Technical Report 1117

A Modern Differential Geometric Approach to Shape from Shading

Bror V. H. Saxberg

MIT Artificial Intelligence Laboratory

This blank page was inserted to preserve pagination.

A Modern Differential Geometric Approach to Shape from Shading

by

Bror Valdemar Haug Saxberg

Revised version of a thesis submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the Massacusetts Institute of Technology in June of 1989

©Massachusetts Institute of Technology 1989

This empty page was substituted for a blank page in the original document.

A Modern Differential Geometric Approach to Shape from Shading

by

Bror V. H. Saxberg

Abstract

Our visual system is a remarkably flexible and reliable source of information about the world. One of the major challenges for appreciating how vision contributes to our understanding of the world is understanding how it copes with the wide variety of lighting conditions, surfaces, and surface markings to provide accurate representations of the surfaces around us. The goal of the research reported here is to gain a better theoretical understanding of what lies behind the visual system's ability to generate robust surface interpretations from single grey scale images of smooth surfaces. In the course of doing this, a new robust shape from shading method is developed.

The image irradiance equation is written using coordinate independent notation and concepts from modern differential geometry and global analysis. This is done to help make explicit the assumptions about the image formation process, and to delay making these assumptions as long as possible. The method of characteristic strips used by Horn (Horn, 1975) can be interpreted as a dynamical system on the five-dimensional space of tangent planes, $\mathcal{C}(\mathbb{R}^3, 2)$. Modern methods for analyzing the behavior of dynamical systems are used to show that solution surfaces for the shape from shading problem are invariant manifolds of the flow generated by the image dynamical system. The rest of the analysis assumes orthographic projection of the image and a space-invariant reflectance function, but does not assume any particular form or symmetry for the reflectance function.

Near critical points in the image dynamical system due to certain critical points in a smooth image, in general (i.e. in the absence of special symmetries) the dynamical system approach implies there will only be four possible smooth solution surfaces for the shape from shading problem. Two of these are the stable and unstable manifolds associated with the image dynamical system critical point. Two implementations for finding the unstable (or stable) manifold in this dynamical system are developed using the image dynamical system directly.

The shading information in a patch containing a piece of the bounding contour is also examined, and it appears to contribute more to an assessment of a reflectance function choice than to the determination of patches of solution surfaces consistent with the image.

Finally directions for future work are suggested, and some guidelines and caveats are provided for the development of image analysis systems based on these ideas.

This empty page was substituted for a blank page in the original document.

Acknowledgements

I'd like to thank my supervisor, Tomaso Poggio, for his warm support over the years. He has always been willing to listen to unusual approaches, and can see where pitfalls lie even from far off.

Alec Norton, a good friend for a number of years, contributed his mathematical insight and geometric expertise, and has been a willing backboard. He has been a sharp-eyed troubleshooter for my work. Both he and the Schatz brothers, David and Peter, have been staunch companions on many a strange journey—I wish all were as lucky in having friends such as these.

I thank the other members of my committee, Berthold Horn and Eric Grimson, for their thoughtful suggestions and their encouragement to do experiments, rather than a purely theoretical project. Their comments have made this a stronger piece of research.

My special thanks to my parents, Borje and Aase Saxberg, for giving us a warm home in which to learn, read, and grow. Their supportive encouragement of good work at school and their patience in the face of occasional attacks of excessive high spirits have made us what we are. My brother, Bo, has been a great friend–he too has been a willing listener to some strange ideas, and I have been fortunate to have him nearby for so many years.

I would be a basket case without the support of my wife, Denise. Without her helping to keep the spirits up and the feet moving over the years, all this work would have been a lot less fun. There's nothing like a home filled with good spirits to keep the creative pot boiling.

Finally, I thank all those who have contributed financially to making this research possible. This research has been done at the Massachusetts Institute of Technology in the Artificial Intelligence Laboratory and in the Center for Biological Information Processing. Support for the A.I. Laboratory's artificial intelligence research is provided by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76–85–C–0010, and in part by ONR contract N00014–85–K–0124. This research was sponsored by a grant from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Artificial Intelligence Center of Hughes Aircraft Corporation; by the Alfred P. Sloan Foundation; by the National Science Foundation; by the Artificial Intelligence Center of Hughes Aircraft Corporation (S1-801534-2); and by the NATO Scientific Affairs Division (0403/87). Tomaso Poggio is supported by the Uncas and Helen Whitaker Chair at MIT, Whitaker College. I have been supported at various times by a Tau Beta Pi Fellowship, a Mortar Board Fellowship, an MIT Fellowship, and NIH Medical Scientist Training Grant #T32GM07753. My special thanks go to IBM, who supported me for three years with an IBM Fellowship: the peace of mind that came from having secure and sufficient funding during that time set the stage for whatever creative work has been done.

This empty page was substituted for a blank page in the original document.

Table of Contents

Abstract **2**

Acknowledgements **3**

Table of Contents **4**

1. Introduction and Background **7**

1.1 Introduction 8

1.2 Background 9

1.3 Overview of the Thesis 18

2. Image Projection **23**

2.1 Mathematical Preliminaries 23

2.1.1 Manifolds 23

2.1.1.1 Equivalence Class Approach 24

2.1.1.2 Manifolds 26

2.1.2 Tangent Spaces 30

2.2 The Image Projection Map 38

2.2.1 The Bounding Contour 42

3. The Image Irradiance Equation **45**

3.1 Mathematical Preliminaries 46

3.1.1 $\mathcal{C}(\mathbb{R}^3, 2)$ 46

3.1.2 $\mathcal{C}(\mathbb{R}^3, 2)$ and a Useful Coordinate System 47

3.1.3 Lifting of a Surface 50

3.1.4 Contact 1-forms 52

3.2 The Reflectance Function and Image Formation 56

3.2.1 Cut vs. Rolled Edges 57

3.2.2 The Invariant Image Irradiance Problem 60

Appendix to Chapter 3 62

A3.1 Smooth overlap for generalized (x, y, z, p, q) charts on $\mathcal{C}(\mathbb{R}^3, 2)$ 62

A3.2 Contact 1-forms 64

4. The Image Dynamical System 67

4.1 Mathematical Preliminaries 69

4.1.1 From First Order PDE to Vector Field 69

4.1.1.1 Differential Ideals and the Frobenius Theorem 69

4.1.1.2 Isovector Fields and Extension of Solutions 73

4.1.2 Vector Fields and Flows as Dynamical Systems 74

4.2 The Image Irradiance Dynamical System 79

Appendix to Chapter 4 84

A4.1 The Dimension of the Characteristic Subspace 84

A4.2 Extension of Solutions by Isovector Field Flow 87

5. Critical Points of the Image Dynamical System 91

5.1 Mathematical Preliminaries 92

5.1.1 Critical Points and Invariant Manifolds 93

5.1.2 Hyperbolic Critical Points 95

5.2 Image Critical Points 98

5.2.1 Location of Critical Points 98

5.2.2 Good Critical Points 104

5.2.3 The Linearized and Reduced Image Dynamical System 106

5.2.4 Eigenvalues of $\tilde{X}'$ 108

5.2.5 Solution Surfaces Near a Good Critical Point 112

5.2.6 The Fundamental Instability of a True Image Dynamical System 116

5.3 The Lambda-Lemma Method for Finding Solution Surfaces 119

5.3.1 The Intuition 120

5.3.2 Fixed Grid Algorithm 122

5.3.2.1 Theory 123

5.3.2.2 Experiments 126

5.3.3 Distorting Grid Algorithm 138

5.3.3.3 Theory 138
5.3.3.4 Experiments 142
5.3.4 Conclusions on Implementations 147
Appendix to Chapter 5 150
A5.1 Similar Signatures for B and $A = P^T BP$. 150
A5.2 Eigenvalues of $A = BC$ When B and C Are Symmetric. 151
A5.3 Eigenvalues of $A = BC$, $C = Q^T BQ$, B Symmetric. 155
6. Image Data Near the Bounding Contour 157
6.1 Introduction 157
6.2 A Motivating Example 160
6.3 The General Case: Linearization Approach 172
6.4 The General Case: Power Series Analysis 188
6.5 Bounding Contour Conclusions 194
Appendix to Chapter 6 196
6.1 Linear Dependence of Power Series Constants 196
7. Conclusions and Pointers 199
7.1 Conclusions 199
7.2 Future Work 202
References 209

Chapter 1

Introduction and Background

Our visual system is a remarkably flexible and reliable source of information about the world. A major challenge is understanding how vision copes with the wide variety of lighting conditions, surfaces, and surface markings to provide accurate representations of the surfaces around us.

The goal of the research reported here is to gain a better theoretical understanding of the visual system's ability to generate robust surface interpretations from single grey scale images. In the course of doing this, we also suggest a new type of shape from shading method. We will write the image irradiance equation using notation and concepts from modern differential geometry and global analysis. The method of characteristic strips used by Horn (Horn, 1975) can be interpreted as a dynamical system on the five-dimensional space of tangent planes, $\mathcal{C}(\mathbb{R}^3, 2)$. We will make use of modern methods for analyzing the behavior of dynamical systems. Some of these tools have already been brought to bear on knotty research areas such as quantum mechanics, relativity, and cosmology (Hawking and Ellis, 1973, Edelen, 1985, Abraham and Marsden, 1985). We show that solution surfaces for the shape from shading problem are invariant manifolds of the flow generated by the image dynamical system, and that the stable and unstable manifolds associated with certain critical points in the image play an important role in determining solution surfaces. We will show two implementations for finding the unstable (or stable) manifolds in

this dynamical system using the image dynamical system directly. We also analyze the shading information in a patch containing a piece of the bounding contour, and conclude that it contributes more to an assessment of a reflectance function choice than to the determination of solution surfaces consistent with the correct reflectance function and image. Finally we suggest directions for future work, and provide some guidelines and caveats for the development of image analysis systems based on these ideas.

1.1 Introduction

Although there are a variety of sources of information available to the human visual system including stereo, color, motion, and shading, we get very clear impressions of the three-dimensional character of a scene from a single grey-level still picture, even of scenes or objects that are not recognized as previously having been seen. This suggests that there is enough information in monocular grey-level images without motion for the visual system to arrive at a three-dimensional interpretation that is very convincing. The visual system is not always correct or even unambiguous in its interpretations: a picture can be interpreted "correctly" as a flat surface with shading variations or a clear window onto a scene; much of the cosmetic industry is dedicated to shape "enhancement" through shading.

