaces is more consistent with the surface consistency constraint, one must be able to compare the two surfaces. To do this, a functional from the space of possible functions to the real numbers is required. In this way, the surface most consistent with the visual information will be that which minimizes the functional. To ensure that the functional has a unique minimal surface, conditions on the form of the functional are derived. In particular, if the functional is a complete semi-norm that satisfies the parallelogram law, or the space of functions is a semi-Hilbert space and the functional is a semi-inner product, then there is a unique (to within possibly an element of the null space of the functional) surface that is most consistent with the visual information.

It can be shown, based on the above conditions plus a condition of rotational symmetry, that there is a vector space of possible functionals that measure surface consistency, this vector space being spanned by the functional of quadratic variation and the functional of square Laplacian. Arguments based on the null spaces of the respective functionals are used to justify the choice of the quadratic variation as the optimal functional.

Possible refinements to the theory, concerning the role of discontinuities in depth and the effects of applying the interpolation process to scenes containing more than one object, are discussed.

1. Introduction

Although our world has three spatial dimensions, the projection of light rays onto the retina presents our visual system with an image of the world that is inherently two-dimensional. We must use such images to physically interact with this three-dimensional world, even in situations new to us, or with objects unknown to us. That we do so easily implies that one of the functions of the human visual system is to reconstruct a three-dimensional representation of the world from its two-dimensional projection onto our eyes.

Methods that could be used to effect this three-dimensional reconstruction include stereo vision (Wheatstone 1838; Helmholtz 1925; Julesz 1971) and structure-from-motion (Miles 1931; Wallach & O'Connell 1953; Johannson 1964). Both of these methods may be considered as correspondence techniques, since they rely on establishing a correspondence between identical items in different images and on using the difference in projection of these items to determine surface shape. That is, correspondence methods compute surface information by:

(i) identifying a location in the physical scene in one image;

(ii) identifying the corresponding location in a second image, taken from a viewpoint different either in space (stereo) or in time (structure-from-motion); and

(iii) computing a three-dimensional surface value, representing the distance of the point relative to some base point, based on the difference in the positions of the two corresponding points in the images.

Many of the current computational theories of these processes (Marr & Poggio 1979; Grimson 1981*a*; Mayhew & Frisby 1981; Ullman 1979*a*; Longuet-Higgins & Prazdny 1980) argue that the correspondence process cannot take place at all points in an image. Rather, the first stage of

the correspondence process is to derive a symbolic description of points in the image at which the irradiance undergoes a significant change (Marr & Hildreth 1980). This symbolic representation (called the *primal sketch* (Marr 1976; Marr & Hildreth 1980)) forms the input to the second stage of the process in which the actual correspondence is computed. As a consequence of the form of the input, the correspondence process can compute explicit surface information only at scattered points in the image. Yet our perception is clearly of complete surfaces. (For example, in figure 1, a sparse random dot stereogram yields the vivid perception of a square floating in space above a background plane, rather than a collection of dots suspended in space.) The problem to be addressed in this paper is that of computing complete surface representations by interpolating an initial representation consisting of sparse surface values.

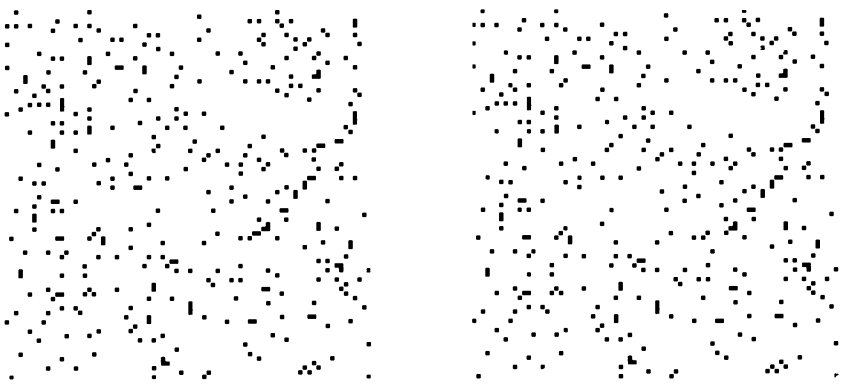


FIGURE 1. A sparse random dot pattern. Although the density of dots is very small, the perception obtained upon fusing this pattern is one of two disjoint planes, rather than dots isolated in depth.

We shall examine this surface interpolation problem at two levels. The first level is to consider the strictly mathematical question of surface reconstruction, independent of its relevance to the human visual system. Suppose that we are given a visual process that determines surface information at points corresponding to relevant changes in the images. In general, there will be many possible surfaces consistent with these initial surface points. For example, consider the boundary conditions provided by a circular arc, along which the depth is constant. The possible surfaces consistent with these known points include a flat disc, a sphere and even the highly convoluted surfaced formed by a radial sine function (see figure 3). How do we distinguish the correct one? Mathematically, we need to be able to compare two possible surfaces, to determine which is 'better'. This can be done by defining a functional Θ from the space of possible surfaces to the real numbers, so that comparing surfaces can be accomplished by comparing corresponding real numbers. Provided that $\Theta(f) < \Theta(g)$ whenever surface f is 'better' than surface g, the 'best' surface to fit through the known points is that which minimizes Θ. There are two problems to solve here: (i) What does it mean for f to be 'better' than g? and (ii) Under what conditions does a unique 'best' surface exist?

Once these questions have been answered and an appropriate functional has been derived, we can turn to the second level, which is to consider a specific algorithm for finding the surface that optimizes the functional. Because our intent is to consider models for the interpolation process as it occurs in the human visual system, we will restrict our attention to biologically feasible algorithms (Ullman 1979*b*; Grimson 1981*b*). In Grimson (1981*b*, *d*, 1982), such algorithms are derived and their performance on a range of images is illustrated.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The motivation for considering the interpolation problem first mathematically, independently of the specifics of the human system, and then algorithmically, incorporating specific biological contraints, is based on the assumption that one can consider the visual system as a symbol manipulation process (Marr 1976, 1982; Marr & Poggio 1977). This implies that the meaning of the symbols being manipulated can be distinguished from the physical embodiment of those symbols. Hence, one can deal with the mathematical consideration of the information processing that is occurring, independently of the implementation of that processing (whether in transistors or neurons). The rationale for this view lies in the belief that any computational theory should address the fundamental questions of the information processing necessary to perform the task, and that such computational theories are independent, to a large extent, of the method used to compute them. The initial goal is thus to determine computational constraints on the interpolation problem, based on the input and output representations of the process, and based on the structure of the computation required to transform one representation into the other. Note that a computational theory of the information processing is applicable both to the human visual system and to applications areas (such as high-altitude photomapping, hand–eye coordination systems, industrial robotics and inspection of manufactured parts) where it is useful to create a complete specification of surface shape.

While we shall initially concentrate on the mathematical aspects of visual surface interpolation, the problem is not completely isolated from the human visual system. If we view the human early visual system as a symbolic manipulator, we can consider visual processing as a series of transformations from one representation to another (Marr 1976, 1982). In particular, three stages can be identified (see figure 2). From the images, we transform to a description, called the primal sketch, of those locations at which the image irradiances change. Next, primal sketch descriptions of several images are matched, by either the stereo or motion computation, to obtain a description of surface information at the zero-crossings. This representation is called the raw $2\frac{1}{2}$-D sketch. Finally, the raw $2\frac{1}{2}$-D sketch is interpolated to obtain complete surface descriptions, called the full $2\frac{1}{2}$-D sketch (Marr 1978; Marr & Nishihara 1978). The first two stages have been considered elsewhere (Marr 1976; Marr & Hildreth 1980; Hildreth 1980; Marr & Poggio 1979; Grimson 1980, 1981*a*, *b*; Ullman 1979*a*). It is the final stage, the problem of surface interpolation, that is considered here.

We note that the form of the input and output representations can influence the design of the transformation. Here we shall assume that the input representation consists of explicit surface information, such as distance or relative distance, along the zero-crossings of the convolved image (these terms will be given technical definitions in § 2). The output representation will be a complete specification of surface information, where by complete we mean that an explicit distance value should be computed at every point on some grid representation of the scene. Our main concern in this paper is with the computational constraints needed to transform the input representation into the output representation.

Although surface values at all points of the image are important, there is another aspect of surface information that should also be made explicit in the output representation. This is the set of discontinuities in surfaces, the occluding contours, both subjective and objective. Marr (1978) argues that the $2\frac{1}{2}$-D sketch should be a viewer-centred representation that includes both explicit surface information, such as depth and surface orientation, and explicit contours of surface discontinuities. In this paper, the concentration is on the problem of creating explicit surface information at all points of the surface. In Grimson (1981*b*, *d*, 1982), the question of

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

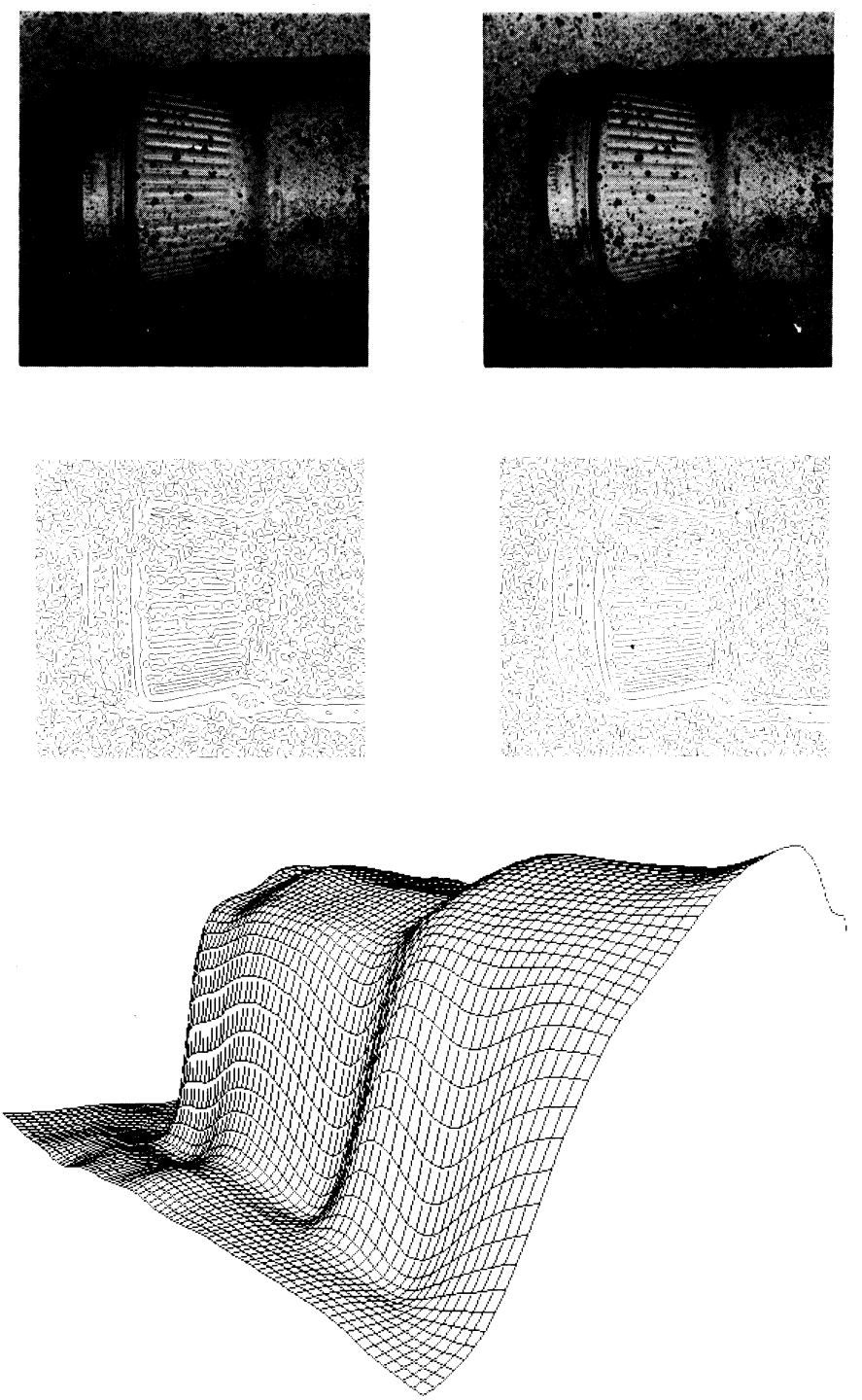


FIGURE 2. Example of processing. The top pair of images is a stereo pair of a scene. The middle pair illustrates the zero-crossings obtained from the images for one size of $\nabla^2 G$. The final image illustrates an interpolated surface description, formed by matching the zero-crossing descriptions, computing the distance to those points based on the difference in projection, and then interpolating the result.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

surface discontinuities is outlined, and possible algorithms are suggested, but an implementation of this stage has not been completed.

Throughout this paper, we shall assume that the surfaces are twice continuously differentiable. While the above remarks indicate that this is not completely valid, it does provide a reasonably good starting point for a computational theory of surface interpolation, since the contours of surface discontinuities will generally be very sparse. A more complete account of the interpolation process should, however, account for the process of making surface discontinuities explicit and for the effect of such discontinuities on our assumption of twice continuously differentiable surfaces.

2. Consequence of the correspondence problem

We indicated above that we should concentrate on correspondence methods that could effect the three-dimensional surface reconstruction: stereopsis (Marr 1982; Marr & Poggio 1979; Mayhew & Frisby 1981; Grimson 1980, 1981*a*) and structure-from-motion (Ullman 1979*a*; Longuet-Higgins & Prazdny 1980). The three main steps of the correspondence problem are: (i) identify a location in the physical scene in one image; (ii) identify the corresponding location in a second image; and (iii) compute a three-dimensional surface value, representing the distance of the point relative to some base point, based on the difference in the positions of the two corresponding points in the images.

If one can identify a location beyond doubt in the two images, then the correspondence problem is trivial. It can be demonstrated, however, that both the stereo computation and the motion computation can take place on very primitive descriptions of the images (Julesz 1960; Ullman 1979*a*). As a consequence, the difficulty of the problem, for human vision, lies in the correspondence problem, which item in one image matches which item in the other. The reason for this is that for any primitive element from one description, there are liable to be many possible matching elements from the other description. This is especially true if image irradiance values are used as the basic descriptions. Consider an image of a mat-painted wall with uniform illumination. Given a small element of that wall from one image, it is virtually impossible to distinguish which small element from the other view matches it. On the other hand, if there is a scratch or texture marking on the wall, it is likely that such a location can be distinguished in the two views. This suggests that the representation upon which the correspondence operation takes place should reflect those positions in an image at which some physical property of the underlying surface is changing. This representation is called the primal sketch (Marr 1976; Marr & Hildreth 1980).