Operationally, we get information from our visual system about positions and surfaces allowing us to navigate and interact with the environment. We are also able to combine the sense of touch with images of an object, allowing us to recognize a handled object blindfolded even if we have previously only seen pictures of it. This suggests some common information structure concerning shapes and surfaces accessible by touch, vision, and coordination. It is not clear, however, how accurately the visual system estimates shape considered as the exact location in space of each point on a surface or as the exact orientation of the surface at each point. It is difficult to create psychophysical experiments to test this because it is difficult to quantitatively probe a subject's internal information about surface shape. There

have been a few experiments along these lines recently (Mingolla and Todd, 1986, Bülthoff and Mallot, 1987) which suggest that our impressions of surface orientation are far more qualitative than we might like to believe. Nonetheless, we are interested in how much information about shape is contained in an image of a surface; answers to questions like these at least provide upper bounds on what the visual system can do in the absence of other information or assumptions.

One approach to studying what we get out of an image through vision is to try to understand what it is possible to discover from an image under different assumptions. Without any assumptions about lighting conditions or surface properties, there is no hope for recovering any information about surfaces in space: one can take a nearly arbitrary smooth surface and paint it to give the same image as another smooth surface without paint. The human visual system is apparently capable of using more than one set of assumptions: consider again the dichotomy between a picture as shaded surface and as window onto a scene. At the same time, it is difficult for us to entertain a continuum of possible interpretations: without extra cues, it is hard for us to interpret a flat picture of a scene as a different, curved, carefully painted surface. Some of Escher's delightfully bewildering works take advantage of such confusions in the visual system (Figure 1.1).

1.2 Background

Berthold Horn's early work on the shape from shading problem (Horn, 1975) examined it as a problem of physics, looking at the process of image formation and how light is reflected from objects and concentrated to form images. He defined a summary function, the reflectance function, that contained all the relevant local information about lighting conditions and surface reflecting properties under the assumption that reflecting properties of a surface patch were dependent solely on the orientation of the surface, and were constant with rotations of the surface around its normal. With additional assumptions of no cast shadows and no mutual illuminations, the brightness of a point in an image depends on the location of the point

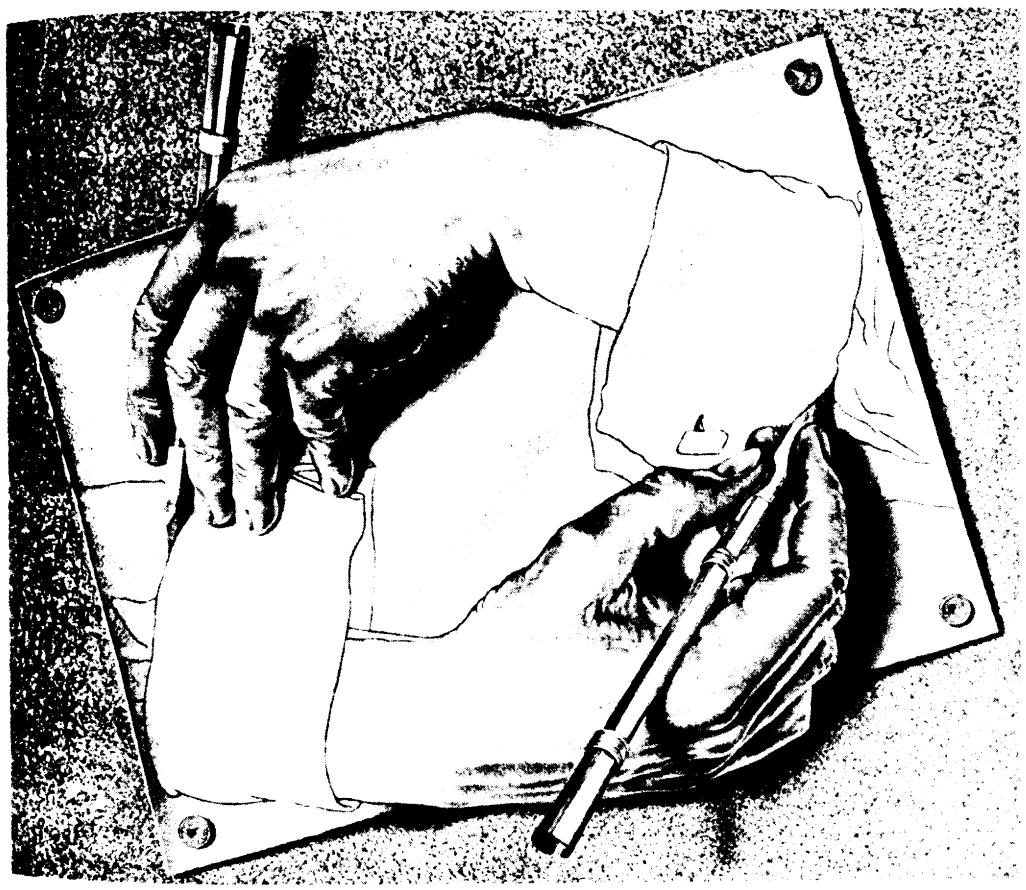


Figure 1.1. "Drawing Hands" by M. C. Escher (lithograph, 1948).

in space, the orientation of the surface in space at that point, and the reflectance function for the lighting conditions and surface material.

Assuming a known reflectance function, Horn was able to characterize the shape from shading problem as a nonlinear first order partial differential equation, the image irradiance equation. Classically, such an equation is solved by the method of characteristics, and Horn used this technique to develop a method for solving for the surface given some initial curve lying on the surface with known surface normals. Essentially, this is a Cauchy problem, and the solution proceeds along adjacent characteristic curves beginning at the known curve of data. These curves together with the surface orientations for the solution surface along them are known as characteristic strips.

Horn used tools from partial differential equations: calculations were generally done in a particular coordinate system, and all derivatives were partial derivatives with respect to the coordinates of that system. Horn assumed orthographic projection onto the image surface, and used orthogonal coordinates x, y, z with z-axis in the direction of assumed parallel projection. He used "gradient space" to represent surface normal directions: the normal to the surface $z = f(x, y)$ in the (x, y, z) coordinate system is given by $(p, q, -1)$, where $p = f_x$, and $q = f_y$. Unfortunately, these coordinates for the surface normal become unbounded when the visual boundary, or bounding contour, of an object is reached, where the tangent plane of the surface becomes parallel to the projection direction. This was later handled by using a different coordinate system based on stereographic projection of the unit sphere (Ikeuchi and Horn, 1981) at the cost of increased complexity of the coordinate expressions.

Horn recognized the importance of the critical points in the image. Given the reflectance function, these could be taken as points in the image with known orientation. As he noted, the characteristic trajectories could not be used to draw out the surface from the critical point, so he constructed a spherical cap consistent with the critical point orientation and used a small closed contour on this cap as his initial condition curve.

Direct integration of the characteristic equations to find the characteristic strips suffers from noise sensitivity in practical implementations. As the solution proceeds along constructed curves from the initial condition curve, these curves can deviate as a result of quantization error and other noise influences. Horn and Brooks (Horn and Brooks, 1986) review and compare a number of related methods by different researchers for making the solution of the shape from shading problem a global one, allowing data from the full image to contribute to finding stable, relatively robust solution surfaces. They provide a recipe for generating shape from shading methods. These are relaxation methods based on minimizing a certain measure, typically the integral over the image region of some combination of the error in the image irradiance

equation and a penalty term for departure from smoothness. In some cases, these methods have been derived using gradient space coordinates; in others, stereographic coordinates. Horn and Brooks derive methods based on direct calculations with the unit normal vector to the surface as well.

Various smoothness, integrability, or regularization terms have been tried by various researchers. Horn and Brooks indicate that enforcing the correct integrability constraint, $p_y - q_x = 0$ in gradient space coordinates, does not immediately yield a convergent relaxation scheme. However, as Frankot and Chellappa (Frankot and Chellappa, 1987) point out, using a different penalty term instead may yield a non-integrable set of surface normals. The resulting set of normals may be a smoothly chosen set of unit vectors, but may not be surface normals for any possible two-dimensional surface.

Pentland takes a different approach (Pentland, 1982, 1984). Rather than assuming full knowledge of the reflectance function, which is not available for human vision, he makes slightly less restrictive assumptions about the reflectance function (e.g. it is Lambertian, but with unspecified direction), and makes more assumptions about the surface structure. Pentland shows that if one assumes that the surface at a point is spherical, i.e. can be fitted by a spherical patch, then Lambertian reflectance gives a unique solution for the surface at that point.

As one might expect and as Pentland points out, not all local image intensity patterns can be accounted for by a spherical patch. If x, y are image coordinates and I is the image intensity, then, according to Pentland, patches where $I_{xx}/I_{yy} < 0$ cannot be fit by spherical surfaces. If we consider the image as a graph of image intensities over the image coordinates, such patches correspond to saddles in the graph considered as a two–dimensional surface. A solution can still be found where the principal curvatures have equal magnitude, but it is no longer a unique solution.

Pentland recognizes that some spherical patch solutions for certain kinds of image data are less likely to be reasonable than others. If one of the second derivatives of

the image intensity is zero, for example, the spherical patch solution involves very special placement of the light source or the surface patch. A more likely solution would have unequal principal curvatures with one of the principal curvatures zero. In a man-made environment of cones, planes and other ruled surfaces, these points would form regions, but it is not clear that the visual system evolved to be specially adapted to these unusual surfaces. On a general smooth surface, these points will occur on one-dimensional curves and so are sparse. [1]

One problem with this approach is that it is an extremely local analysis. In general, the only non-planar surfaces composed completely of equal curvature points (i.e. umbilic points) are pieces of spheres. Although one may be able to generate a spherical solution patch locally consistent with the image for many points in an image, it is not clear that the surface normals of these patches will be able to form a smooth surface. Nor is it clear that such a surface would have much to do with the original surface imaged. To compensate for the lack of generality of this surface structure assumption, Pentland is led to statistical methods to estimate surface orientation based on correlations in natural images: heuristic estimators for surface orientation are proposed.

Pentland's tools are again those of partial differential equations. A coordinate system is chosen, and derivatives are partial derivatives taken with respect to these coordinates. Pentland in places changes to a coordinate system parallel to the principal curvature directions, and indicates the change in coordinates with matrices when needed. To simplify the expressions, he assumes the surfaces of interest are exactly second order, an approach which may lose significant contributions to surface derivatives from higher order terms.

[1] It must be said that the natural environment in which we evolved is not smooth either. Recently, Pentland (Pentland, 1986) has suggested making use of fractal models of surfaces to help understand how a visual system might handle images of rough or textured images. To do such analyses carefully requires even more geometrical sophistication than the smooth approach: problems of differentiability and definition loom large. Perhaps because of our own visual experience which seems to give us smooth models of surfaces in our world in spite of its true complicated nature, most theoretical vision research has shied away from grappling with the complexities involved.

Frankot and Chellappa (Frankot and Chellappa, 1987) suggest a way of taking a set of non-integrable surface normals and producing an integrable set by projecting the surface normals onto the nearest set of integrable normals, where "nearest" is defined by a global integral distance measure defined in the gradient space coordinate system (and dependent on that coordinate system). Together with a method derived from Horn and Brooks, 1986, they construct a more quickly convergent algorithm guaranteed to form an integral set of normal vectors, and suggest their method could be applied to Pentland's normals to generate an integrable surface. Here too all the development is done in a single coordinate system, the gradient space coordinate system.

Pentland suggests classifying points based on the image derivatives (Pentland, 1984). Haralick and co-workers (Haralick et al, 1983) have carried this out more fully, looking at the surface formed by the graph of image intensities versus image coordinates. They define a set of features—peaks, pits, ridges, ravines, saddles, flats, and hillsides—in terms of the first and second derivatives of the image intensity that can describe each point on the image surface. Hillsides are defined as points that are not any of the other features, and will be the label that most points in an image receive.

Haralick et al consider the notion of invariance to various transformations as important. They point out that one of the advantages of this set of labels is that the labeling is independent of monotonic transformations of image intensity values: pits stay pits, and so forth. In addition the features are defined in terms of the gradient of the image intensity and the second derivative of the image intensity, both of which can be considered as tensors with invariant definitions with respect to changes in the image coordinate systems. Although hillsides could be further classified, these subclassifications do not remain invariant under monotonic image brightness transformations.

Haralick et al show a procedure for deriving these labels from an image using cubic polynomial patches fitted to the image and then analyzed for features. Pong

et al (Pong et al, 1985) begin analytical work on this representation by applying it to synthetic images of simple solids, calculating theoretically where features should lie on the image and comparing this to the features derived directly from the image.

Koenderink and van Doorn (Koenderink and van Doorn, 1980) have looked at the field of isophotes, or constant brightness contours, of an image and discussed some of their geometric features. They note that the Weingarten map, which maps the surface in space to the Gaussian sphere of unit surface normals, maps constant brightness contours on the surface to level sets of the reflectance function on the Gaussian sphere. In order to make sense of the infinity of possible arrangements of constant brightness contours potentially created by a given reflectance function, they restrict themselves to Weingarten maps that are "stable," or generic, meaning their fundamental geometry is not changed by small perturbations. This still leaves the wide class of Weingarten maps that are mostly one to one, with one-dimensional curves of points that are fold singularities of the Weingarten map, and isolated points that are cusps of the Weingarten map.

They classify various regions of such a surface using parabolic lines (folds in the Weingarten map), and discuss different causes for critical points in the image: specularities, certain points on the parabolic lines, and critical points of the reflectance function itself. These latter two cases will be discussed in more detail in Chapter 5.

Blicher (Blicher, 1983, Blicher, 1985) has used some of the modern language of differential topology to discuss stereo matching and other correspondance problems and their limitations as well as examining invariant features of a grey-level image. Chen and Penna (Chen and Penna, 1986) have used some language from modern differential geometry to state certain ideas about the study of photometric stereo and non-rigid objects.

Several difficulties and issues are common to the motivation and execution of shape from shading methods. One problem is the attention focused on representation. In order to test methods for deriving shape from shading, some coordinate

system must be chosen in which to represent various parameters of the surface. The computer must work with numbers, and numbers for surface calculations come from coordinate systems. The researcher typically uses this same coordinate system to derive and reason about methods. A single coordinate system is typically not able to cover an entire surface, and so there are problems with orientations that happen to fall outside the span of the major coordinate system. Algebraic tractability becomes a major factor in looking at results: with some coordinate systems, certain expressions become very complicated and are discarded as infeasible.

Modern differential geometry and global analysis provide tools for studying surfaces directly and independently of coordinate systems. This is not to say coordinate systems are not useful; rather, much of the geometrical background may be defined using properties intrinsic to the surfaces involved, and a coordinate system chosen to exploit other particular features of the problem at hand.

Enforcement of integrability is another issue common to different shape from shading methods. Most of the methods struggle to ensure that the surface normals produced are from a true two-dimensional surface, recognizing that not all collections of surface normals can be considered as coming from a surface. Frankot and Chellappa (Frankot and Chellappa, 1987) make the most explicit separation between the two groups of surface normal distributions, those that are integrable and those that are not, projecting one group onto the other. In other methods, some "integrability constraint" is frequently added in as a "regularization" term that is supposed to roughly deal with getting a smooth surface as a solution. Unfortunately, solutions derived without strictly enforcing integrability may still not be integrable.

As we shall discuss in Chapter 5, our results suggest integrability of a solution surface containing a critical point comes from the fact that it is an invariant smooth manifold containing the critical point. The algorithms proposed in Chapter 5 based on the image dynamical system attempt to find these invariant manifolds directly. Integrability of the corresponding surface normals comes for free in the noise-free case,

and can be used to monitor the progress of the algorithm in locating a reasonable solution when the image is noisy.

The question of uniqueness of solution for the different methods has also proved difficult. One would like some knowledge about how many interpretations of an image are possible given the information in the scene. Bruss (Bruss, 1980) found a uniqueness result for Horn's method in the case of a reflectance function of a particular form: essentially those with level surfaces that are concentric ellipses in the gradient space coordinate system. Many reflectance functions do not have this form, however. Deift and Sylvester (Deift and Sylvester, 1981) investigated in some detail uniqueness results for the degenerate case of an image of a Lambertian hemisphere lit from the viewing direction. They found different classes of non-spherical, even non-symmetrical, local solutions which are C^2 almost everywhere. If the solution surface is required to be C^2 everywhere, the expected spherical solutions are unique. They used methods of functional analysis, working in a polar coordinate system to analyze this specific case, and did not address questions about stability of the unusual solutions. General results on uniqueness and properties of solutions are lacking from the literature. Brooks (Brooks, 1982) discusses the general problem of ambiguity of solutions surfaces in images, and shows families of solutions for certain degenerate cases, e.g. a plane and hemisphere lit from the viewing direction. He also briefly examines the relationship between uniqueness of solution and the kind of image patch in the case of a hemisphere: certain patches of the image provide much more constraint than others.

Our results in Chapter 5 indicate that generically there will be at most four and at least two smooth solution surfaces through a patch of a smooth image containing a certain type of critical point due to a known reflectance function. These solutions correspond to the different invariant manifolds of the critical point seen as a critical point of the image dynamical system: the stable and unstable manifolds on which the image dynamical system has source and sink behavior, and potentially two other

invariant manifolds on which the image dynamical system has saddle behavior. The critical points in the image for which this result holds are those due to local reflectance function maxima (the usual case) or minima; we call these very good critical points of the image.

The stability of scene interpretation to variations in assumptions has not been fully explored. Ikeuchi and Horn (Ikeuchi and Horn, 1981) did a number of experimental tests of the performance of their shape from shading algorithm under violations of the assumed conditions, but this seems rare in the published literature.

In Chapter 5 we test our algorithms with noise and against errors in the reflectance function by looking at the effect of incorrect light source directions. In Chapter 7 we suggest some routes for future theoretical exploration of stablity of solution surfaces under variations in assumptions. In Section 5.3 we discuss the fundamental instability of the global image dynamical system: generically, the invariant manifolds of various critical points in a dynamical system do not flow smoothly together; however, an image due to a physical surface and a known reflectance function generates an image dynamical system which has the physical surface as an invariant manifold through all the critical points of the system. This creates a potential problem for computational solutions, since non-generic dynamical systems can be very difficult to simulate. However, this also suggests a potential answer to the problem of determining the reflectance function (or a number of parameters of it): the wrong choice of reflectance function is almost certain to prevent invariant manifolds from flowing smoothly into one another, and so may be rejectable on these grounds.

1.3 Overview of the Thesis

In the rest of this thesis, we explore in more detail the first order partial differential equation used by Horn for the shape from shading problem. As indicated above, one can classically solve such a problem by solving a related set of ordinary differential equations, the characteristic equations. For a typical space-invariant reflectance function such as the Lambertian, the one-dimensional curves which solve

the characteristic equations are trajectories through (x, y, p, q) space, where p and q are gradient space coordinates giving the orientation of the surface. These are the characteristic trajectories or strips.

In Chapters 2 and 3 the shape from shading problem and the image irradiance equation are placed in a coordinate-independent and geometrically inspired setting. Coordinate independent definitions of the bounding contour (edges of the image due to the surface rolling away from the viewer) and the behavior of the image at the bounding contour are discussed. In Chapter 4 we show a somewhat dauntingly algebraic but coordinate-independent approach to deriving the characteristic equations.

A modern approach to the study of nonlinear ordinary differential equations like the characteristic equations considers them as a dynamical system. A dynamical system is essentially a vector field on some space; trajectories of the dynamical system are parameterized paths in the space that have derivatives equal to the vector field everywhere along them. For example, if the vector field represents the velocity at each point of fluid flowing in a pipe, the trajectories will be the time-dependent paths of points flowing along with the fluid. We can speak of the flow of a vector field as the collection of all the trajectories together. In Chapter 4 we consider the characteristic equations of the shape from shading problem as a vector field defining an image dynamical system.

In Chapter 5 we examine critical points of the image and the image dynamical system. The qualitative study of dynamical systems emphasizes the role of critical points, places where the vector field is zero, in determining the overall structure of the flow lines. If the critical point is not "degenerate" in some sense, then the behavior of the nonlinear dynamical system will be quite close to the behavior of a linear approximation to the dynamical system near the critical point. It turns out that critical points of the image intensities due to critical points in the reflectance function give rise to critical points in the image dynamical system.

Another important tool in the modern study of dynamical systems discussed in Chapter 5 is the idea of invariant surfaces for the system. For any point on such a

surface, the trajectory of that point (perhaps for a small time interval) also lies on the surface. Smooth solution surfaces for the shape from shading problem must be invariant surfaces of the image dynamical system: they are two-dimensional surfaces made up entirely of segments of characteristic trajectories.