Marr & Hildreth (1980; see also Hildreth 1980) have refined the preceding intuitive argument into more rigorous computational arguments, in conjunction with evidence from neurophysiology and psychophysics. They argue that the primal sketch representation is computed by determining those locations in an image at which the corresponding surface location undergoes a change in one of its physical properties, for example, reflectivity, texture or surface material. Such changes will generally correspond to a step change in image irradiance, at some scale. There are many ways of detecting the irradiance changes. Marr & Hildreth argue on computational and psychophysical grounds for using the following scheme. (See Richter & Ullman (1980) for neurophysiological arguments in support of this scheme.)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

(i) Convolve the image with a set of filters given by the Laplacian applied to a Gaussian,

$$\nabla^2 G(r,\theta) = \frac{r^2 - 2\sigma^2}{\sigma^4} \exp\left(-\frac{r^2}{2\sigma^2}\right),$$

where σ is a constant determined from psychophysical data.

(ii) Locate all non-trivial *zero-crossings* in the convolved irradiances (see Marr & Poggio 1977).

A non-trivial zero-crossing is a point at which the convolved irradiance values change from positive to negative or vice versa.

These zero-crossings form the basic representation upon which the later visual processing takes place.

Given this representation of the images, the stereo correspondence problem can now be solved (Marr & Poggio 1979; Grimson 1981*a*). After some additional processing, the structure-from-motion computation (Ullman 1979*a*) can also be performed. As a consequence, explicit three-dimensional surface information (such as distance, or surface orientation) can be computed only at points corresponding to zero-crossings in the primal sketch. This would yield a sparse surface representation. Yet clearly our perception is of complete surfaces (see for example figure 1). In addition, a 'nice' boundary is found for the central square. This implies that once the correspondence problem is solved, either by the stereo computation or by the motion computation, an interpolation must be performed between the surface values given at the zero-crossings, to obtain a complete surface description, and contours of surface discontinuities should be explicitly delineated.

3. The surface consistency constraint

We now turn to the problem of determining computational constraints involved in the process of creating complete surface specifications, by interpolating between known points. As basic input to the interpolation process, we are given the zero-crossings of a convolved image, with depth information computed along these zero-crossing contours. Suppose one were to attempt to construct a complete surface description based only on the surface information known along the zero-crossings. An infinite number of surfaces would consistently fit the boundary conditions provided by these surface values. Yet there must be some way of deciding which surface, or at least which small family of surfaces, could give rise to the zero-crossing descriptions. This means that there must be some additional information available from the visual process which, when taken into account, will identify a class of nearly indistinguishable surfaces that represent the visible surfaces of the scene.

To determine what information is available from the visual process, one must first carefully consider the process by which the zero-crossing contours are generated. The Marr–Hildreth theory of edge detection (Marr & Hildreth 1980; Hildreth 1980) relies on the fact that sudden changes in the reflectance of a surface, for example, caused by surface scratches or texture markings, will give rise to zero-crossings in the convolved image. Sudden or sharp changes in orientation or shape of the surface will under most circumstances also give rise to zero-crossings. This fact can be used to constrain the possible shapes of surfaces that could give rise to particular surface values along zero-crossing contours.

We illustrate the basic arguments with an example. Suppose that we are given a closed zero-

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

crossing contour, within which there are no other zero-crossings. An example would be a circular contour, along which the depth is constant. There are many surfaces that could fit this set of boundary conditions (see figure 3). One such surface is a flat disk. However, we could also fit other smooth surfaces to this set of boundary conditions. For example, the highly convoluted surface formed by $\sin (x^2+y^2)^{\frac{1}{2}}$ would be consistent with the known disparity values. Yet in principle, such a rapidly varying surface should give rise to other zero-crossings. This follows from the observation that if the surface orientation undergoes a periodic variation, then it is likely that the irradiance values will also undergo such a variation. Since the only evident zero-crossings are at the borders of the object, this implies that the surface $\sin (x^2+y^2)^{\frac{1}{2}}$ is not a valid representative surface for this set of boundary conditions.

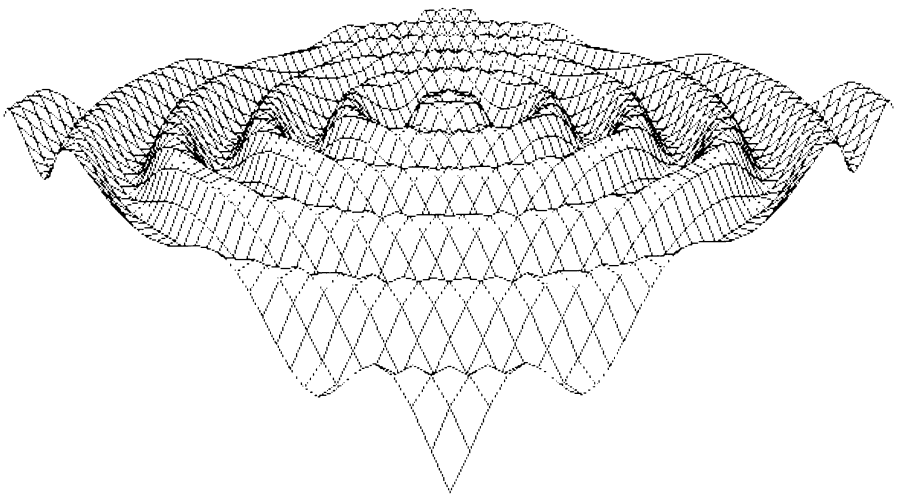


FIGURE 3. Possible surfaces fitting depth values at zero-crossings. Given boundary conditions of a circular zero-crossing contour, along which the depth is constant, there are many possible surfaces that could fit the known depth points. Two examples are a flat disk and the highly convoluted surface formed by $\sin (x^2+y^2)^{\frac{1}{2}}$, shown here. (From Grimson (1981*b*).)

Hence, the hypothesis is that the set of zero-crossing contours contains implicit information about the surface as well as explicit information. If we can determine a set of conditions on the surface shape that cause inflexions in the irradiance values, then we may be able to determine a likely surface shape, given a set of boundary conditions along the zero-crossing contours.

3.1. *No news is good news*

The implicit information about surface shape contained in the image irradiances can be formalized in the *surface consistency constraint*, namely:

The absence of zero-crossings constrains the possible surface shapes.

Just as the presence of a zero-crossing tells us that some physical property is changing at a given location, the absence of a zero-crossing tells us the opposite, that, in general, no physical property is changing, and in particular that the surface topography is not changing in a radical manner. We informally refer to this constraint as *no news is good news* since it says that, in general, the surface cannot change radically without informing us of this fact by means of zero-crossings.

To make explicit any constraints on the shape of the surface for locations in the image not associated with a zero-crossing, one must carefully examine the image formation process (Horn

1970, 1975, 1977). Many factors are involved in the formation of image irradiances. Changes in any of those factors can cause a change in the image irradiances, and hence a zero-crossing in the convolved image. For example: a change in surface material can correspond to a change in albedo, and hence to a zero-crossing in the convolved image; a discontinuity in depth can correspond to a change in the illumination striking the surface, and hence to a zero-crossing; and a discontinuity in surface orientation can correspond to a change in the amount of illumination reflected toward the viewer, and hence to a zero-crossing. We are interested in showing that the inverse is also true, in particular that, in regions in which the illumination and albedo are roughly constant, the absence of a zero-crossing implies that the surface shape cannot be changing in a radical manner.

While it is difficult (because of the many factors involved) to obtain a precise analytic expression for the probability of a zero-crossing occurring at any point in an image, we can obtain some reasonable estimates of this probability. In Grimson (1981*c*) it is shown that under some relatively minor assumptions concerning the surface material and the strength of the illuminant, an analytic expression for the probability of a zero-crossing in some region of an image can be derived. In essence, the expression verifies our earlier intuitive argument, that is, the probability of a zero-crossing is directly related to the amount of variation in the local surface orientation (or, more informally, to the fluctuation or 'wiggle' in the surface).

If we know that in some region of the image there are no zero-crossings, then the above relation can be inverted to imply that the surface should contain a minimal amount of variation in surface orientation (provided that we assume that the surface material and the illuminant are roughly constant over this region). This provides a constraint on the possible surfaces that could be interpolated through a set of known points, and is referred to as the *surface consistency constraint*.

4. The computational problem

We are now ready to consider the computational problem associated with the task of constructing complete surface specifications consistent with the information contained in the zero-crossings. The modules of early visual processing, such as stereo or structure-from-motion, provide explicit information about the shapes of the surfaces at specific locations in the images, corresponding to the zero-crossings of the convolved images. The surface consistency constraint indicates implicit information about the shapes of the surfaces between the zero-crossings, stating that between known depth values the surface cannot change in a radical manner, since such changes would usually give rise to additional zero-crossings. These two factors will now be combined, to obtain a complete surface specification.

4.1. *Using the surface consistency constraint*

Suppose that we are given a set of known depth points. We want a method for finding a surface to fit through these points that is 'most consistent' with the surface consistency constraint. We shall find the most consistent surface in two ways. In the *surface interpolation* problem we construct a surface that exactly fits the set of known points. The problem can be relaxed somewhat into a *surface approximation* problem, by only requiring that the surface should approximately fit the known data and be smooth in some sense.

Given the initial boundary conditions of the known depth values along the zero-crossing contours, there is an infinite set of possible surfaces that fit through those points. We need to be

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

able to compare pairs of members of this set of all possible surfaces fitting through those points, to determine which surface is more consistent. If we can do this, then the 'most consistent' surface can be found by comparing all possible surfaces. A traditional method for comparing surfaces is to assign a real number to each surface. Then, to compare the surfaces, we need only compare the corresponding real numbers. The assignment of real numbers to possible surfaces is accomplished by defining a functional, mapping the space of possible surfaces into the real numbers, $\Theta: X \mapsto \mathscr{R}$. This functional should be such that the more consistent the surface the smaller the real number assigned to it. To satisfy the surface consistency constraint, the functional should measure variation in surface orientation. In this case, the most consistent surface will be the surface that is minimal under the functional. (For further details and background information about the use of functionals, see, for example, Rudin (1973).)

The key mathematical difficulty is to guarantee the *existence* and *uniqueness* of a solution. In other words, we need to guarantee that there is at least one surface that minimizes the surface consistency constraint, and to guarantee that any other surface passing through the known points, for which the functional measure of surface consistency has the same value, is indistinguishable from the first surface. This issue is not just a mathematical nicety, however, but is essential to the solution of many computational problems. Suppose that we devise an iterative algorithm to solve some problem. What happens if we cannot guarantee the existence of a solution? The iterative process could be set off to solve an equation and never converge to an answer, which is clearly undesirable. Further, suppose that a solution exists but is not guaranteed to be unique. Then an iterative process might converge to one solution when started from one point, and converge to another solution when started from a different point. Although small variations in the different solutions might be acceptable, the solutions should not differ in a manner inconsistent with our intuition about the problem. Thus, in visual surface interpolation, the real trick is to find a functional that accurately measures the variation in surface orientation, as well as guarantees the existence of a unique best surface (or a family of indistinguishable surfaces).

How can we guarantee the existence and uniqueness of a solution? In our particular case of surface interpolation, we shall be using the calculus of variations to determine a system of equations that the most consistent surface must satisfy, by applying the calculus to the situation of fitting a thin plate through a set of known points. While this system of equations characterizes the minimal surface, it does not guarantee uniqueness. The form of the boundary conditions (the set of known points) will determine the size of the family of minimal surfaces. Unfortunately, determining the types of input for which a unique solution exists is generally very hard. Instead, we shall exploit a general case of the mathematical existence of a solution with the weakest possible conditions on the functional. That is, we shall determine a weak set of conditions on the functional that are needed to ensure that a unique most consistent surface, or at least a unique family of surfaces that are most consistent, will exist. We shall show that, if the functional is an inner product on a Hilbert space of possible surfaces, then a unique most consistent surface will exist. (A Hilbert space is an extension of normal Euclidean space, basically an infinite dimensional vector space in which a dot product operation exists and in which functions are usually used in place of the normal notion of vector.)

In general, it is extremely difficult to find a functional that measures surface consistency and satisfies the conditions of an inner product. Hence, we shall show that if the functional is a semi-inner product on a semi-Hilbert space of possible surfaces, then the most consistent surface is unique up to possibly an element of the null space of the functional. (The null space is simply the set of surfaces that cannot be distinguished by the functional from the surface that is zero every-

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

where.) In this way, the family of most consistent surfaces can be found. Based on the form of the null space, we can determine whether or not the differences in minimal surfaces are intuitively indistinguishable, and what conditions on the known boundary values will guarantee a unique minimal surface, from this family.

Having derived conditions on the functional, we need to show that there is such a functional. The surface consistency constraint implies that the functional should measure variation in surface orientation over an area of the surface. Although the condition of a semi-inner product is a mathematical requirement needed to guarantee a solution, it does not restrict in an unreasonable way the kinds of surfaces that we consider, and gives rise to at least two very natural functionals, both of which can be derived from the calculus of variations: one measures the integral of the square Laplacian applied to the surface and the other measures the quadratic variation of the local x and y components of the surface orientation.

Given that there are at least two possible functionals, are there others? It can be shown (Brady & Horn 1982) that, if we require a functional that is (i) a monotonic function of the variation in surface orientation, (ii) a semi-inner product, and (iii) rotationally symmetric, then there is a vector space of possible functionals, spanned by the square Laplacian and the quadratic variation. In other words, there is a family of possible functionals, given by all linear combinations of these two basic functionals.

Given that there is more than one possible functional, how do they differ? Using the calculus of variations, and some results from mathematical physics, we shall show that the surfaces that minimize these functionals will be roughly identical in the interior of a region and will differ only along the boundaries of a region. Also, the null spaces of the functionals will differ, implying different families of most consistent surfaces corresponding to each functional. We know that the minimal surface is unique up to possibly an element of the null space. Since we require that the solution surface be either unique, or a member of an indistinguishable family of solutions, the size of the null space is important in judging the value of a functional. Based on this, we shall argue that the quadratic variation is to be preferred over the square Laplacian. If we require that the surface pass through the known points, we can show that the form of the stereo data will force a unique most consistent surface for quadratic variation, while this is unlikely for functionals such as the square Laplacian.