The two ideas of critical point and invariant surface come together quite literally at a critical point. The set of all points whose trajectories end up at a particular critical point as time goes to infinity form an invariant manifold called the stable manifold of that critical point; the set of points which end up at the critical point as time goes to negative infinity form an invariant manifold called the unstable manifold. Both the stable and unstable manifolds contain the critical point. There may also be other invariant manifolds containing the critical point. For hyperbolic critical points (defined in Chapter 5), for example those generically caused by maxima and minima of the reflectance function, near the critical point all these invariant manifolds are quite close to the invariant manifolds of the linearized dynamical system—in particular, the number of invariant manifolds and their tangents at the critical points are determined by the linearized system, as well as the type of flow restricted to the invariant manifold. As discussed in Chapter 5, in the shape from shading problem we can use the fact that there are usually only a finite number of invariant manifolds around very good critical points to get existence and uniqueness results on patches of the image around the critical points.

In Section 5.3 we discuss how we can use the dynamical system itself to help find certain of the invariant manifolds around a critical point. Essentially we float an infinitely distensible surface in the flow of the vector field. If we are assured that some point on this surface will flow to the critical point (i.e. the initial surface cuts the stable manifold), and the rest of the points on the initial surface are allowed to follow the trajectories through them, then the surface will be stretched and deformed over time to approximate the unstable manifold near the critical point.

In Section 5.3 we give a couple of examples of highly parallel algorithms based on this idea, and analyze their performance in the presence of noise and mistaken

reflectance functions by implementing the algorithms on a 16K CM-1 Connection Machine. The methods turn out to be stable in the presence of noise, and robust in handling errors in the reflectance function.

Chapter 6 discusses the role of the bounding contour in constraining shape from shading solutions. The results suggest that the bounding contour does not provide "patch" constraints of the kind given by critical points, but does provide constraints on the reflectance function.

Finally, Chapter 7 summarizes the results, and discusses future extensions, including caveats and suggestions for implementing a system to connect the patches of solutions generated by the local analyses made here.

We begin looking at the shape from shading problem with global analytic tools by discussing the image projection map in the next chapter.

Chapter 1

Introduction and Background

Chapter 2

Image Projection

In this chapter we look at how the image projection takes points in space and maps them to points in an image, and some of the consequences of this mapping. We begin with a discussion of some of the foundations of modern differential geometry, the concept of a manifold and a tangent bundle, and follow with an analysis of the image projection map, the bounding contour, and suggest possible extensions to time varying images.

2.1 Mathematical Preliminaries

2.1.1 Manifolds

In this section a few of the basic tools and ideas from modern differential geometry are presented. A good introduction to the modern approach to differential geometry can be found in (Spivak, 1979); a brisker introductory approach with connections to dynamical systems and other areas of modern physics can be found in (Abraham, Marsden, and Ratiu, 1983). The reader is referred to either of these for the technical details omitted here.

In both classical and modern differential geometry, the notion of coordinate invariance plays an important role. In the classical view still adopted by many physics textbooks, a formula that maintains its character after a coordinate change is in some sense an intrinsically defined object: if we replace a coordinate system (x, y, z) with a

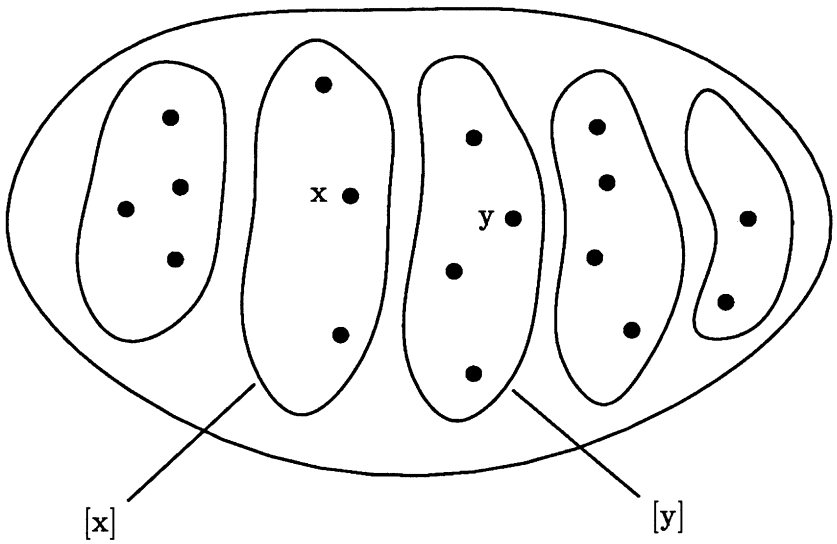


Figure 2.1. Equivalence classes as the partition of a set.

coordinate system (x', y', z'), a formula written with coordinates x, y, z becomes a formula in the functions $x(x', y', z'), y(x', y', z'), z(x', y', z')$; if, after simplification, the resulting formula looks exactly the same as the original, replacing the symbols x, y, z with x', y', z', then the formula is considered invariant to this coordinate change.

2.1.1.1 Equivalence Class Approach

A modern way to approach coordinate invariance is to use equivalence classes. A set X is partitioned into equivalence classes when it can be divided up into a group of non-overlapping sets, X_i, such that $\cup X_i = X$, $X_i \cap X_j = \emptyset$ if $i \neq j$ (Figure 2.1). This gives rise to a number of useful properties. Note that every element of X is in one of the equivalence classes; we can refer to the set containing x as $[x]$. If $[x] = [y]$, and $[y] = [z]$, then we must have $[x] = [z]$, since the two equivalence classes $[x]$ and $[z]$ share an element, y, so their intersection is non-empty; since the partition is non-overlapping, this must mean $[x] = [z]$. Clearly, if $[x] = [y]$ then $[y] = [x]$; also $[x] = [x]$.

These last three properties are those of an equivalence relation: an equivalence relation on a set X is a predicate $\sim$ acting on two elements of the set with the

properties that $x \sim x$ is true; $x \sim y$ true implies $y \sim x$; and finally $x \sim y$ and $y \sim z$ imply $x \sim z$. Equivalence classes and equivalence relations are essentially the same idea: given an equivalence relation, we can define the set $[x]$ to be the set of all $y \in X$ such that $y \sim x$. The equivalence relation properties make the collection of sets $\{[x]\}$ so defined into a collection of equivalence classes on X. Similarly, given an equivalence class partition of X we can define the equivalence relation $x \sim y$ as true if and only if $[x] = [y]$, and can refer to the collection of equivalence classes as $X/\sim$ or as $\tilde{X}$.

Why are equivalence classes useful? They become very useful when properties or operations on the individual elements give rise to properties or operations on the equivalence classes considered as elements of the collection of equivalence classes. Essentially, the properties and operations that stick around after moving to equivalence classes can be considered as invariant under the equivalence relation defined by the equivalence classes. Consider a function $f : X \to X$. [1] If we can define a map $\tilde{f} : \tilde{X} \to \tilde{X}$ such that the functional diagram

$$\begin{array}{ccc} \tilde{X} & \xrightarrow{\tilde{f}} & \tilde{X} \\ \uparrow{\scriptstyle \pi} & & \uparrow{\scriptstyle \pi} \\ X & \xrightarrow{f} & X \end{array}$$

commutes,[2] where $\pi : X \to \tilde{X}$ is just the equivalence class map $x \mapsto [x]$, then in some intuitive way we can consider f as "really" acting on the equivalence classes of X. The equivalence classes provide a way of looking at some of the behavior of f, how it moves elements between equivalence classes, while ignoring other parts, how it moves elements within an equivalence class. An extreme example might be if

[1] The notation $f : A \to B$ means that f takes elements of the set A and gives values in the set B; thus, for $a \in A$ we have $f(a) \in B$.

[2] A diagram like that above defines composition functions between the different domains. For example, π is a map from X to $\tilde{X}$, $\tilde{f}$ is a map from $\tilde{X}$ to itself, and $\tilde{f} \circ \pi$ defines a map from X to $\tilde{X}$ whose value at p is $\tilde{f}(\pi(p))$. The diagram is said to commute when following two sets of arrows from the same starting set to the same finishing set gives the same function: i.e., $\tilde{f} \circ \pi = \pi \circ f$.

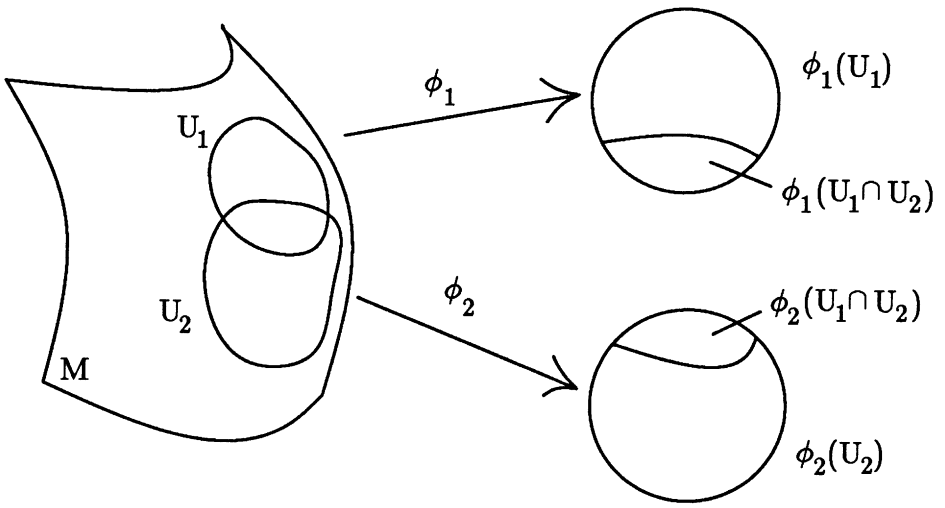


Figure 2.2. The modern manifold: charts of an atlas. ϕ_1 and ϕ_2 are charts mapping open sets U_1 and U_2 on M into $\mathbb{R}^2$.

$\tilde{f} = id$, that is $f([x]) = [x]$ for all $[x] \in \tilde{X}$. This means that f does not move elements of X outside of their equivalence classes under $\sim$ at all; in this case the equivalence classes of X are actually invariant sets of f.

2.1.1.2 Manifolds

The modern definition of a surface, or manifold, essentially relies on equivalence class definitions to get away from individual coordinate representations. First let us look at features of the modern definition of a manifold.

The idea is to model the local behavior of the surface by coordinate charts (maps from the manifold to $\mathbb{R}^n$) which behave "reasonably" on pieces of the surface that are acted on by more than one chart (Figure 2.2). We start by requiring the manifold M to have an atlas of charts, meaning that there is a set of open neighborhoods U_α of M that cover M and a corresponding set of invertible maps $\phi_\alpha : M \to \mathbb{R}^n$ mapping each patch onto $\mathbb{R}^n$.[3] These coordinate patches and coordinate maps give us local windows onto the surface. One analogy is a collection of television cameras covering

[3] This is for an n-dimensional manifold. One can use other kinds of linear spaces as so-called model spaces in place of $\mathbb{R}^n$ to define infinite dimensional manifolds. This can be used to make the calculus of variations into a true calculus with derivatives, etc. (Marsden and Hughes, 1983)

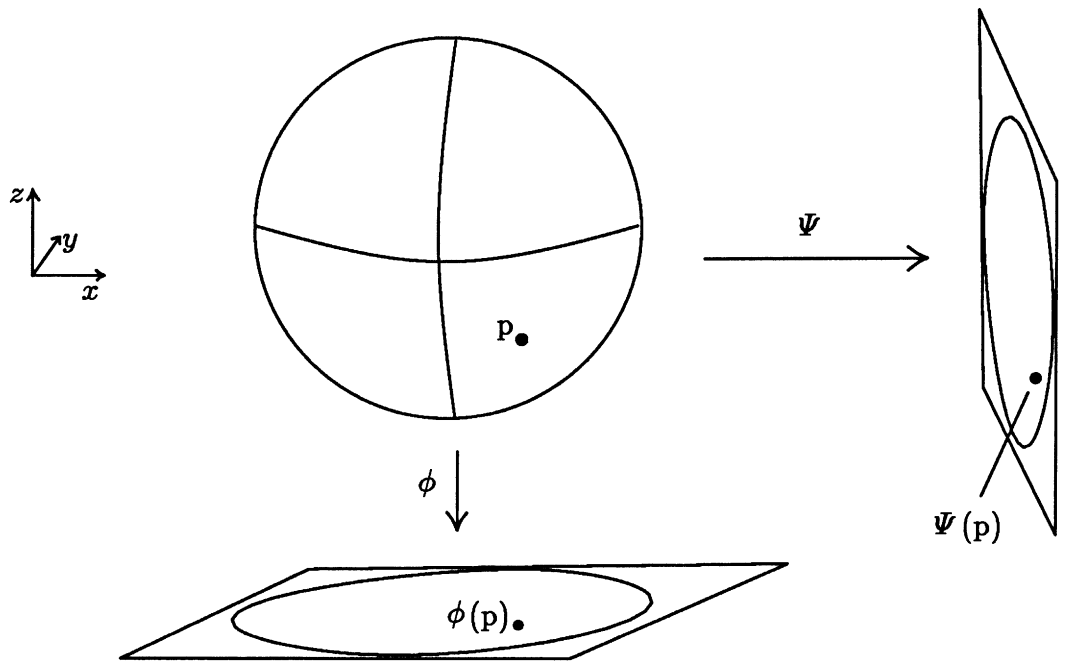


Figure 2.3. The sphere and two charts, ϕ and ψ.

a sporting event: when you look at the wall of television images at the studio, you are seeing different views of the same event.[4] We do need some assurance of consistency between the images: someone might have changed a channel. In the mathematical realm, we can require the charts to be consistent on their overlap: if we have $\phi_1 : U_1 \to \mathbf{IR}^n$ and $\phi_2 : U_2 \to \mathbf{IR}^n$ as coordinate charts on M, we can define a new map $\phi_2 \circ \phi_1^{-1} : V_1 \to V_2$, where $V_1 = \phi_1(U_1 \cap U_2)$ and $V_2 = \phi_2(U_1 \cap U_2)$ are both sets in $\mathbf{IR}^n$. For consistency, we can require this overlap map between two real spaces to be continuous with continuous inverse, or C^k with C^k inverse, or smooth (C^∞) with smooth inverse—whatever degree of smoothness we are interested in using to define and work with the manifold.

As an example, consider the two dimensional sphere embedded in $\mathbf{IR}^3$ (Figure 2.3). One coordinate chart we might use is the map $\phi : \{(x, y, z) \in S^2 | z > 0\} \to \mathbf{IR}^2$ given by $\phi(x, y, z) = (x, y)$: this is orthogonal projection onto the x–y plane. A point

[4] Marshall McLuhan would approve of our approach: our perception of an event is often defined these days by the different views provided by different television cameras.

on the sphere is given by $(x, y, \sqrt{1-x^2-y^2}) \in \mathbb{R}^3$ within the domain of this chart; in fact, $\phi^{-1}(x,y) = (x, y, \sqrt{1-x^2-y^2})$ since $\phi \circ \phi^{-1} = id$ using these definitions.

We could also take the chart $\psi : \{(x,y,z) \in S^2 | x > 0\} \to \mathbb{R}^2$ given by $\psi(x,y,z) = (y,z)$. Are ψ and ϕ compatible on their overlap? We look at $\psi \circ \phi^{-1}$ on the appropriate domain of $\mathbb{R}^2$: we have

$$\begin{aligned} \psi \circ \phi^{-1}(r,s) &= \psi(r, s, \sqrt{1-r^2-s^2}) \\ &= (s, \sqrt{1-r^2-s^2}). \end{aligned}$$

Is the map $(r,s) \mapsto (s, \sqrt{1-r^2-s^2})$ smooth? Here we have to be careful to include the domain on which we are working: we are interested in the smoothness of this map not on all of $\mathbb{R}^2$, but only on $\phi(U_1 \cap U_2)$ where $U_1 = \{(x,y,z) \in S^2 | z > 0\}$ and $U_2 = \{(x,y,z) \in S^2 | x > 0\}$. We have $\phi(U_1 \cap U_2) = \{(r,s) | r > 0 \ , r^2 + s^2 < 1\}$. We can now look at the smoothness of the overlap map $(r,s) \mapsto (s, \sqrt{1-r^2-s^2})$ on the overlap: the Jacobian of this map is

$$\begin{bmatrix} 0 & 1 \\ -r/\sqrt{1-r^2-s^2} & -s/\sqrt{1-r^2-s^2} \end{bmatrix},$$

and indeed for $r > 0$, $r^2 + s^2 < 1$, this Jacobian is smooth and is smoothly invertible; hence the overlap map is smooth and smoothly invertible, and the coordinate charts are smoothly compatible.

Given one such atlas of charts for a manifold, we can now throw into the atlas all other charts that are compatible with it to create a maximal atlas. We include any other open neighborhood U on M with associated invertible map $\phi : U \to \mathbb{R}^n$ which is compatible with all the original charts. If we have two of these new charts, ϕ and ψ, they will then be compatible with each other: on appropriate subdomains, we have

$$\psi \circ \phi^{-1} = (\psi \circ \phi_i^{-1}) \circ (\phi_i \circ \phi^{-1})$$

for each i, where ϕ_i are the original charts for the manifold; since the original atlas has neighborhoods covering the whole manifold, we can decompose the intersection

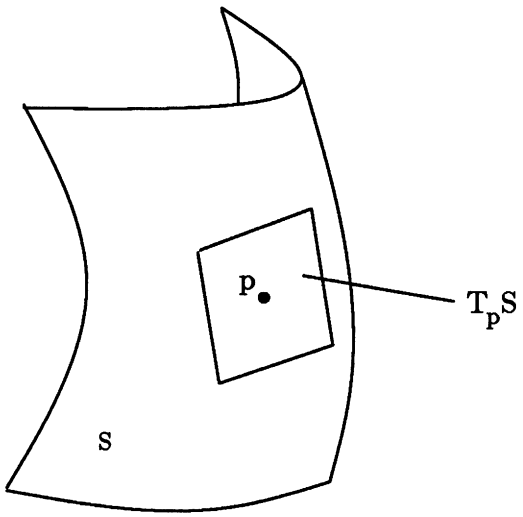


Figure 2.4. The geometric tangent space T_pS is a plane in $\mathbb{R}^3$ tangent to the surface at p.

of the domains of ϕ and ψ completely into patches within different U_i. Since ψ and ϕ are compatible with all the ϕ_i by definition, and the compositions above have the same degree of smoothness, we have ψ and ϕ also compatible.

A manifold is defined as a set of points with an atlas of charts compatible on their overlaps. The modern way to work with a manifold M is to always be thinking about this potentially quite abstract object while working with useful views of it provided by the coordinate charts in an atlas. The same manifold M may look quite different using different coordinate charts. Formulas in the two coordinate systems may look very different but refer to the same operations or points on the manifold. For example consider a plane parameterized using polar coordinates (r, θ) or cartesian coordinates (x, y). An arc of a circle can be given by $r = c$ or by $y = \sqrt{c^2 - x^2}$, depending on the coordinate system. The choice of a particular coordinate system can be made solely to simplify a particular calculation involving the arc. The arc is a geometric object independent of the charts used to view it.

2.1.2 Tangent Spaces

We have defined a manifold intrinsically, that is without requiring the manifold to sit "inside" a larger space.[5] We can define the tangent space at a point on the manifold intrinsically as well. At first glance this seems unreasonable: our intuition about what a tangent space is typically comes from the embedded picture of a two-dimensional manifold in three dimensions (Figure 2.4): we think of it as a plane just tangent to a point on the surface, and this plane clearly sits in the full three-dimensional space. However, using equivalence classes we can dispense with the embedding space and still maintain the idea of the tangent space as a kind of linear approximation to the original surface.

As with many good ideas in mathematics, there is more than one way to define the tangent space intrinsically, and all of these are equivalent. One can define it as equivalence classes of curves on the surface M, or as equivalence classes of derivative operators on real-valued functions of the manifold M, for example. Here we will discuss another approach tied more closely to the coordinate charts themselves.