In Grimson (1981 *b*, *d*, 1982) examples of the types of minimal surfaces obtained under quadratic variation and the square Laplacian are shown, and it is argued that the mathematical distinction in size of null space has a practical consequence, as the types of surfaces that minimize the square Laplacian are inconsistent with our intuitive notion of the best surface to fit to the known points, while the surfaces that minimize the quadratic variation are much more consistent with our intuitive notion of the best surface.

4.1.1. *Possible functionals*

We know from our intuitive arguments concerning the radial sine function of figure 3 that any appropriate functional should measure the amount of 'wiggle' in the surface, that is, the amount of variation in the local surface orientation. This suggests that the functional should measure some factor of the second-order derivatives of the surface. In this section a number of possible functionals are outlined.

Example 1. One possibility is to measure the curvature of the surface, which implicitly reflects variation in surface orientation. The curvature of a surface is usually measured in one of two ways.

For any point on the surface, consider the intersection of the surface with a plane containing

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

the normal to the surface at that point. This intersection defines a curve, the curvature of which can be measured as the arc rate of rotation of its tangent. For any point there are infinitely many normal sections, each defining a curve. As the normal section is rotated through 2π radians, all possible normal sections will be observed. There are two sections of particular interest, that which has the maximum curvature and that which has the minimum. It can be shown that the directions of the normal sections corresponding to these sections are orthogonal. These directions are the *principal directions* and the curvatures of the normal sections in these directions are the *principal curvatures*, denoted κ_a and κ_b. It can be shown that the curvature of any other normal section is defined by the principal curvatures.

There are two standard methods for describing the curvature of the surface, in terms of the principal curvatures. One is the first (or mean) curvature of the surface

$$J = \kappa_a + \kappa_b.$$

The other is the second or Gaussian curvature of the surface

$$K = \kappa_a \kappa_b.$$

For a surface defined by the vector $[x, y, f(x, y)]$, these curvatures are given by

$$J = \frac{\partial}{\partial x}\left[\frac{f_x}{(1+f_x^2+f_y^2)^{\frac{1}{2}}}\right] + \frac{\partial}{\partial y}\left[\frac{f_y}{(1+f_x^2+f_y^2)^{\frac{1}{2}}}\right]$$

and

$$K = \frac{f_{xx}f_{yy} - f_{xy}^2}{(1+f_x^2+f_y^2)^2}.$$

Thus, there are two possibilities for the functional. One is to measure the first (or mean) curvature of the surface,

$$\Theta_1(f) = \left[\iint J^2 \,dx\,dy\right]^{\frac{1}{2}}$$
$$= \left[\iint \frac{[f_{xx}(1+f_y^2) + f_{yy}(1+f_x^2) - 2f_x f_y f_{xy}]^2}{(1+f_x^2+f_y^2)^3}\,dx\,dy\right]^{\frac{1}{2}}.$$

If f_x and f_y are assumed to be small, then Θ_1 is closely approximated by the functional

$$\Theta_2(f) = \left[\iint (\nabla^2 f)^2 \,dx\,dy\right]^{\frac{1}{2}}.$$

Example 2. A second possibility for reducing curvature is to reduce the second or Gaussian curvature,

$$\Theta_3(f) = \left[\iint K^2 \,dx\,dy\right]^{\frac{1}{2}}.$$

Note that by using the above approximation of small f_x and f_y we obtain the functional

$$\Theta_4(f) = \left[\iint (f_{xx}f_{yy} - f_{xy}^2)\,dx\,dy\right]^{\frac{1}{2}}.$$

We shall return to this form later.

Example 3. Another possibility is to consider second-order variation (called quadratic variation here) in each of the surface variables. The quadratic variation in $p = f_x$ is given by

$$\iint (p_x^2 + p_y^2)\,dx\,dy$$

and the quadratic variation in $q = f_y$ is given by

$$\iint (q_x^2 + q_y^2)\, \mathrm{d}x\, \mathrm{d}y.$$

If the surface is twice continuously differentiable, then $p_y = q_x$, and by combining these two variations we obtain the quadratic variation

$$\Theta_5(f) = \left[\iint (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)\, \mathrm{d}x\, \mathrm{d}y\right]^{\frac{1}{2}}.$$

Duchon (1975, 1976) refers to the surfaces that minimize this expression as *thin plate splines* since the expression Θ_7 relates to the energy in a thin plate forced to interpolate the data. We shall return to this point later.

It turns out that for forms such as Θ_1, Θ_3, it is very difficult to show that under general input conditions (that is, a random assortment of zero-crossings with associated depth values) there exists a unique minimal surface. For forms such as Θ_2, Θ_4, Θ_5, however, we can exploit a general class of functionals for which the uniqueness of a minimal solution is defined.

4.1.2. *The problem is well defined*

If surfaces are to be compared, by using a functional from the space of surfaces to the real numbers, with the purpose of finding the surface that best satisfies the surface consistency constraint, it is necessary to ensure that such a goal is attainable. What conditions on the form of the functional, or on the structure of the space of functions, will guarantee the existence of such a 'best' surface? One key constraint on the functional is given by the following theorem. The main point of the theorem is that one method (although not the only one) of ensuring that the problem is well defined is to require the functional to have the characteristics of a semi-norm.

THEOREM 1. *Suppose that there exists a complete semi-norm Θ on a space of functions H, and that Θ satisfies the parallelogram law (for definition, see proof of theorem). Then, every non-empty closed convex set $E \subset H$ contains a unique element v of minimal norm, up to an element of the null space. Thus, the family of minimal functions is*

$$\{v + s \mid s \in S\},$$

where

$$S = \{v - w \mid w \in E\} \cap \mathcal{N}$$

and $\mathcal{N}$ is the null space of the functional

$$\mathcal{N} = \{u \mid \Theta(u) = 0\}.$$

Proof. (See, for example, Rudin 1973.) Any space with a semi-norm defined on it can be associated with an equivalent normed space. Let W be a subspace of a vector space H. For every $v \in H$, let $\pi(v)$ be the coset of W that contains v,

$$\pi(v) = \{v + u : u \in W\}.$$

These cosets are elements of a vector space H/W called the quotient space of H modulo W. In this space, addition is defined by

$$\pi(v) + \pi(w) = \pi(v + w)$$

and scalar multiplication is defined by

$$\alpha\pi(v) = \pi(\alpha v).$$

The origin of the space H/W is $\pi(0) = W$. Thus, π is a linear map of H onto H/W with W as its null space.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Now consider the semi-norm Θ on the vector space H. Let

$$\mathcal{N} = \{v: \Theta(v) = 0\}.$$

This can easily be shown to be a subspace of H. Let π be the quotient map from H onto $H/\mathcal{N}$, and define a mapping $\Theta': H/\mathcal{N} \mapsto \mathcal{R}$,

$$\Theta'[\pi(v)] = \Theta(v).$$

If $\pi(v) = \pi(w)$ then $\Theta(v-w) = 0$. Since $|\Theta(v) - \Theta(w)| \leqslant \Theta(v-w)$, then $\Theta'[\pi(v)] = \Theta'[\pi(w)]$ and Θ' is well defined on $H/\mathcal{N}$. It is straightforward to show that Θ' is a norm on $H/\mathcal{N}$.

Now we can prove the statement of the theorem. The set E, a subset of H, can be transformed into a set E' in the quotient space $H/\mathcal{N}$ while preserving the convexity and closure properties.

The parallelogram law states

$$[\Theta'(v+w)]^2 + [\Theta'(v-w)]^2 = 2[\Theta'(v)]^2 + 2[\Theta'(w)]^2$$

Let

$$d = \inf\{\Theta'(v): v \in E'\}.$$

Choose a sequence $v_n \in E'$ such that $\Theta'(v_n) \mapsto d$. By the convexity of E', we know that

$$\tfrac{1}{2}(v_n + v_m) \in E' \quad \text{and so} \quad [\Theta'(v_n + v_m)]^2 \geqslant 4d^2.$$

If v and w are replaced in the definition of the parallelogram law by v_n and v_m, then the right side tends to $4d^2$. But $[\Theta'(v_n + v_m)]^2 \geqslant 4d^2$; so one must have $[\Theta'(v_n - v_m)]^2 \mapsto 0$ to preserve the equality. Thus, $\{v_n\}$ is a Cauchy sequence in $H/\mathcal{N}$. Since the norm is complete, the sequence must converge to some $v \in E'$, with $\Theta'(v) = d$.

To prove the uniqueness, if $v, w \in E'$ and $\Theta'(v) = d, \Theta'(w) = d$ then the sequence $\{v, w, v, w, \ldots\}$ must converge, as we just saw. In other words, $[\Theta'(v-w)]^2 \mapsto 0$. Since Θ' is a norm, this implies that $v - w = 0$ or $v = w$ and hence the element is unique.

We have proven that, under the norm Θ' on the quotient space $H/\mathcal{N}$, the set E' has a unique minimal element. Hence, the structure of the quotient space implies that, under the semi-norm Θ on the space H, the set E has a unique minimal element v, up to possibly an element of the null space $\mathcal{N}$. In other words, the family of minimal elements is

$$\{v + s |\ s \in S\}$$

where

$$S = \{v - w |\ w \in E\} \cap \mathcal{N}.$$

This theorem specifies one set of mathematical criteria needed to ensure that there exists a unique minimal element. Thus, if the surface consistency constraint could be specified by a functional that satisfied the conditions of a complete semi-norm, obeying the parallelogram law, it might be possible to show that there is a unique coset of 'most consistent' surfaces. We should really prefer to be guaranteed a unique surface, rather than some set of surfaces. One way to tighten the result of the theorem is to require that the functional is a norm.

COROLLARY 1.1. *If Θ is a complete norm on a space of functions H, which satisfies the parallelogram law, then every non-empty closed convex set $E \subset H$ contains a unique element v of minimal norm.*

Proof. If the functional is a norm, the null space is the trivial null space, and the result holds uniquely.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The theorem can be rephrased in terms of the surface interpolation problem as follows.

COROLLARY 1.2. *Let the set of known points be given by*

$$\{(x_i, y_i) \mid i = 1, \ldots, N\}$$

where the associated depth value is F_i. *Let* $\mathcal{F}$ *be a vector space of 'possible' functions on* $\mathcal{R}^2$ *and let*

$$U = \{f \in \mathcal{F} \mid f(x_i, y_i) = F_i \quad i = 1, \ldots, N\}$$

so that U *is the set of functions that interpolate the known data* $\{F_i\}$. *Let* Θ *be a semi-norm, which measures the 'consistency' of a function* $f \in X$, *that is, we shall say that* f *is 'better' than* g *if* $\Theta(f) < \Theta(g)$. *If* Θ *is a complete semi-norm and satisfies the parallelogram law, then there exists a unique (to within possibly a function of the null space of* Θ*) functions* $s \in U$ *that is least inconsistent and interpolates the data. Hence the interpolation problem is well defined.*

Proof. Clearly U is a convex set since for any $f, g \in U$,

$$[\lambda f + (1 - \lambda) g](x_i, y_i) = (\lambda + 1 - \lambda) F_i = F_i,$$

for any data point (x_i, y_i). Furthermore, U is closed, since if $f_n \in U$ and $f_n \mapsto f$ then $f(x_i, y_i) = F_i$ and $f \in U$. Then the previous corollary states that U has a unique (to within an element of the null space) element of minimal norm, which is exactly the desired 'most consistent' surface.

This corollary is a translation of theorem 1 into the problem of interest to us, finding the surface most consistent with the known data from the stereo algorithm. It specifies a set of conditions under which the interpolation problem is well defined. Here, the notion of well defined refers to finding a solution to the interpolation problem that is unique, and by unique we mean up to possibly an element of the null space of the semi-norm. As a consequence, the extent and structure of the null space of any semi-norm chosen to incorporate the surface consistency constraint will be important in determining the utility of that semi-norm. There are two reasons for this. One is that the null space defines how much 'wiggle' or fluctuation in the surface is invisible to the functional. The second is that when combined with suitable conditions on the set of known points, the structure of the null space can be used to determine how unique a minimal solution is (that is, whether there is exactly one minimal solution, or a family of solutions, and how much the members of that family differ).

Thus, theorem 1 and corollary 1.1 specify two different sets of sufficient, but not necessary, criteria for ensuring differing types of uniqueness. In both cases, the criteria apply directly to the structure of the functional. Of course, the real trick is to find a functional Θ that captures our intuition of variation in surface orientation and meets the requirements needed to guarantee a unique solution.

4.1.3. *The space of functions*

Theorem 1 describes a set of sufficient conditions for obtaining a unique family of minimal surfaces. The fundamental point is that we require a complete parallelogram semi-norm to ensure a unique solution. These conditions precisely define a semi-inner product, and hence the space of functions over which we seek a minimum must be a semi-Hilbert space.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

COROLLARY 1.3. *If $\mathcal{F}$ is a semi-Hilbert space of possible surfaces, and $\Theta(v) = \mu(v,v)^{\frac{1}{2}}$ is an inner product semi-norm, where $\mu(v,v)^{\frac{1}{2}}$ is the semi-inner product of the space $\mathcal{F}$, then there exists a unique surface in $\mathcal{F}$ (possibly to within an element of the null space of the semi-norm) that minimizes the semi-norm Θ over all surfaces.*

Proof. By the definition of Hilbert space, the semi-norm is complete. It is easy to show that it satisfies the parellelogram law from the definition of $\Theta(v) = \mu(v,v)^{\frac{1}{2}}$. Thus, if the space of functions is a semi-Hilbert space, then, by theorem 1, the interpolation problem is guaranteed to have a unique minimal solution, possibly to within an element of the null space.

COROLLARY 1.4. *If $\mathcal{F}$ is a Hilbert space of possible surfaces, and $\Theta(v) = \mu(v,v)^{\frac{1}{2}}$ is an inner product norm, where $\mu(v,v)^{\frac{1}{2}}$ is the inner product on the space $\mathcal{F}$, then there exists a unique surface in $\mathcal{F}$ that minimizes the norm Θ over all surfaces.*

4.1.4. *Return to the examples*

Returning briefly to the set of possible functionals derived in §4.1.1, we see that, while Θ_1 and Θ_3 are not semi-norms, Θ_2, Θ_4 and Θ_5 do satisfy the conditions of theorem 1. Hence we are guaranteed a unique solution up to possibly an element of the null space.