Let us consider two charts in an (assumed) smooth atlas for a manifold M, ϕ_1 and ϕ_2, with a non-empty overlap on M. The overlap map $\beta_{12} \stackrel{\Delta}{=} \phi_2 \circ \phi_1^{-1}$ on $\phi_1(U_1 \cap U_2)$ is smooth by assumption. The derivative of β_{12}, $D\beta_{12}$, at a point $\phi_1(p) \in \mathbb{R}^n$ defines a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$.[6] If we pick v_1 as a vector in $\mathbb{R}^n$, then we can consider $v_2 = D\beta_{12}(v_1)$ as related to v_1 through the overlap map β_{12}. We can make this into an equivalence relation in the following way: consider all ordered pairs (ϕ_i, v_i) of charts[7] and vectors in $\mathbb{R}^n$. We define the equivalence relation $\sim$ on this set as

[5] Whitney's theorem (Guilleman and Pollack, 1974) says that any manifold can be modeled as a submanifold of a higher dimensional space; it is not always conveniant or reasonable to do this, however; e.g, curved space-time does not necessarily sit "inside" some other larger "hyperspace".

[6] The differential Df_x of a function $f : \mathbb{R}^m \to \mathbb{R}^n$ at a point $x \in \mathbb{R}^m$ is the unique linear map such that $||f(x+h) - f(x) - Df_x(h)||$ is $O(||h^2||)$ for all $h \in \mathbb{R}^n$. See (Abraham, Marsden, and Ratiu, 1983) or the more introductory (Lang, 1968) for details: it is essentially the Jacobian without a particular coordinate choice.

[7] Including domains, although we don't write them for clarity of notation.

$(\phi, v) \sim (\psi, w)$ if and only if $w = D(\psi \circ \phi^{-1})v$ at $p \in M$; i.e. w and v are related by the derivative of the overlap map. To see that this is an equivalence relation, we have to verify the equivalence relation properties:

1) $(\phi, v) \sim (\phi, v)$: we have the simple overlap map $\phi \circ \phi^{-1} = id$, so $v = D(id)(v) = id(v)$.

2) $(\phi, v) \sim (\psi, w)$ implies $(\psi, w) \sim (\phi, v)$: from the first relation we have $w = D(\psi \circ \phi^{-1})(v)$. Using the inverse function theorem and the smooth invertibility of $\psi \circ \phi^{-1}$, this means that $D(\phi \circ \psi^{-1})w = v$, so that $(\psi, w) \sim (\phi, v)$.

3) If $(\phi_1, v_1) \sim (\phi_2, v_2)$ and $(\phi_2, v_2) \sim (\phi_3, v_3)$, then $(\phi_1, v_1) \sim (\phi_3, v_3)$: we have

$$v_2 = D(\phi_2 \circ \phi_1^{-1})v_1$$
$$v_3 = D(\phi_3 \circ \phi_2^{-1})v_2.$$

We can compose the overlap maps and use the chain rule to get

$$\begin{aligned} v_3 &= D(\phi_3 \circ \phi_2^{-1}) \circ D(\phi_2 \circ \phi_1^{-1})(v_1) \\ &= D(\phi_3 \circ \phi_1^{-1})(v_1), \end{aligned}$$

so $(\phi_1, v_1) \sim (\phi_3, v_3)$.

We can now define a tangent vector for M as an equivalence class $[(\phi, v)]$ under the above relation. We think of (ϕ, v) or just v as the coordinate representation by ϕ of the tangent vector. Why is this a reasonable thing to do? Again we can think about the charts as windows onto the real object we are looking at. If a tangent vector $\mathbf{v}$ of M is to be reasonably defined, we expect that the "snapshots" of $\mathbf{v}$ in different charts should be nicely related; what could be nicer than to have them related directly by the overlap maps that connect the different pictures?

We can see how this works by looking at tangent vectors to curves on an n dimensional surface embedded in $\mathbb{R}^m$ (Figure 2.5). A curve on the surface is given by a map $\alpha : \mathbb{R} \to M$ from an interval of the real line to the manifold. Since $M \subset \mathbb{R}^m$, α describes a curve in $\mathbb{R}^m$, and we can define its derivative at any

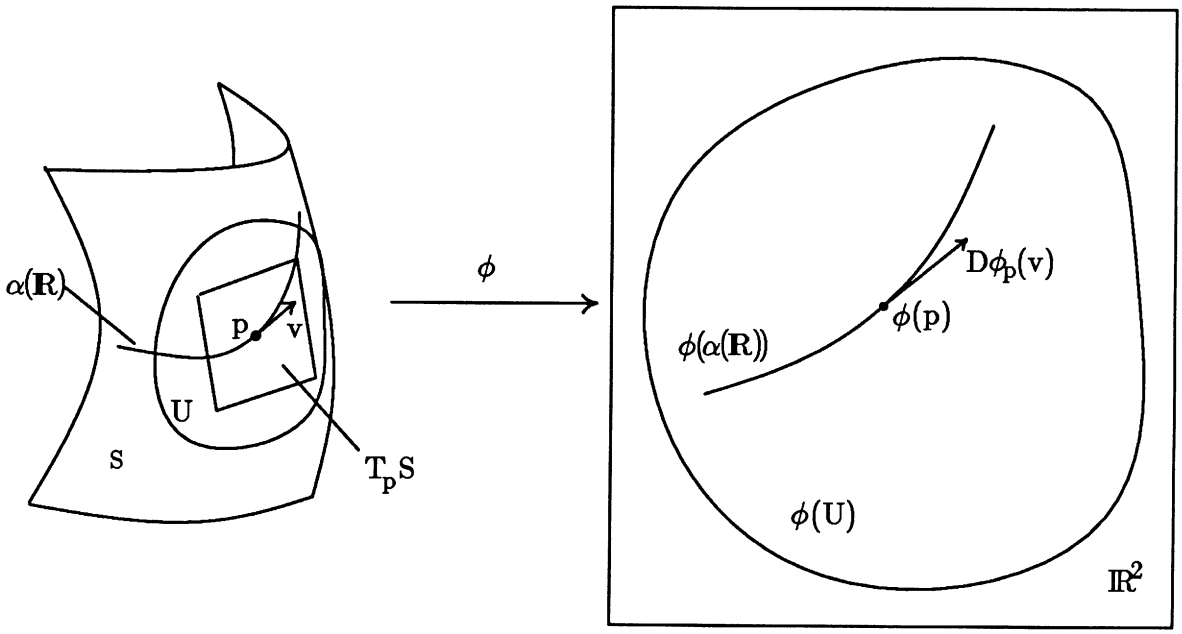


Figure 2.5. Connection between modern and geometric definition of tangent vectors. Here $\alpha(s)$ is a curve in S with tangent vector $\mathbf{v}$ at p. It is mapped by the chart ϕ to a curve in $\mathbb{R}^2$ with tangent vector

$$D\phi_p(\mathbf{v}) = (\partial/\partial t)\phi(\alpha(t))|_{t=0}.$$

The set of image vectors $D\phi_p(\mathbf{v})$ for all charts ϕ is the equivalence class definition of the tangent vector.

point along the curve in space by the usual time derivative $\alpha'(t)$.[8] Now let us use a coordinate chart ϕ for M containing a piece of the curve α: we can look at the image of the curve under the coordinate chart, $\phi \circ \alpha(t)$; this is a curve in the real space $\mathbb{R}^n$. The time derivative of this combined map gives a vector $\frac{\mathrm{d}}{\mathrm{d}t}(\phi \circ \alpha)(t)$ in $\mathbb{R}^n$ which can be thought of as the image of the tangent vector $\alpha'(t)$ under the map ϕ; indeed we would like to write

$$\frac{\mathrm{d}}{\mathrm{d}t}(\phi \circ \alpha)(t) = D\phi \circ \alpha'(t)$$

using the chain rule if we could make sense of $D\phi$ on the manifold M.

[8] Note that in terms of the derivative operation, D, $\alpha'(t) \triangleq D\alpha_t(1)$.

If we have another chart ψ on the same area of M, we have a different image of the tangent vector $\alpha'(t)$: we have $\frac{\mathrm{d}}{\mathrm{d}t}(\psi \circ \alpha)(t)$. How are these two projected coordinate vectors related? Let us look at $t = 0$. We define

$$\alpha_1(t) = \phi \circ \alpha(t)$$
$$\alpha_2(t) = \psi \circ \alpha(t)$$
$$v_1 = \alpha_1'(0)$$
$$v_2 = \alpha_2'(0).$$

We can use the overlap map to relate the two curves α_1 and α_2: we have

$$\psi \circ \phi^{-1} \circ \alpha_1(t) = \alpha_2(t)$$

(ignoring domain details), so taking the time derivative at $t = 0$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\psi \circ \phi^{-1} \circ \alpha_1)(0) = v_2$$
$$D(\psi \circ \phi^{-1})(v_1) = v_2.$$

Since $\psi \circ \phi^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ is a map between vector spaces, $D(\psi \circ \phi^{-1})$ makes sense, and we have used the usual real vector space chain rule on the composition of real vector space maps $(\psi \circ \phi^{-1}) \circ (\alpha_1)$. From the definition of the vector equivalence relation given earlier, we have $(\phi, v_1) \sim (\psi, v_2)$, or $[(\phi, v_1)] = [(\psi, v_2)]$. The members of the equivalence class $[(\phi, v)]$ are indeed views of the geometric tangent vector in different coordinate systems.[9]

From this we see that when the manifold is embedded in a vector space, the equivalence class definition of a tangent vector puts together all the coordinate images of a "true" tangent vector into one equivalence class. We avoid the need for an embedding space in general by defining a tangent vector to be the set of all possible related image vectors under all possible consistent charts where the relation is through the overlap map. This is a bit like defining a tennis player's stroke as all

[9] For technical details see (Spivak, 1979).

possible television images of it: we do not need to postulate the existence of Rod Laver in order to admire the stroke and analyze it in detail; all we need are all possible different views of it. Vision can be considered as giving us two time-varying charts of the two-dimensional surfaces around us; judging our reaction to animated artificial figures, the available views alone can be very powerful in determining our perception of reality.[10]

We can do more with the equivalence class definition of tangent vectors at a point on the manifold. The equivalence classes themselves form a vector space: for a a scalar we can define

$$a[(\phi, v)] \triangleq [(\phi, av)]$$
$$[(\phi, v)] + [(\psi, w)] \triangleq [(\phi, v + D(\phi \circ \psi^{-1})w)],$$

where the slightly complicated expression for the sum just reflects the need to move the representative vectors into the same chart before adding them; we have $[(\phi, v)] + [(\phi, w)] = [(\phi, v + w)]$ when the representative vectors are in the same coordinate chart. The chain rule and linearity of the derivative map on vector spaces can be used to show that these definitions make sense; for example, if $(\phi_1, v_1) \sim (\phi_2, v_2)$ then we know

$$v_2 = D(\phi_2 \circ \phi_1^{-1})(v_1),$$

so

$$av_2 = D(\phi_2 \circ \phi_1^{-1})(av_1),$$

where a is a scalar, and hence $(\phi_1, av_1) \sim (\phi_2, av_2)$ so that the definition of $a[(\phi, v)]$ does not depend on the choice of representative (ϕ, v) for the equivalence class. Essentially, the basic vector space structure of $\mathbf{IR}^n$ is preserved by the mapping from representatives to equivalence classes even though individual identities of representatives are (usefully) confused by the mapping.