For the square Laplacian, Θ_2, the null space is the space of all harmonic functions. In light of our surface consistency constraint, this may not be the most appropriate functional. For the quadratic variation, Θ_5, the null space is the space of all linear functions. Here, the collection of surfaces that are invisible to the functional seems to make sense, since we should expect planes to have no measurable surface variation, and any other surface to contain at least some variation. Note that the quadratic variation is capable of distinguishing between possible surfaces to a much finer level than the square Laplacian, since only surfaces that differ by a plane will appear identical to the quadratic variation, while surfaces differing by a harmonic function (which includes many more functions than the planes) will be indistinguishable to the square Laplacian.

4.2. *Where do we stand?*

We have seen that for the general surface interpolation problem there are two constraints on possible functionals. One is that the functional must measure a monotonic function of the variation in surface orientation. The other is that the functional should satisfy the conditions of a complete parallelogram semi-norm, or, equivalently, a semi-inner product. If the functional satisfies these conditions, then we know that there will be a unique family of surfaces that minimize this functional and hence form a family of best possible surfaces to fit through the known information. In the examples sketched above we saw that there are at least two possible candidates for this functional, namely the square Laplacian,

$$\Theta_2(f) = \left[\iint (\nabla^2 f)^2 \,\mathrm{d}x\,\mathrm{d}y\right]^{\frac{1}{2}}$$

and the quadratic variation,

$$\Theta_5(f) = \left[\iint (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)\,\mathrm{d}x\,\mathrm{d}y\right]^{\frac{1}{2}}.$$

There are several points still to consider. Are there other possible functionals? How do the minimal solutions to these functionals differ? What criteria can be applied to determine which functional is best suited to our surface interpolation problem? What is the best functional under

those criteria? In §§4.3–4.5, we shall consider these questions in detail. The point that we shall develop is that the appropriate functional to apply is the quadratic variation, and thus the surface that minimizes this functional is 'most consistent' with the imaging information.

4.3. *Are there other functionals?*

We have determined at least two functionals that meet our conditions. Are there other possible functionals, and, if so, how do their minimal solutions differ from those of the square Laplacian and the quadratic variation?

To answer this question, we rely on a result of Brady & Horn (1982), sketched below. Recall that the basic conditions on the functional were that it measure a monotonic function of the variation in surface orientation and that it be a semi-inner product. The first requirement suggests that the functional must involve terms that are functions of the second-order partial derivatives of the surface, since such terms will be related to the variation in surface orientation. The second requirement is needed to ensure the uniqueness of the solution. The conditions for $\mu(f,g)$ to be a semi-inner product are:

(i) $\mu(f,g) = \mu(g,f)$;
(ii) $\mu(f+g,h) = \mu(f,h)+\mu(g,h)$;
(iii) $\mu(\alpha f,g) = \alpha\mu(f,g)$;
(iv) $\mu(f,f) \geqslant 0$.

Given a semi-inner product $\mu(f,g)$, we can define the desired functional by $\Theta(f) = \mu(f,f)^{\frac{1}{2}}$.

The difficult condition to satisfy is (iii), which implies that the semi-inner product should not contain any constant terms. The conditions taken together imply that we should consider any quadratic form as a possible semi-inner product:

$$\mu(f,g) = \iint \alpha f_{xx}g_{xx} + \beta f_{xy}g_{xy} + \gamma f_{yy}g_{yy} + \delta(f_{xx}g_{xy} + f_{xy}g_{xx}) + \epsilon(f_{xx}g_{yy} + f_{yy}g_{xx}) + \zeta(f_{xy}g_{yy} + f_{yy}g_{xy}).$$

Thus, the corresponding functional will have the quadratic form:

$$\Theta(f) = \iint \alpha f_{xx}^2 + \beta f_{xy}^2 + \gamma f_{yy}^2 + 2\delta f_{xx}f_{xy} + 2\epsilon f_{xx}f_{yy} + 2\zeta f_{xy}f_{yy}.$$

The final condition that we apply to the functional is that it be rotationally symmetric. This follows from the observation that if the input is rotated the surface that fits the known data should not change in form, other than also being rotated.

Minimizing the quadratic form of the functional $\Theta(f)$ can be considered as finding the minimum over the integral of the function $(\Delta f)^T M \Delta f$ where Δf is the vector

$$\begin{pmatrix} f_{xx} \\ f_{xy} \\ f_{yy} \end{pmatrix}$$

and M is the symmetric matrix

$$\begin{bmatrix} \alpha & \delta & \epsilon \\ \delta & \beta & \zeta \\ \epsilon & \zeta & \gamma \end{bmatrix}.$$

If R is a rotation matrix, then the condition of rotational symmetry is given by

$$(R\Delta f)^T M (R\Delta f) = \Delta f^T M \Delta f.$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Vector algebra implies that we must have

$$R^T MR = M \quad \text{or} \quad R^T M = MR^{-1}.$$

Equating elements shows that the matrix M must have the form

$$\begin{bmatrix} \frac{1}{2}\beta+\epsilon & 0 & \epsilon \\ 0 & \beta & 0 \\ \epsilon & 0 & \frac{1}{2}\beta+\epsilon \end{bmatrix}.$$

There are two important consequences of this fact. The first is that the set of all possible functionals forms a vector space, since if M_1 and M_2 satisfy the conditions then so does $\sigma M_1 + \nu M_2$. The second is that this vector space of operators is spanned by the square Laplacian and the quadratic variation since

$$\begin{bmatrix} \frac{1}{2}\beta+\epsilon & 0 & \epsilon \\ 0 & \beta & 0 \\ \epsilon & 0 & \frac{1}{2}\beta+\epsilon \end{bmatrix} = \epsilon \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} + \tfrac{1}{2}\beta \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The first term of the sum corresponds to the square Laplacian while the second corresponds to the quadratic variation. Thus, for $\epsilon = 1$ and $\beta = 0$, the functional reduces to square Laplacian. For $\epsilon = 0$ and $\beta = 2$, the functional reduces to quadratic variation. Finally, if $\epsilon = \frac{1}{2}$ and $\beta = -1$ we obtain a functional that corresponds to the approximation to the integral of square Gaussian curvature derived previously.

Thus, we have answered our second question. There are other possible functionals, but they are all linear combinations of the two basic functionals, the square Laplacian and the quadratic variation.

4.4. *How do the functionals differ?*

Given that there are many possible functionals, all linear combinations of the square Laplacian, Θ_2, and the quadratic variation, Θ_5, we must consider how the solutions to the square Laplacian and the quadratic variation differ. In other words, is there any noticeable difference in the surfaces that minimize these two functionals, subject to fitting through the stereo data? To answer this question, we shall rely on the calculus of variations (see, for example: Courant & Hilbert 1953; Forsyth 1960). The salient points are outlined below.

4.4.1. *Calculus of variations*

The calculus of variations is frequently used to solve problems of mathematical physics, and is applicable to our surface interpolation problem. In particular, we can use the calculus of variations to formulate differential equations associated with problems of minimum energy. Suppose that we are given a thin elastic plate whose equilibrium position is a plane and whose potential energy under deformation is given by an integral of the quadratic form in the principal curvatures of the plate. We can consider the interpolation problem as one of determining the surface formed by fitting a thin elastic plate over a region $\mathscr{R}$ (with boundary Γ) and through the known points. With use of a small deflexion approximation, the potential energy is given by

$$\iint_{\mathscr{R}} \left[(\nabla^2 f)^2 - 2(1-\mu)\,(f_{xx}f_{yy} - f_{xy}^2) \right] \mathrm{d}x\,\mathrm{d}y.$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The solution to the interpolation problem is then the surface that has the minimum potential energy.

The calculus of variations can be used to characterize this problem by providing a set of differential equations (called the Euler equations) that the solution surface must satisfy. It can be shown (see Courant & Hilbert 1953, p. 251) that the Euler equations for the interior of any region $\mathscr{R}$ are given by

$$\nabla^4 f = f_{xxxx} + 2f_{xxyy} + f_{yyyy} = d(x, y)$$

where $d(x, y)$ represents the density of the known surface points. Along the boundary contour Γ of the region, the solution surface must satisfy the equations (called the *natural boundary conditions*)

$$M(f) = -\nabla^2 f + (1-\mu)\,(f_{xx}x_s^2 + 2f_{xy}x_s y_s + f_{yy}y_s^2) = 0$$

$$P(f) = \frac{\partial}{\partial n}\nabla^2 f + (1-\mu)\frac{\partial}{\partial s}(f_{xx}x_n x_s + f_{xy}(x_n y_s + x_s y_n) + f_{yy}y_n y_s) = 0,$$

where $\partial/\partial n$ is a derivative normal to the boundary contour, $\partial/\partial s$ is a derivative with respect to arc length along the boundary contour and x_s, y_s and x_n, y_n are the direction cosines of the tangent vector and the outward normal respectively. The constant μ denotes a constant factor associated with the elastic material of the plate, called the Poisson ratio.

There are two subcases of particular interest. In the first case, suppose that $\mu = 1$. The energy equation reduces to

$$\iint_{\mathscr{R}} (\nabla^2 f)^2 \,\mathrm{d}x\,\mathrm{d}y,$$

which is simply the square Laplacian condition derived previously. The Euler equation is the biharmonic equation $\nabla^4 f = 0$ while the natural boundary conditions are

$$\nabla^2 f = 0, \quad \partial\nabla^2 f/\partial n = 0,$$

along the boundary contour Γ. In the second case, suppose that the constant factor is given by $\mu = 0$. The energy equation reduces to

$$\iint_{\mathscr{R}} (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)\,\mathrm{d}x\,\mathrm{d}y,$$

which is simply the quadratic variation condition, also derived previously. The Euler equation is identical to that of the square Laplacian, namely the biharmonic equation $\nabla^4 f = 0$. The natural boundary conditions are different, however. They are given by

$$-\nabla^2 f + (f_{xx}x_s^2 + 2f_{xy}x_s y_s + f_{yy}y_s^2) = 0,$$

$$\frac{\partial}{\partial n}\nabla^2 f + \frac{\partial}{\partial s}\left(f_{xx}x_n x_s + f_{xy}(x_n y_s + x_s y_n) + f_{yy}y_n y_s\right) = 0.$$

In the simple case of a square boundary, oriented with respect to the coordinate axes, the boundary conditions reduce to

$$f_{yy} = 0, \quad 2f_{xxy} + f_{yyy} = 0$$

along the boundary segments parallel to the x axis and

$$f_{xx} = 0, \quad 2f_{yyx} + f_{xxx} = 0$$

along the boundary segments parallel to the y axis.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

These boundary conditions can be straightforwardly simplified to

$$f_{yy} = 0, \quad f_{xxy} = 0$$

along the boundary sdgments parallel to the x axis and

$$f_{xx} = 0, \quad f_{yyx} = 0$$

along the boundary segments parallel to the y axis.

Thus, for both the square Laplacian and the quadratic variation, the Euler equations are identical in the interior. The natural boundary conditions, however, are different. This suggests that the solutions to the functionals will in general be different, especially along the boundaries of the surfaces, although the difference can, of course, propogate into the interior of the surface. In Grimson (1981*b*, 1981*d*, 1982), examples of solving these equations are shown and this difference is seen to be important.

There is a second manner in which the minimal solutions to the functionals will differ, in part related to the difference in boundary conditions of the two solutions. The minimal surface obtained from either functional will be uniquely determined only to within possibly an element of the null space of the functional. This will be an important factor in determining which functional is best suited to our problem, since we should like the boundary conditions provided by the stereo data to completely determine a unique solution. The null spaces of the two functionals differ greatly, since the null space of the quadratic variation is the space of all linear functions, while the null space of the square Laplacian is the much larger space of all harmonic functions. We shall consider the effect of this difference later.

4.5. *The best functional*

Given that the set of possible functionals forms a vector space spanned by the square Laplacian and the quadratic variation, what criteria can be applied to determine the 'best' functional?

We suggested earlier that there are at least two criteria that may be used to determine the 'best' functional. One is the size of the null space, since this determines the resolution of the functional, that is, the level at which the functional cannot distinguish between two different surfaces. We saw that the quadratic variation was unable to distinguish between two surfaces only when they differed by a plane, while the square Laplacian could not distinguish between two surfaces differing by any harmonic function. Thus the size of the null space is important.

Let us denote the null space of the square Laplacian by $\mathcal{N}_1$ (the space of all harmonic functions) and the null space of the quadratic variation by $\mathcal{N}_2$ (the space of all linear functions). Note that $\mathcal{N}_2$ is a subspace of $\mathcal{N}_1$. Now the null space for any linear combination of these two operators must contain at least the space spanned by the intersection of the two null spaces $\mathcal{N}_1$ and $\mathcal{N}_2$. Hence the null space of any other operator must consist at least of the linear functions. Thus no possible operator can have a null space smaller than that corresponding to quadratic variation, and this suggests that the quadratic variation may be the best function to use.

The importance of the null space is that it helps determine the family of surfaces that are minimal under the functional. The requirement that we impose on the best functional is that the member of this family corresponding to the minimal surface be uniquely determined, when combined with the requirement that the surface must pass through the known points provided by the stereo algorithm. Clearly the smaller size of the null space the fewer the requirements that we must impose on the output of the stereo algorithm to ensure a unique solution.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

We may view this criterion in the following manner. We start with the space of all possible functions, namely, the space of all second differentiable functions of two real variables, denoted $C^2(\mathscr{R}^2)$. If we restrict our attention to those surfaces that pass through the boundary conditions imposed by the stereo or structure-from-motion data, we define a convex subset $U \subset C^2(\mathscr{R}^2)$. If we define a functional on this space, the set of surfaces that are minimal under the functional is given by

$$\{v+s \mid s \in S\}$$

where

$$S = \{v-u \mid u \in U\} \cap \mathscr{N},$$

for some minimal surface $v \in U$, where $\mathscr{N}$ is the null space of the functional. We are guaranteed a unique solution to the interpolation problem if S is empty (or, equivalently, consists only of the null surface, defined to be zero everywhere). The key question becomes: can we have two surfaces that fit through the known points, have the same measure of surface consistency (the same value as measured by the functional) and differ by an element of the null space? If not, the minimal surface is guaranteed to be unique. Thus, the structure of the boundary conditions provided by the stereo algorithm (or the structure-from-motion algorithm) may be important in deciding which functional is more suitable. Clearly, the smaller the subspace of minimal surfaces, the more likely we are to have a unique minimal surface fitting the known data, as the set S is more likely to be empty.

Recall that the null space of the square Laplacian

$$\Theta_2(f) = \left[\iint (\nabla^2 f)^2 \,\mathrm{d}x\,\mathrm{d}y\right]^{\frac{1}{2}}$$

is the set of all harmonic functions. We wish to know what form of the boundary conditions will uniquely determine the harmonic function. This problem is known as the Dirichlet problem in classical analysis, and it has long been known that if the boundary conditions consist of a series of closed, bounded Jordan curves then the harmonic function is uniquely determined. These are, of course, sufficient, but not necessary conditions. It would appear, however, from these conditions that it is unlikely that the boundary conditions provided by the stereo algorithm will be sufficient to uniquely determine the component of the null space. This follows from the observation that the stereo algorithm is capable of providing boundary values at scattered points in the image, corresponding to the zero-crossings of the convolved image, while the Dirichlet problem is uniquely determined if the boundary values form a closed, bounded Jordan curve. Thus, for the square Laplacian, the best that we can do is to determine a family of most consistent surfaces, which differ by harmonic functions. Referring back to our earlier question, we see that in this case we could have two (or more) surfaces that fit through the known points, have the same measure of surface consistency and differ by an element of the null space. The variation in such a family of surfaces is not consistent with our intuitive notion of indistinguishable surfaces, that is, the difference in the shape of two surfaces that have identical minimal values for the square Laplacian measured over the surface can be noticeably large (see Grimson 1981*b*, *d*, 1982). As a consequence, we consider the square Laplacian to be a poor choice for the functional.

On the other hand, the null space of the quadratic variation

$$\Theta_5(f) = \left[\iint (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)\,\mathrm{d}x\,\mathrm{d}y\right]^{\frac{1}{2}}$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

is the set of all linear functions. The boundary conditions required in this case to uniquely determine the component of the null space are much simpler. In particular, if the stereo algorithm provides at least three non-collinear points, the element of the null space is uniquely determined to be the null surface (the surface that is zero everywhere). It is clear that in almost all imaging situations the stereo algorithm will be capable of providing the necessary boundary conditions, and thus the most consistent surface is uniquely determined.

Thus we have seen that the only possible functionals that can be used to measure the *surface consistency constraint* form a vector space spanned by the square Laplacian operator and the quadratic variation operator. The minimal surface for any such operator satisfies the biharmonic equation in the interior of the region being interpolated, but along the boundaries of the region it may satisfy other differential equations than the minimal solution of any other operator. In general, this implies that the solution surfaces corresponding to different operators will generally differ in shape. To distinguish between possible operators, we examined the form of their null spaces. We showed that the operator with the smallest null space was the quadratic variation. Further, the stereo data is in general sufficient to uniquely determine the component of the null space corresponding to the minimal surface. That is, the surface that minimizes the quadratic variation, subject to passing through the known points provided by the stereo or structure-from-motion algorithms, is uniquely determined.

4.6. *The computational problem*

By combining the results of §§ 4.4 and 4.5, it is now possible to state the computational theory of the problem of interpolating visual surface information.

THE INTERPOLATION OF VISUAL INFORMATION. *Suppose that we are given a representation consisting of surface information at the zero-crossings of a primal sketch description of a scene. Within the context of the visual information available, the best approximation to the original surface in the scene is given by the minimal solution to the quadratic variation in gradient* (*or surface orientation*)

$$\Theta(f) = \left[\iint (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)\,\mathrm{d}x\,\mathrm{d}y\right]^{\frac{1}{2}}.$$

Such approximations are guaranteed to be uniquely 'best' to within an element of the null space of the functiona Θ. *For quadratic variation, the null space is the set of all linear functions. Provided that the set of known point. supplied by the stereo algorithm or by the structure-from-motion algorithm includes at least three non-collinear points, the component of the surface due to the null space is uniquely determined to be the null surface. Hence, the surface most consistent with the visual information is uniquely determined.*

It is worth noting that, although the above statement is phrased in terms of zero-crossings obtained from images convolved with $\nabla^2 G$ filters, the heart of the statement is much broader in scope. The key point is that, to interpolate any surface representation that contains explicit information only at sparse points in the representation, we need to find the 'most conservative' surface consistent with the input information. This implies that between the known surface points the surface should vary as little as possible. Thus, whether those known points correspond to zero-crossings, edges or some other basic descriptor of image changes, the surface interpolation algorithm should construct the surface that minimizes variation in the surface between the known points.

It is interesting to compare the criteria for surface interpolation developed here, as well as the specific theory of surface interpolation stated above, with the work of Barrow & Tenenbaum (1981).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

5. Examples

We have developed a computational model of the process of visual surface interpolation, which stated that the 'best' surface to fit through a set of known points (provided for example by stereo or motion correspondence) was that which minimized the functional of quadratic variation. In

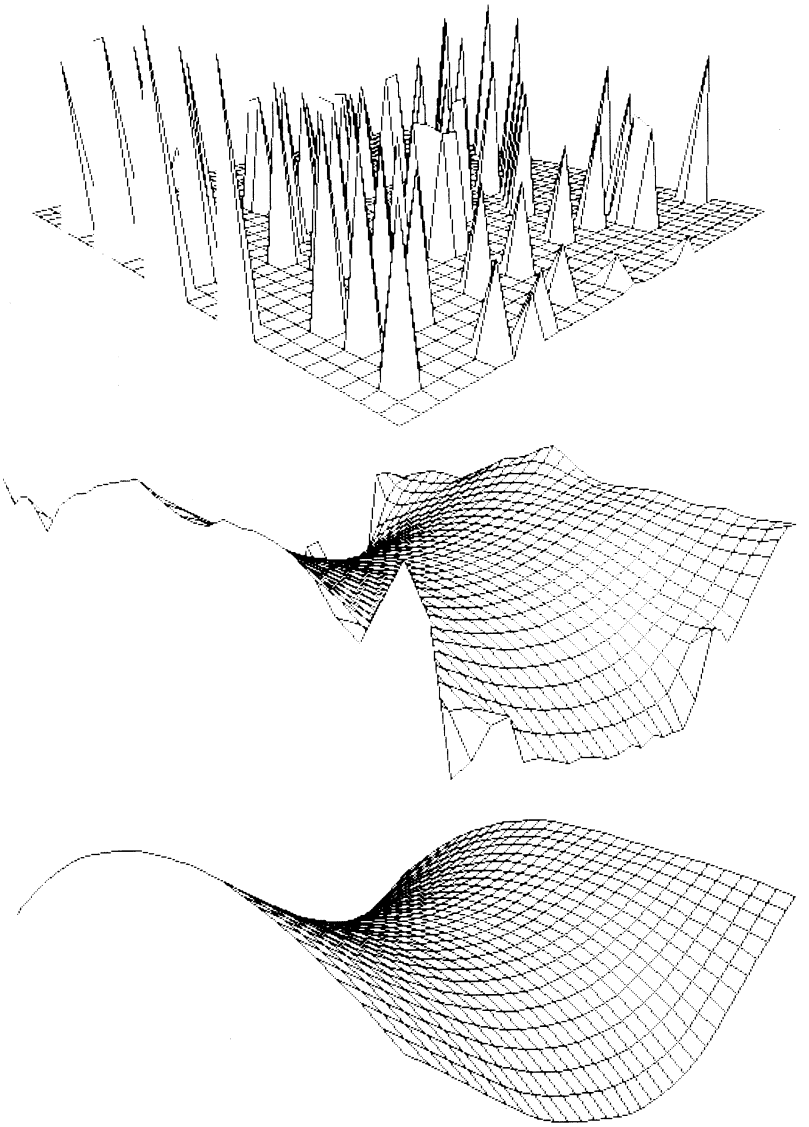


Figure 4. Synthetic example. The top figure shows a synthetic set of boundary conditions, consistent with a hyperbolic paraboloid. The points are chosen at random with a density of 10%. The middle figure shows the surface obtained by minimizing the square Laplacian, while the bottom figure shows the surface obtained by minimizing the quadratic variation. (From Grimson (1981*b*).)

Grimson (1981*d*, 1982), algorithms for computing this surface, given a set of known points, are presented, along with a series of examples and a discussion of their performance. Figures 4–6 illustrate the results of applying the algorithm. Figure 4 shows a synthetic example. The top figure shows the boundary conditions, a set of points lying on a hyperbolic paraboloid, chosen at random. The middle figure shows the surface obtained by minimizing the integral of square

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Laplacian, while the bottom figure shows the results of minimizing the integral of quadratic variation. Figure 5 shows a random dot stereogram and the surface obtained by applying the Grimson (1981 *a*) implementation of the Marr & Poggio (1979) stereo algorithm to this stereo pair and interpolating the result by means of quadratic variation. In figure 6, a natural stereogram is illustrated. The disparity values obtained by applying the Marr–Poggio stereo algorithm are processed in two ways. In figure 6*b* the disparity values are interpolated by using quadratic variation. In figure 6*c*, the disparity values are approximated, by applying the quadratic variation subject to the condition that the surface pass near but not necessarily through the known disparity values. A more complete discussion of the processes may be found in Grimson (1981 *b*, *d*, 1982).

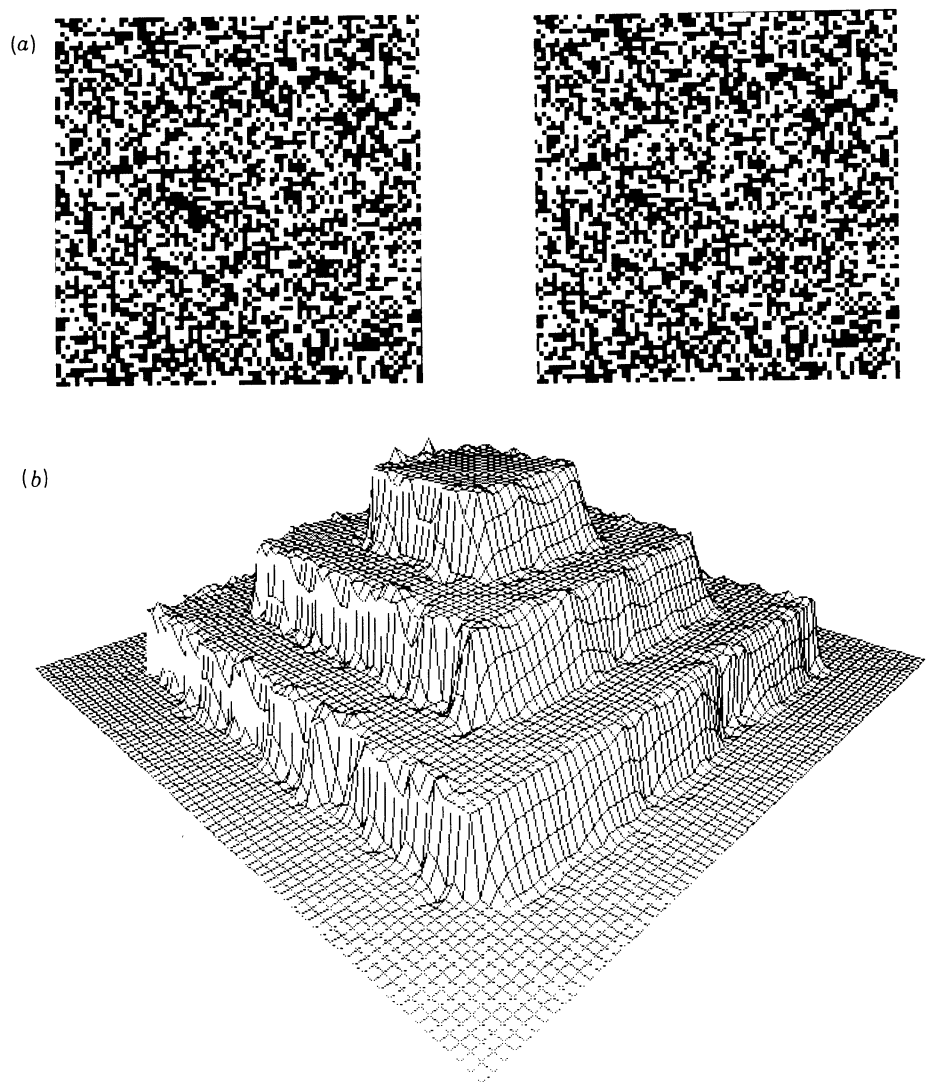


FIGURE 5. A random dot wedding cake. The top figure shows a random dot stereogram of a wedding cake. The bottom figure shows the surface obtained by processing the stereo pair with the Grimson implementation of the Marr–Poggio stereo algorithm and interpolating the result by means of the quadratic variation. (From Grimson (1981 *b*).)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