[10] Jan Koenderink has recently proposed a representation scheme for the human visual system based on coordinate independent objects (Koenderink, 1988).

We will call the vector space of all equivalence classes of representatives the tangent space T_pM at p of M. Given a map $f : M \to N$ taking elements of the manifold M over to the manifold N, we can define a linear derivative map $Df_p : T_pM \to T_qM$ taking elements of the tangent space of M at p to elements of the tangent space of N at $q = f(p)$:

$$Df_p\mathbf{v} = Df_p([\phi, v]) \triangleq [\psi, D(\psi \circ f \circ \phi^{-1})(v)],$$

where ψ is a coordinate chart on N and ϕ is a coordinate chart on M. Essentially, we are "writing the map f in coordinates" when we look at $\psi \circ f \circ \phi^{-1}$: we use the chart ϕ to examine a piece of the domain and the chart ψ to examine a piece of the range. Note that $\psi \circ f \circ \phi^{-1}$ is a map from $\mathbb{R}^n$ to $\mathbb{R}^m$ (assuming N has dimension $\mathbb{R}^m$), and so the derivative of this real vector function is the standard vector space derivative. The definition of Df as an equivalence class mapping makes sense because the chain rule connects different representatives for the same equivalence classes: if $\tilde{\psi}$ is another coordinate chart on N, $\tilde{\phi}$ another coordinate chart on M, and $[(\tilde{\phi}, \tilde{v})] = [(\phi, v)]$, we have

$$\begin{aligned}[\psi, D(\psi \circ f \circ \phi^{-1})(v)] &= [\tilde{\psi}, D(\tilde{\psi} \circ \psi^{-1}) \circ D(\psi \circ f \circ \phi^{-1})(v)] \\ &= [\tilde{\psi}, D(\tilde{\psi} \circ f \circ \phi^{-1})(v)]\end{aligned}$$

using the definition of equivalence, and

$$\begin{aligned}[\psi, D(\psi \circ f \circ \phi^{-1})(v)] &= [\tilde{\psi}, D(\tilde{\psi} \circ f \circ \tilde{\phi}^{-1}) \circ D(\tilde{\phi} \circ \phi)(v)] \\ &= Df_p([(\tilde{\phi}, D(\tilde{\phi} \circ \phi)(v)] \\ &= Df_p([(\tilde{\phi}, \tilde{v}]),\end{aligned}$$

using the definition of Df_p and using the definition of equivalence between $[(\tilde{\phi}, \tilde{v})]$ and $[(\phi, v)]$. Thus the definition of Df_p makes sense. We can consider the derivative $D(\psi \circ f \circ \phi^{-1})$ as the coordinatized view of Df_p in just the way we considered $v \in \mathbb{R}^n$ as the coordinatized version of $\mathbf{v} = [(\phi, v)]$.

If we take the simplest case where M is the space $\mathbb{R}^n$, then the identity map on the open "subset" $\mathbb{R}^n$ of $\mathbb{R}^n$ makes a perfectly good chart. $T_p\mathbb{R}^n$ at a point $p \in \mathbb{R}^n$

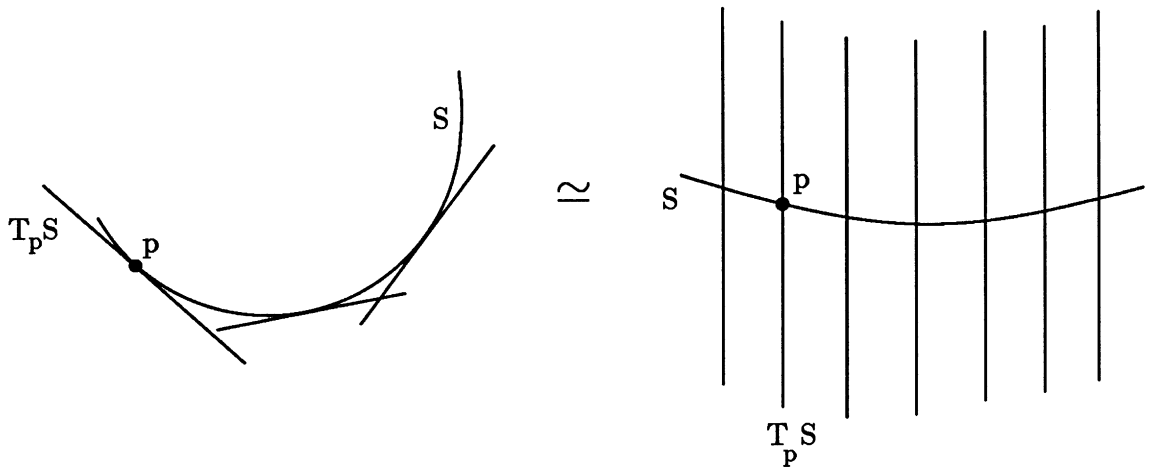


Figure 2.6. The tangent bundle to a curve: by turning the 1-d tangent spaces on their sides, we can see how $TS \simeq \mathbb{R} \times \mathbb{R}$.

can now be identified with $\mathbb{R}^n$ at each p, since we have a single global chart. Even if we add other charts to the atlas (e.g. polar coordinates or other things), we can still consider this as the canonical coordinate system, and in general we can identify n-vectors $\mathbb{R}^n$ as elements of $T_p\mathbb{R}^n$ at some $p \in \mathbb{R}$. The full collection of tangent vectors for all $p \in \mathbb{R}^3$, $T\mathbb{R}^3$, is equivalent to $\mathbb{R}^3 \times \mathbb{R}^3$: we consider a tangent vector $\mathbf{v}_p \in T_p\mathbb{R}^3$ to be just $(p, \mathbf{v}) \in \mathbb{R}^3 \times \mathbb{R}^3$.

On a general manifold M, given a coordinate chart $\phi : U \subseteq M \to \mathbb{R}^n$ there are certain vector fields often labeled as special. These are the vector fields whose representatives in $\mathbb{R}^n$ under the chart are the standard orthonormal basis vectors. If we label the components of the chart as $\phi(p) = (x^1(p), \ldots, x^n(p))$, then the associated vector fields are called $\partial/\partial x^i$. Using the idea of the derivative of a mapping between manifolds, we can define the vectors $\partial/\partial x_i$ by

$$D\phi_p\left(\frac{\partial}{\partial x^i}\right) = e_i,$$

where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ where the 1 appears in the i-th position.[11]

[11] The notation $\partial/\partial x^i$ comes about because if $f : M \to \mathbb{R}$ is a real function on the manifold M, then $Df\left(\partial/\partial x^i\right)$ in coordinates is given by $D(f \circ \phi)(e_i) = \partial f/\partial x_i$, where in the last expression f is assumed written in coordinates with respect to the chart ϕ. In fact, there is another definition of the tangent space that considers tangent vectors as real-valued linear operators on real-valued functions of the manifold. (Warner, 1971).

In general we can collect all the tangent spaces T_pM together and make the collection $TM = \cup T_pM$ itself into a manifold. To do this, we need to find an atlas for TM: we can construct one from an atlas for M. We take a chart for M, (U, ϕ) (remember U is the open set of M on which the chart map ϕ is defined), and consider the derivative map $D\phi_p$ for all p in the open set U: since $\phi : U \rightarrow \mathbf{IR}^n$ and both U and $\mathbf{IR}^n$ are manifolds, we can define $D\phi_p$ as an equivalence class mapping as we did above. We can use the map

$$\tilde{\phi}(\mathbf{v}_p) \triangleq (\phi(p), D\phi_p(\mathbf{v}_p)) : TU \rightarrow \mathbf{IR}^n \times \mathbf{IR}^n$$

as our new coordinate chart on the new open set $TU = \bigcup_{p \in U} T_pM$ of TM; note that we have identified $D\phi_p(\mathbf{v}_p) \in T_p\mathbf{IR}^n$ with the appropriate n-vector in $\mathbf{IR}^n$. An overlap map for this new chart would look like

$$\tilde{\beta}_{12}(x, v) = (\phi_2 \circ \phi_1^{-1}(x), D(\phi_2 \circ \phi_1^{-1})_x(v)),$$

which has the usual overlap map as the first component, and the derivative of the overlap map as the second component; assuming the charts are smooth, for example, this combined overlap map is also smooth.[12] Essentially, every tangent bundle of an m-dimensional manifold locally (i.e. viewed through a chart) looks equivalent to $\mathbf{IR}^m \times \mathbf{IR}^m$. Globally this is not the case: for example, there is a result which says there can be no completely non-zero vector field on the two-dimensional sphere (the so-called Hairy Ball Theorem (Abraham, Marsden, and Ratiu, 1983)); this is certainly not true of $\mathbf{IR}^2$, and so TS^2 and $\mathbf{IR}^2 \times \mathbf{IR}^2$ are not equivalent, even though pieces of TS^2 are equivalent to pieces of $\mathbf{IR}^2 \times \mathbf{IR}^2$.

Basically, we can reason about the tangent vectors defined as equivalence classes as if they were the familiar tangent vectors in space on which our intuition is based.

[12] One can define the tangent space in yet another way through overlap maps alone, considering them as "glue" between patches of $\mathbf{IR}^n$.

Modern differential geometry provides a bag of tools and techniques defined intrinsically (via the equivalence class definitions of tangent space, etc.). One reason for using these tools is to ensure that constructions do not depend on a particular coordinate system; instead, they can be thought of as part of the intrinsic geometry of a situation. In practice, in the early examination of a physical problem it is very hard to reason usefully without eventually picking coordinates. From a pragmatic point of view, the intrinsic definitions of the tools and constructions of differential geometry and global analysis means that they are available even if we decide to pick bizarre coordinates to exploit special features of the problem under consideration.