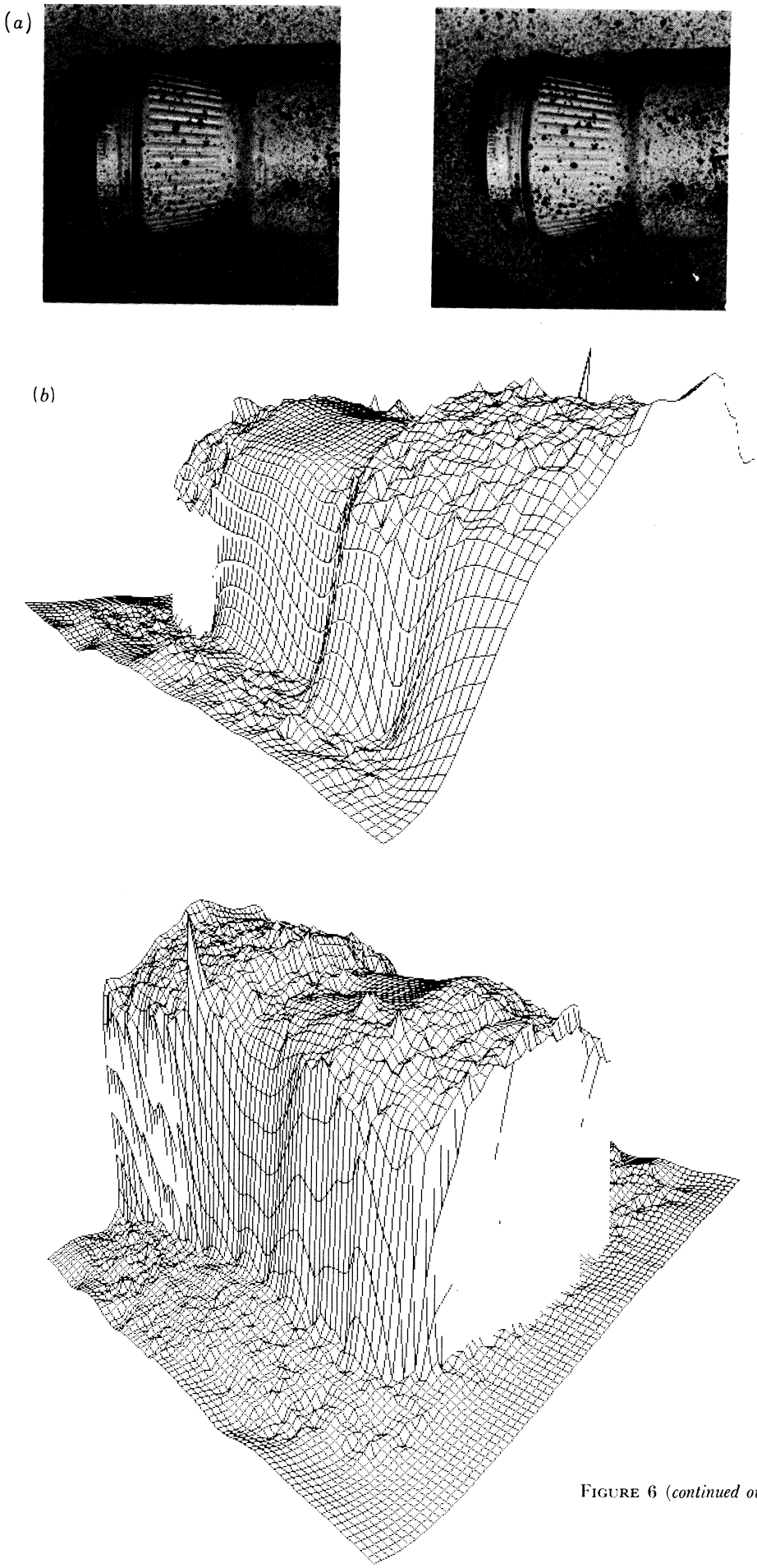


FIGURE 6 (*continued overleaf*).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

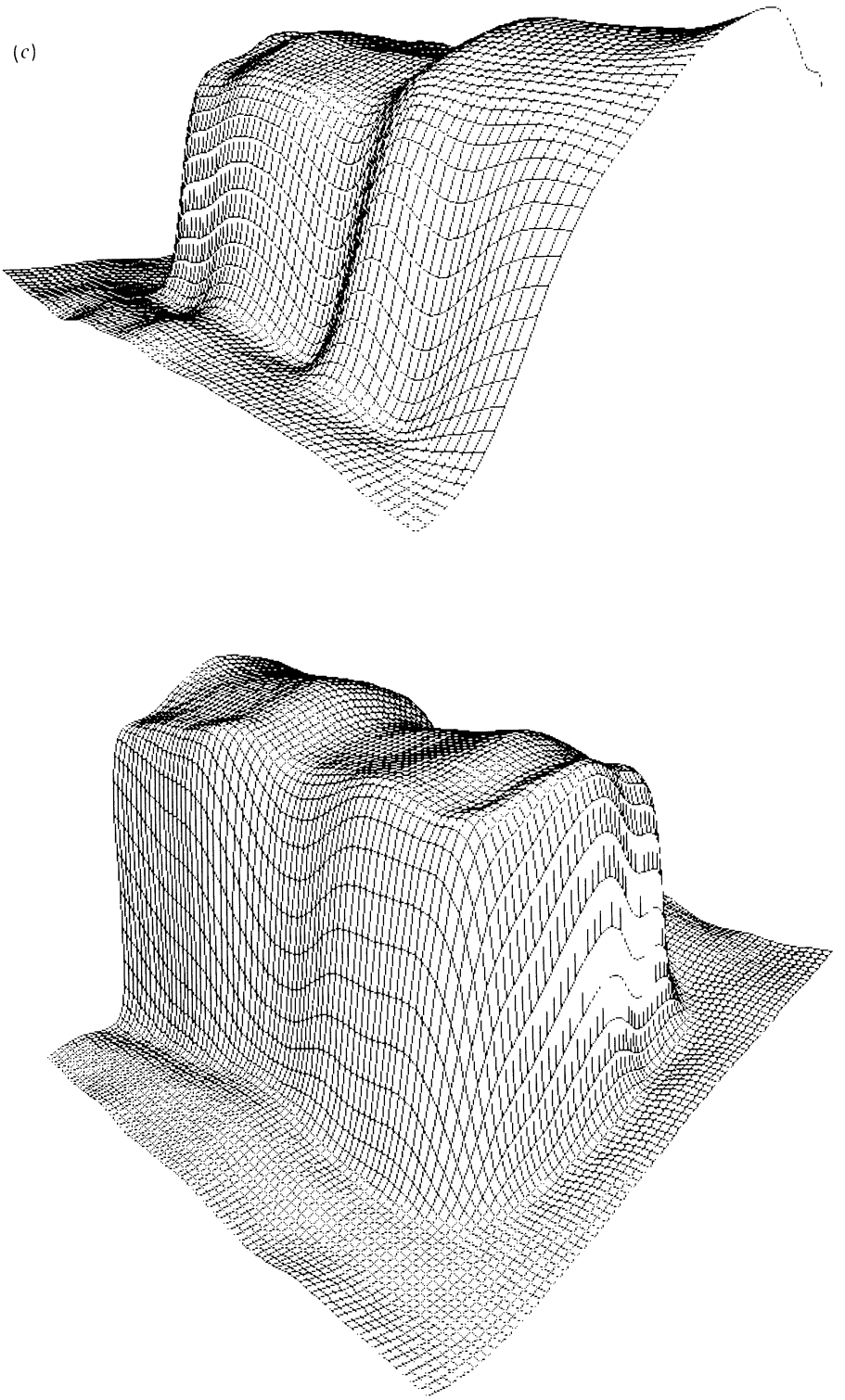


FIGURE 6. The coffee jar. (*a*) Stereogram of a coffee jar that has been painted with random dot spray paint. (*b*) Two views of the interpolated bottle. (*c*) The same two views of an approximated surface, where the effects of incorrect stereo data have been reduced by requiring the surface to pass near but not through the known points. (From Grimson (1981*b*).)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

6. Analysis and refinements

6.1. *Discontinuities*

One of the implicit assumptions of the interpolation process described in this paper is that the pieces of surface are in fact pieces of a single surface. Of course, this will frequently not be so. In this section, we consider what modifications are necessary to account for the existence of several surfaces within a scene. In particular, we address the issue of explicitly computing discontinuities in the surface representation, and the effects of explicit discontinuities on the form of the reconstructed surface.

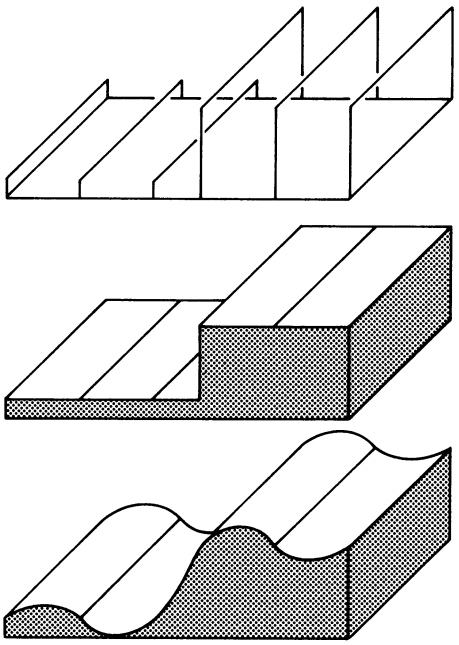


Figure 7. Discontinuities in the surfaces. (*a*) A set of known data points. Intuitively, the correct reconstructed surface would be a pair of planes, with a discontinuity between them, as shown in (*b*). If the interpolation algorithm attempts to reconstruct a surface through the boundary points, without a discontinuity, the result is as shown in (*c*). The sharp change in depth results in a rippling of the surface. (From Grimson (1981*b*).)

One of the problems associated with the failure to make surface discontinuities explicit is that information about the shape of one surface affects the shape of an adjacent surface. This is illustrated in figure 7. A set of known depth points is given in figure 7*a*. Intuitively, the most likely surface to fit through these points would be a pair of planes with a discontinuity in depth between them, shown in figure 7*b*. However, the requirement that a smooth surface fit through these points results in a warping and rippling of the surface that is undesirable, as shown in figure 7*c*. Thus, the lack of explicit discontinuities can affect the shapes of the interpolated surfaces in an unacceptable manner.

To make discontinuities explicit, there are several questions to ask about the process. How are the discontinuities detected? Where are they placed in the representation? When does the detection of discontinuities take place in the overall interpolation process? In §§ 6.1.1–6.2 sections, we shall discuss two possible methods for detecting the discontinuities, and their role in the overall interpolation.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

6.1.1. *Occlusions in the stereo algorithm*

Consider the geometry indicated in figure 8. There are regions of the left image that will not have a corresponding region in the right image, and vice versa. Consequently any zero-crossings in this portion of one image will have no counterpart in the other image and the stereo algorithm should not assign any match to such zero-crossings. Hence one possible mechanism for detecting occlusions would be to search for portions of the image that contain unmatched zero-crossings. Then the interpolation can be restricted to take place only over those sections of the image that are bounded by zero-crossings with known disparity values.

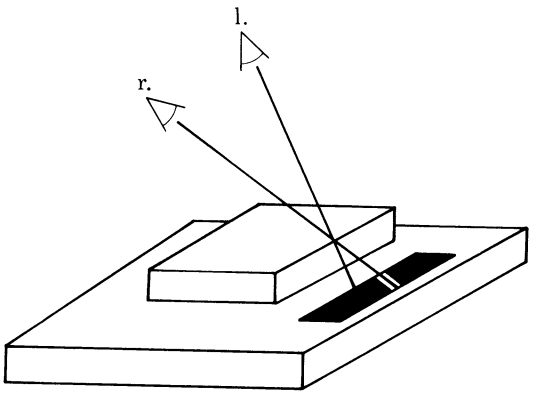


FIGURE 8. Occlusions. The upper surface occludes portions of the lower surface in each eye. These portions are different for the two eyes. The cross-hatched area of the lower surface indicates the region of the surface visible to the left eye (l.) but not to the right (r.). (From Grimson (1981 *b*).)

This method would detect the discontinuities before the interpolation, since it uses stereo information directly to locate the occlusions. A problem with the method is that it will not detect all discontinuities, only those in the horizontal direction. Discontinuities that occur in the vertical direction do not cause occlusions. Hence any method for detecting discontinuities that relies only on the unmatched zero-crossings will be incomplete.

6.1.2. *The primal sketch revisited*

An integral part of most computational theories proposed as models of aspects of the human visual system is the use of computational constraints based on assumptions about the physical world (Marr 1976, 1982; Marr & Poggio 1979; Marr & Hildreth 1980; Ullman 1979*a*). In some of the computational theories, the constraints are explicitly checked for validity within the algorithm (e.g. Ullman's rigidity constraint in recovering structure-from-motion). In others, the constraints are simply assumed to be true, and are not explicitly checked (e.g. the Marr & Poggio uniqueness constraint in stereopsis). Can any aspect of the surface consistency constraint be explicitly checked and used by the algorithm?

The basic notion of the surface consistency constraint (Grimson 1981 *b*, *c*) is that the surface cannot undergo a radical change in shape without having an accompanying zero-crossing in the convolved image. Implicit in this constraint is the assumption that the portion of the image being examined in fact corresponds to a single object. Thus, one could propose that if the shape of the interpolated surface forces a zero-crossing in a location for which none exists in the primal sketch

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

then such a zero-crossing indicates a location at which the assumption of a single object is violated. Such zero-crossings could then be taken as indicative of a surface discontinuity.

Perhaps the simplest method of detecting such discontinuities is again to use ideas inherent in the primal sketch (Marr & Hildreth 1980; Hildreth 1980). The primal sketch creates descriptions of points in the image associated with inflexions in intensity, for a range of resolutions. Since the image intensities may be considered as a type of three-dimensional surface, the primal sketch operators essentially detect discontinuities in the image intensities for a range of resolutions. Thus, we could apply the same type of analysis to the detection of surface discontinuities, where now the surface on which the operators apply is the reconstructed depth surface, rather than the intensity surface.

It is worth noting that not only should the operators be of the form used in the extraction of the primal sketch, but it may also be useful to use a range of operators, as in the primal sketch. One reason for using a range of zero-crossing detectors was that surface changes, and hence intensity changes, could take place over a wide range of scales. This is still true for surface descriptions such as have been constructed for the coffee jar or the wedding cake. Thus, surface discontinuities corresponding to occluding edges will frequently tend to correspond to large surface changes, while internal surface discontinuities, due to a warping of the surface, will tend to correspond to small surface changes. By using a range of $\nabla^2 G$ operators, we can extract both occluding contour discontinuities and ripples or warpings of the surface itself.

Note that this method requires that the surface interpolation has already taken place, before it can be applied. Since one of the general requirements on an algorithm is that it be rapid, we must consider the consequence of detecting discontinuities after the interpolation of the surfaces. There are two main reasons for the explicit detection of discontinuities. One is that such an explicit representation of this information will allow higher level processes, such as recognition, or extraction of axes for three-dimensional shape analysis, to operate more easily, since the process serves to make implicit information explicit. However, a second reason is to create more accurate surface representations, by removing the type of effect illustrated in figure 7*c*. If the process used to isolate discontinuities takes place after interpolation, and if the interpolation process requires the discontinuities to improve the interpolated surface approximation, we must propose an interpolater that passes over the surface information twice; first to produce an initial description, and secondly to refine the description after the detection of discontinuities. We must then question whether such a two-pass process will affect our constraint of rapid algorithms. Fortunately, the answer is no, since the surface approximation obtained without explicitly accounting for the discontinuities is very close to the limiting surface except in the areas of the discontinuities (that is, any effects of the discontinuities are quickly damped out as one moves across the surface). Thus the initial starting position for the second pass of the interpolation algorithm is very close to the limiting surface, and only a few iterations will be needed to refine the surface approximation.

6.1.3. *Forced inflexions*

A third possibility for detecting surface discontinuities is to use the surface consistency constraint explicitly. Recall that this constraint stated that, if the surface albedo is roughly constant, the illumination is constant, and the surface has continuous first- and second-order partial derivatives, then a large variation in the surface orientation will generally result in a zero-crossing in the convolved image. Suppose that the stereo data are such that an inflexion is forced in the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

interpolated surface. If we assume that the surface albedo and the illumination are constant, then the surface consistency constraint implies that the surface cannot have continuous first- and second-order derivatives. In other words, there must be a crease or a discontinuity in the surface.

6.2. *Interpolation over occluded regions*

Even though occluded regions of the image can only be viewed from one eye, the human system still associates a depth value with these regions. This has an interesting implication for the interpolation algorithm. For most occluded regions, the only depth information available is at the edges of the occluded region. Psychophysical experiments have shown that the occluded region is always perceived at the depth of the lower surface. Thus in figure 8 the occluded region would be perceived at the level of the lower surface. Note that this is consistent with the physics of the situation, since if the occluded region were perceived at the level of the upper surface then it should be visible to the right eye, and this is not the case.

This observation suggests that, when an occlusion is detected, it is explicitly located along the occluding boundary corresponding to the edge of the nearer object. This allows the occluded region itself to be associated with the lower surface, and the interpolation algorithm will fill in surface values for the occluded region from this lower surface.

This raises an interesting psychophysical prediction. The psychophysical literature has examined planar surfaces and their occlusions, as in figure 8. If the interpolation method developed here is given an explicit discontinuity along one edge of the occluded region, it will correctly fill in the region as an extension of the lower plane. Of interest is the case in which the occluded region is not planar. For example, consider a cylindrical object. If the interpolation algorithm is given this type of input, it will fill in the occluded portions of the surfaces as a smooth continuation of the curved cylinder. If the interpolation algorithm correctly models interpolation by the human visual system, then this predicts that the surface perception for human observers in this situation should also be that of a smooth cylinder. While informal experiments indicate that this is true, the prediction has not yet been rigorously tested psychophysically.

6.3. *Noise removal*

Although in general the Marr–Poggio stereo algorithm is very good at matching zero-crossings correctly (especially for random dot patterns), incorrect disparity values may sometimes be assigned to regions of the image. These incorrect values can be considered as noise superimposed on the correct surface. Since the surface interpolator explicitly attempts to fit a surface through all the disparity points, such noise points can affect the shape of the surface approximation. Indeed, the effect of these noise points can spread over a noticeable portion of the surface before the nearby disparity values can damp out its effect. Thus, it would be preferable to remove these noise points, or at least neutralize their effect on the approximated surface shape. One possibility is that if a two-pass interpolator is used, as suggested in the previous section, the detection of surface discontinuities will isolate such noise points from the rest of the surface, and the second pass of the interpolator will adjust the surface approximation to remove the influence of the noise points on the first pass approximation. Certainly this will be true for noise points with disparity values far removed from the correct values. For noise points whose disparity values are only slightly different from the correct surface disparities, the difference does not really matter. However, the final result would be that the noise points, while being isolated from the rest of the correct surface, would still remain in the final surface description. It would be preferable to completely remove such points.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Is it possible to identify and remove noise points from the disparity map? If the noise points are isolated spatially, then it is possible to identify them as undesirable. This follows from the form of the primal sketch operators. The case to consider is that in which we must distinguish between a set of noise points in a disparity map and a small object separated in depth from the rest of the scene. For the small object, the size of the zero-crossing contour is limited by the size of the available operator, and hence there is a minimum size of zero-crossing contour that the operator will yield about the object. If the number of zero-crossing points that differ significantly from their neighbours is less than this minimum, we may conclude that the points are noise, and thus remove them. This will result in an improved surface approximation.

6.4. *Acuity*

It can be seen from the example of the interpolated wedding cake in figure 5 (see also Grimson 1981*d*, 1982) that the interpolated surface contains a bumpy quality that may not be consistent with the original object. How can this be explained? The effect occurs in part because the disparity values are specified only to within a pixel. This yields a fairly coarse disparity map which results in the observed bumps. Hence, one method of removing the bumps would be to improve the accuracy of the disparities obtained by the algorithm. Note that some improvement in disparity accuracy is necessary if the algorithm is to be consistent with the human system. If we roughly equate pixels with receptors, then a pixel corresponds to roughly 27″. The implementation of the stereo algorithm computed disparity to within a pixel, while humans are capable of stereo acuity to a resolution of 2–10 s (Howard 1919; Woodburne 1934; Berry 1948; Tyler 1977).

To account for finer disparity values, it is necessary to localize the zero-crossing to a better accuracy than has been done so far. Since the convolution values are only specified at each pixel, one method for more accurately specifying the zero-crossing positions is to interpolate between the known convolution values (Crick *et al.* 1980; Marr *et al.* 1979; Hildreth 1980). Perhaps the simplest method is to rely on the observation of Hildreth that for most cases even a simple linear interpolation will give extremely accurate localization of the zero-crossings (see also MacVicar-Whelan & Binford 1981). The addition of finer resolution depth information may improve the performance of the algorithm.

This example also raises a question of scale. Depending on the application of the surface specification, different amounts of resolution may be required. For example, if the ultimate goal of the surface specification is to obtain a rough idea of the position and shape of the surfaces in a scene, the spatial resolution at which surface information must be made explicit may not be critical. In this case, the known data from the stereo algorithm may be sampled at a coarser resolution, before the interpolation takes place. This should result in a smoother surface approximation. Further, although the reconstructed surface is less exact in terms of fine variation of the surface shape, the overall shape of the bottle in figure 6 is still preserved in this interpolation.

6.5. *Psychophysics*

I close by listing a series of psychophysical questions of relevance to the interpolation process.

(i) What is the form of the surface perceived in occluded regions? In particular, the minimization of quadratic variation suggests that, if a portion of a curved object is occluded, then the surface in the occluded region should also be curved and should minimize the quadratic variation across that region.

(ii) Figure 7 suggests that, if discontinuities are not explicitly demarked in the interpolation

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

process, a warping of the reconstructed surface (similar to Gibb's phenomena) will result. While, in principle, such ripples in the surface are undesirable, it is worth asking whether the human system specifically accounts for discontinuities before interpolation occurs. This may be rephrased by asking whether in stereoscopic situations similar to figure 7 we perceive a warping of the surface in depth similar to Mach bands?

(iii) We have suggested that there are several possible functionals that could be used to determine the most consistent surface. Based on algorithmic and mathematical arguments, we choose the quadratic variation. Can we test the shape of the reconstructed surface psychophysically? In particular, can we distinguish psychophysically between the minimum surface under quadratic variation and the minimum surface under some functional, such as the square Laplacian? Is the reconstructed surface psychophysically consistent with the surface computed by quadratic variation?

(iv) What is the spatial resolution of the reconstructed surface? That is, what is the spacing of the grid upon which the values of the reconstructed surface are computed?

The answers to these questions will help verify or correct the theory of visual surface interpolation developed in this paper.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643 and in part by National Science Foundation Grant MCS 77-07569.

The author wishes to express his gratitude for many useful comments and discussions to David Marr, Tommy Poggio, Shimon Ullman, Berthold Horn, Mike Brady, Whitman Richards, Tomas Lozano-Perez, Marilyn Matz and Ellen Hildreth.

References

Barrow, H. G. & Tenenbaum, J. M. 1981 Interpreting line drawings as three-dimensional surfaces. *Artif. Intell.* **17** (special issue on computer vision, pp. 75–116).

Berry, R. N. 1948 Quantitative relations among vernier, real depth, and stereoscopic depth acuities. *J. exp. Psychol.* **38**, 708–721.

Brady, J. M. & Horn, B. K. P. 1982 Rotationally symmetric operators for surface interpolation. *MIT artif. Intell. Lab. Memo* no. 654. (Also to appear in *Comp. Graph. Image Processing*.)

Courant, R. & Hilbert, D. 1953 *Methods of mathematical physics*, vol. 1. New York: Interscience Publishers.

Crick, F. H. C., Marr, D. & Poggio, T. 1980 An information processing approach to understanding the visual cortex. In *The organization of the cerebral cortex* (ed. E. O. Schmitt, F. G. Worden & G. S. Dennis), pp. 505–533. Cambridge, Massachusetts: M.I.T. Press.

Duchon, J. 1975 Fonctions-spline du type plaque mince en dimension 2. *Univ. Grenoble tech. Rep.* no. 231.

Duchon, J. 1976 Fonctions-spline a energie invariante par rotation. *Univ. Grenoble tech. Rep.* no. 27.

Forsyth, A. R. 1960 *Calculus of variations*. New York: Dover Publications.

Grimson, W. E. L. 1980 Computing shape using a theory of human stereo vision. Ph.D. thesis, Department of Mathematics, Massachusetts Institute of Technology.

Grimson, W. E. L. 1981*a* A computer implementation of a theory of human stereo vision. *Phil. Trans. R. Soc. Lond.* B **292**, 217–253.

Grimson, W. E. L. 1981*b* *From images to surfaces: a computational study of the human early visual system*. Cambridge, Massachusetts: M.I.T. Press.

Grimson, W. E. L. 1981*c* The implicit constraints of the primal sketch. *MIT artif. Intell. Lab. Memo* no. 663.

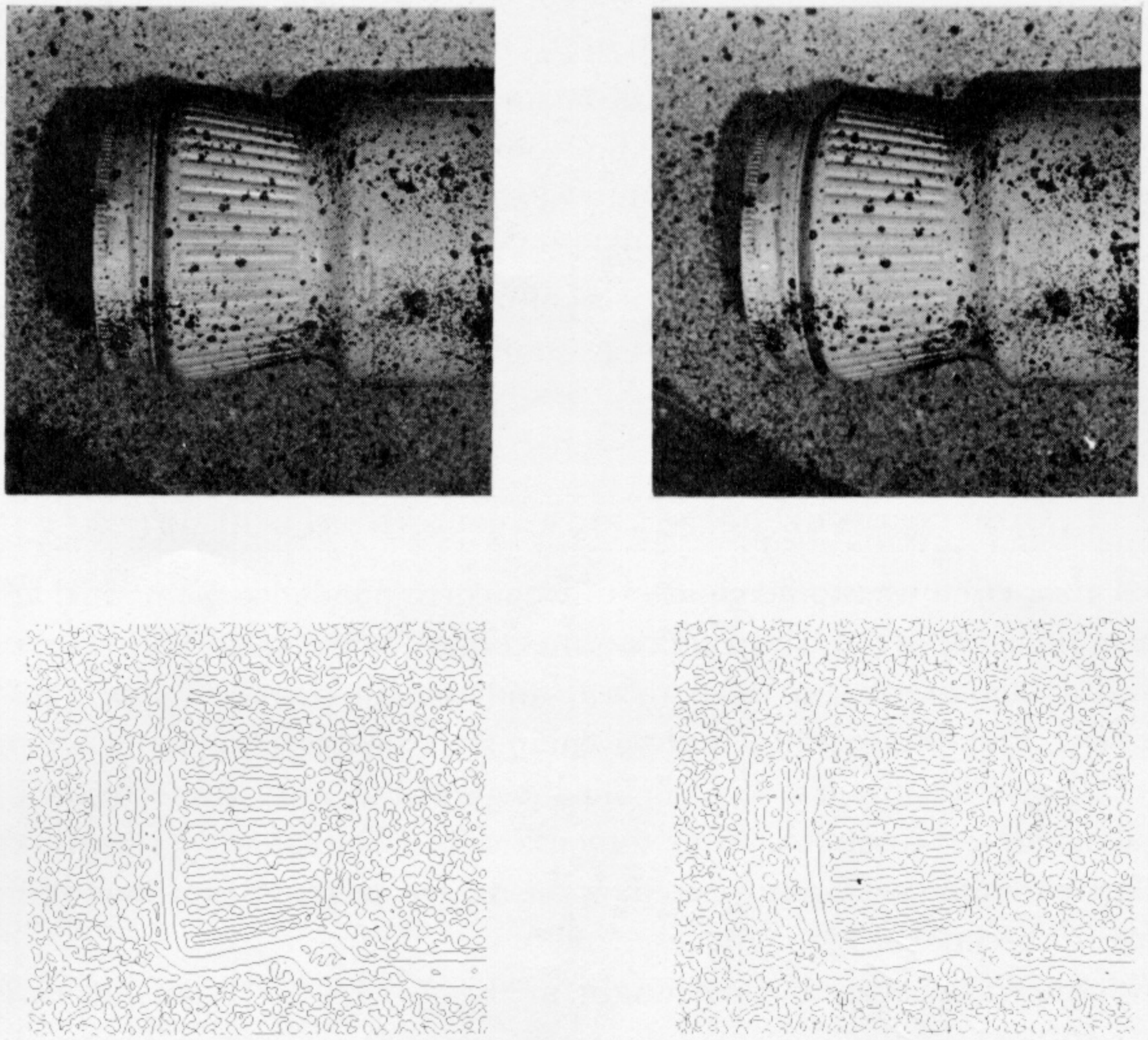
Grimson, W. E. L. 1981*d* A computational theory of visual surface interpolation. *MIT artif. Intell. Lab. Memo* no. 613.

Grimson, W. E. L. 1982 An implementation of a computational theory of visual surface interpolation. *Comput. Graph. Image Processing*. (In the press.)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Helmholtz, H. 1925 *Physiological optics*, vol. 3. New York: Optical Society of America.
Hildreth, E. C. 1980 Implementation of a theory of edge detection. S.M. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.
Horn, B. K. P. 1970 Shape from shading: a method for obtaining the shape of a smooth spaque object from one view. *MIT Project MAC tech. Rep.* no. MACTR-79.
Horn, B. K. P. 1975 Obtaining shape from shading information. In *The psychology of computer vision* (ed. P. H. Winston), pp. 115–155. New York: McGraw-Hill.
Horn, B. K. P. 1975 Obtaining shape from shading information. In *The psychology of computer vision* (ed. P. H.
Howard, J. H. 1919 A test for the judgement of distance. *Am. J. Ophthal.* **2**, 656–675.
Johannson, G. 1964 Perception of motion and changing form. *Scand. J. Psychol.* **5**, 181–208.
Julesz, B. 1960 Binocular depth perception of computer-generated patterns. *Bell Syst. Tech. J.* **39**, 1125–1162.
Julesz, B. 1971 *Foundations of cyclopean perception*. University of Chicago Press.
Longuet-Higgins, H. C. & Prazdny, K. 1980 The interpretation of a moving retinal image. *Proc. R. Soc. Lond.* B **208**, 358–397.
MacVicar-Whelan, P. J. & Binford, T. O. 1981 Intensity discontinuity location to subpixel precision. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* (*Vancouver, Canada*), pp. 752–754.
Marr, D. 1976 Early processing of visual information. *Phil. Trans. R. Soc. Lond.* B **275**, 483–534.
Marr, D. 1978 Representing visual information. *Lect. Life Sci.* **10**, 101–180.
Marr, D. 1982 *Vision: a computational investigation in the human representation and processing of visual information*. San Francisco: W. J. Freeman.
Marr, D. & Hildreth, E. C. 1980 Theory of edge detection. *Proc. R. Soc. Lond.* B **207**, 187–217.
Marr, D. & Nishihara, H. K. 1978 Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond.* B **200**, 269–294.
Marr, D. & Poggio, T. 1977 From understanding computation to understanding neural circuitry. *Neurosci. Res. Progr. Bull.* **15** (3), 470–488.
Marr, D. & Poggio, T. 1979 A theory of human stereo vision. *Proc. R. Soc. Lond.* B **204**, 301–328.
Marr, D., Poggio, T. & Hildreth, E. 1979 The smallest channel in early human vision. *J. opt. Soc. Am.* **70** (7), 868–870.
Mayhew, J. E. W. & Frisby, J. P. 1981 Psychophysical and computational studies towards a theory of human stereopsis. *Artif. Intell.* **17**, 379–386.
Miles, W. R. 1931 Movement interpretations of the silhouette of a revolving fan. *Am. J. Psychol.* **43**, 392–505.
Richter, J. & Ullman, S. 1980 A model for the spatio-temporal organization of X and Y-type ganglion cells in the primate retina. *MIT artif. Intell. Lab. Memo* no. 573. *Biol. Cybernetics*. (In the press.)
Rudin, W. 1973 *Functional analysis*. New York: McGraw-Hill.
Tyler, C. W. 1977 Spatial limitations of human stereoscopic vision. *Soc. photo-opt. Instrum. Engrs* **120**.
Ullman, S. 1979*a* *The interpretation of visual motion*. Cambridge, Massachusetts: M.I.T. Press.
Ullman, S. 1979*b* Relaxation and constrained optimization by local processes. *Comput. Graph. Image Processing* **10**, 115–125.
Wallach, H. & O'Connell, D. N. 1953 The kinetic depth effect. *J. exp. Psychol.* **52** (5), 571–578.
Wheatstone, C. 1838 Contributions to the physiology of vision. Part I. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Phil. Trans. R. Soc. Lond.* **128**, 371–394.
Woodburne, L. S. 1934 The effect of a constant visual angle upon the binocular discrimination of depth differences. *Am. J. Psychol.* **46**, 273–286.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES



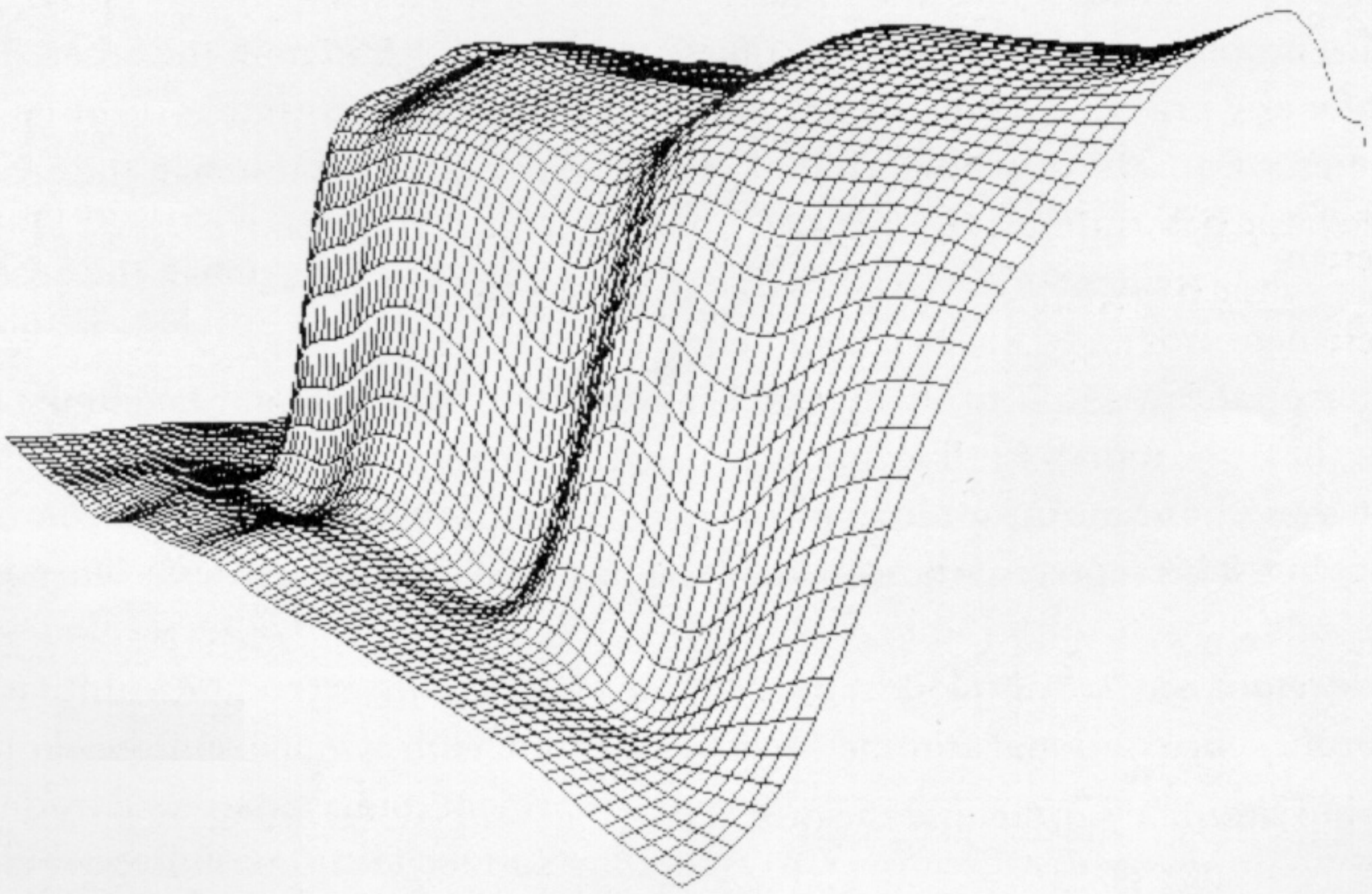
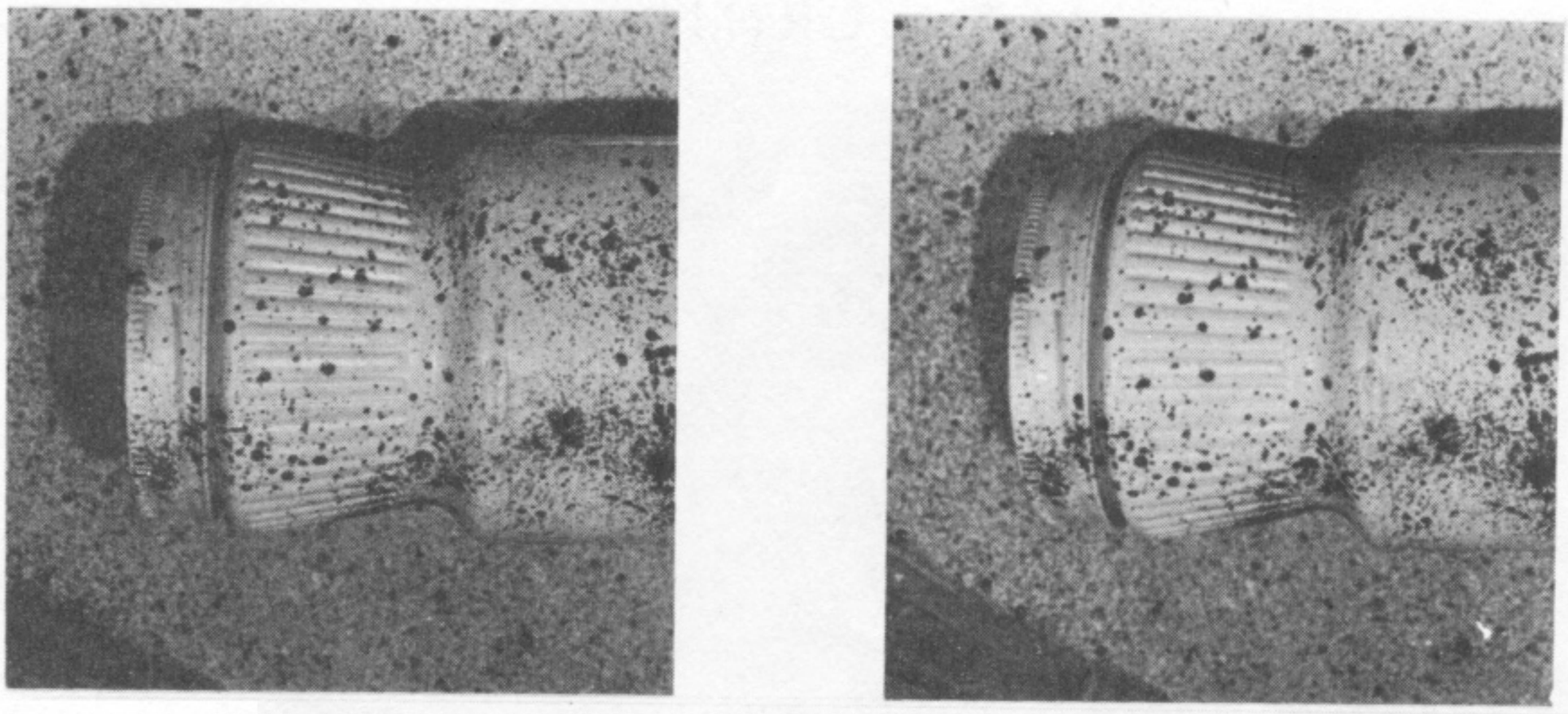


FIGURE 2. Example of processing. The top pair of images is a stereo pair of a scene. The middle pair illustrates the zero-crossings obtained from the images for one size of $\nabla^2 G$. The final image illustrates an interpolated surface description, formed by matching the zero-crossing descriptions, computing the distance to those points based on the difference in projection, and then interpolating the result.

(*a*)



(*b*)

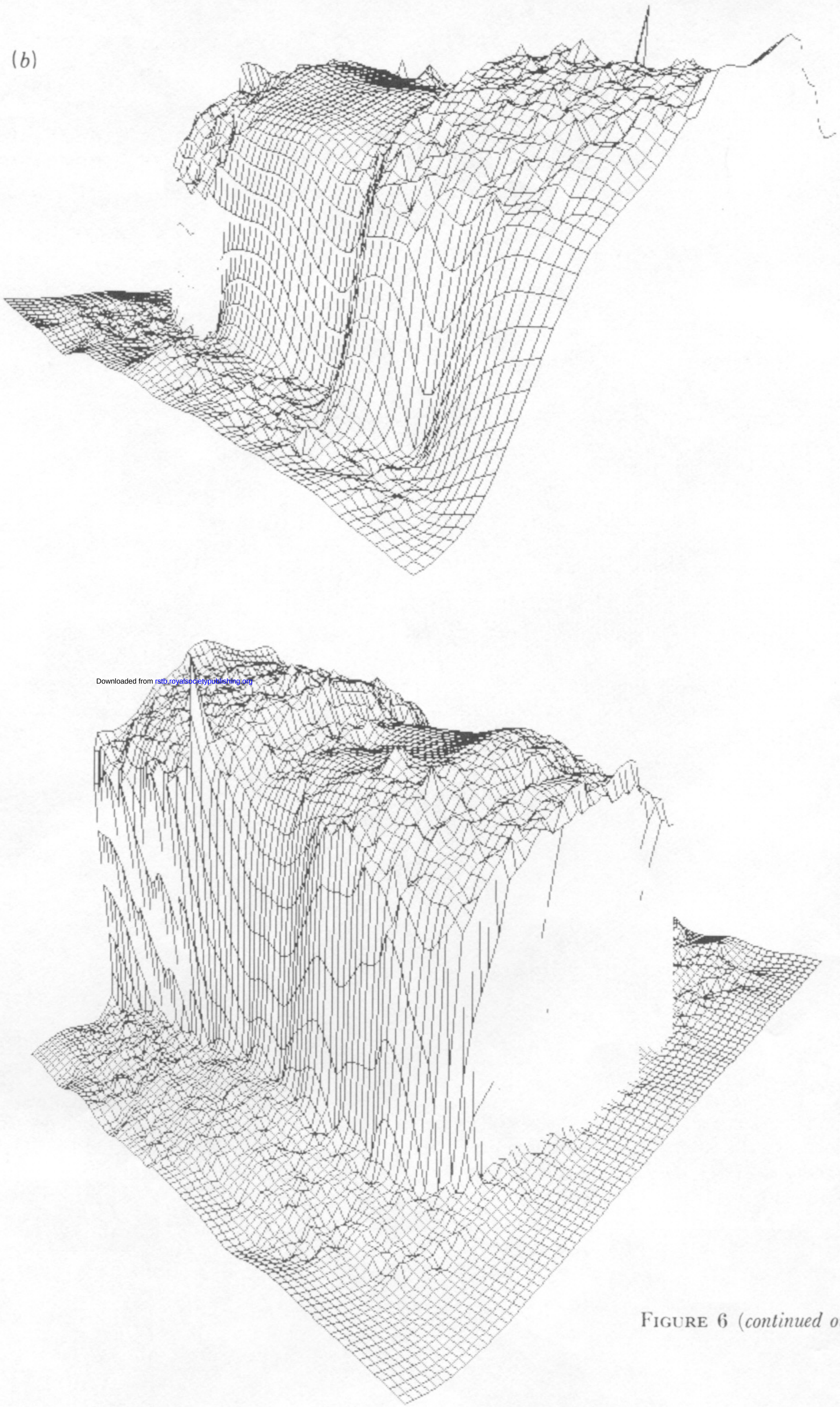


FIGURE 6 (*continued overleaf*).