2.2 The Image Projection Map

As suggested by the previous pictures used to help explain some of the ideas of modern differential geometry, vision and geometry have very close ties. It might be said that vision provides us with a pair of time-varying charts of the surfaces around us, together with brightness information at each point of the image. Particular coordinate systems do not have such a close relationship with vision: we are not aware of some pre-existing coordinate system in reasoning about surfaces or interpreting an image. It seems to make sense to try to use some of the invariant language and notation of the modern approach to describe features of the image formation and interpretation process. Often coordinate representations of things can look complicated and messy even though the underlying geometric idea may not be so complicated. If we begin with invariant definitions of some of the fundamental ideas, we can move to various coordinates suited for particular analyses of details with (hopefully) a minimum of confusion.

Let us begin by looking at the image projection map. A projection system can be thought of as a way of mapping points in $\mathbb{R}^3$ to points on some two dimensional imaging surface I. [13] (See Figure 2.7.) We can call this map $\pi^I : \mathbb{R}^3 \rightarrow I$, and will

[13] Really we map from some large open set of $\mathbb{R}^3$. We will mostly ignore this technical detail. (Blicher, 1985) tries to keep track of all the different open sets in his definitions, and it can be confusing.

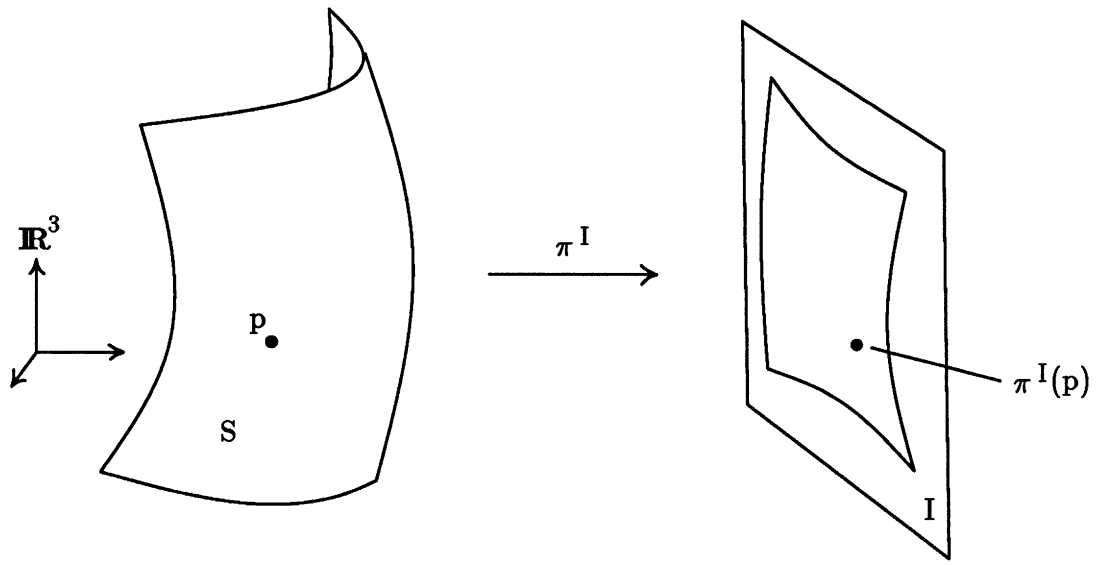


Figure 2.7. Image projection from $\mathbb{R}^3$ to the image I.

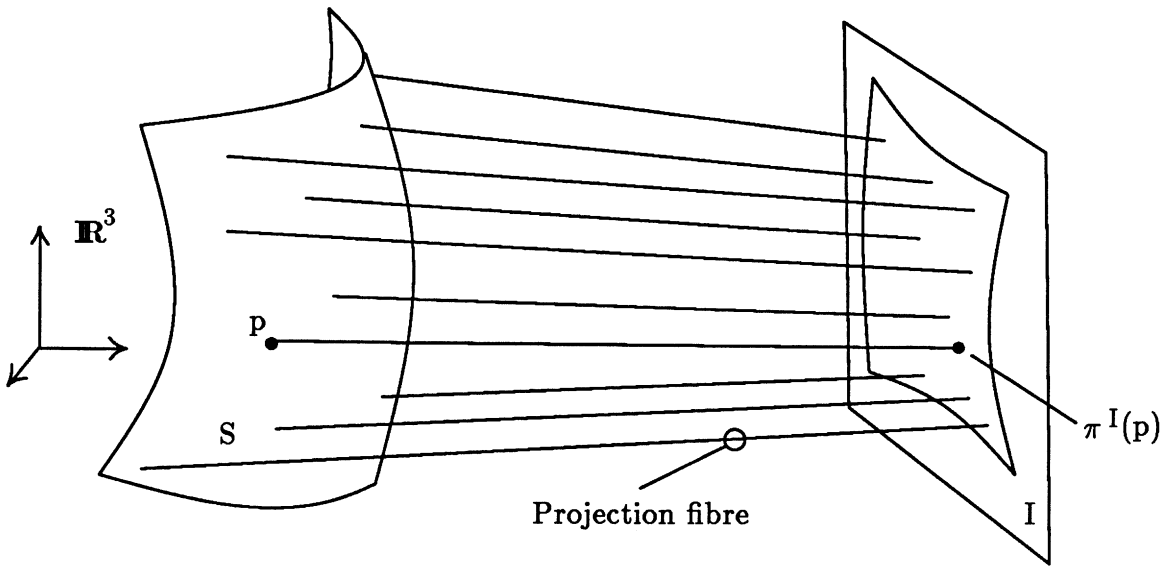


Figure 2.8. Image projection as a fibre map

assume it is a smooth map. Typical projection maps used are orthogonal projection onto a plane, central projection onto a plane, or central projection onto a sphere centered at the projection center. Although central projection onto a sphere perhaps best models the physical projection of light onto the retina while central projection onto a plane best models projection of light onto film in a camera, it is not clear what

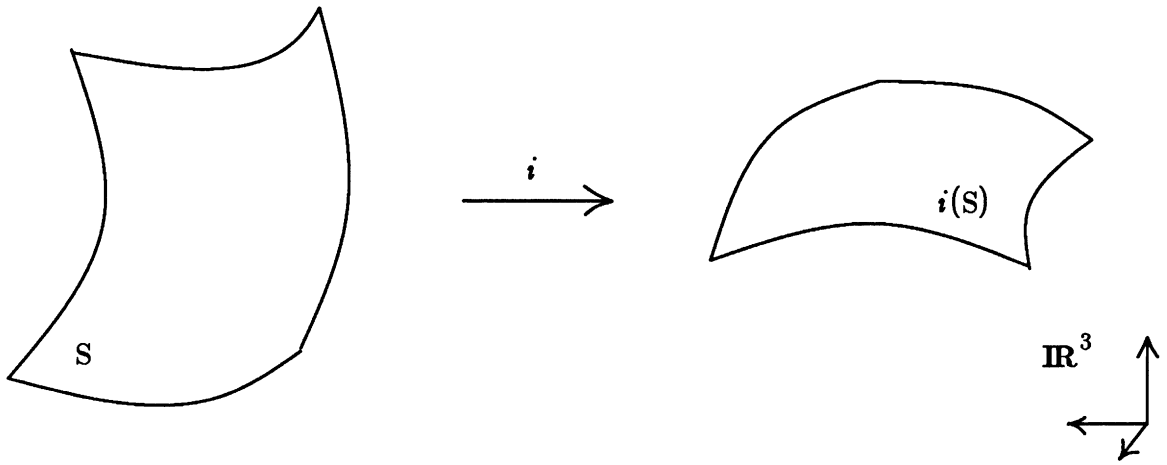


Figure 2.9. Embedding a two dimensional surface in $\mathbb{R}^3$.

the "right" projection for vision should be: do we attempt to include the inhomogeneities of the retinal neural distribution in the projection by centrally projecting onto some distorted surface? Ideally we would like results that are insensitive to variations in the projection system; we will try to work in that direction in what follows. [14]

The distinction between images made at a large distance (compared to the focal distance) under central projection to a plane or orthogonal (scaled) projection to a plane is quite small, so orthogonal projection is often used because its coordinate representation is much simpler. We will do the same as soon as the going gets rough.

[14] One abstract way of viewing projection is to eliminate the imaging surface altogether. We can usually imagine the projection situation to be as follows: we have a collection of non-intersecting rays running from $\mathbb{R}^3$ to the imaging surface. The point at which a ray hits the imaging surface defines the image projection of all the points on the ray. (See Figure 2.8) The projection can be thought of as a projection from $\mathbb{R}^3$ to the set of rays; the imaging surface just provides us with a convenient way to coordinatize the set of rays, since each point on the imaging surface corresponds to a unique projection ray. This perspective suggests that the shape of the imaging surface is of secondary importance in understanding the image projection: given the projection of $\mathbb{R}^3$ onto the projection rays, or fibres, we can generate the values of any image projection formed by intersecting the bundle of rays with an arbitrary surface. Note that this is not true of brightness values at each point on the image: the brightness of a point on the image does depend on the orientation of the imaging surface.

A two dimensional surface S in $\mathbb{R}^3$ can be thought of as a two-dimensional manifold N embedded in $\mathbb{R}^3$ by a map $i : N \to \mathbb{R}^3$.[15] In the case of a non-moving scene or a still photograph, we can consider i to be the inclusion map, so that $N = S$, and $i(p) = p$. However, in the case of a moving scene, we may want to model the motion as a time varying embedding of some "fixed" surface N, so that we have $i : N \times \mathbb{R} \to \mathbb{R}^3$, with $i_t(N) = S_t$ giving the position of the embedded surface in space at time t. For various purposes, we might try to specify what i can be: e.g. an isometry, a rigid rotation, or a general embedding. We can speak about the velocity field in space of each point $p \in N$ as $(\mathrm{d}/\mathrm{d}t)i_t(p)$.

As a two-dimensional manifold on its own, N has an intrinsic tangent bundle TN. This is related to the usual tangent planes in $\mathbb{R}^3$ of the embedded surface $S = i(N)$ by the embedding map: $Di : TN \to T\mathbb{R}^3$ takes the two dimensional tangent space T_pN to the two-dimensional tangent space $T_{i(p)}i(N) = T_{i(p)}S$. One way to see this is to think of paths on the surfaces: a path $\alpha : \mathbb{R} \to N$ becomes a path $i \circ \alpha : \mathbb{R} \to \mathbb{R}^3$ in space on the embedded surface S. Di then carries the tangent vector $\alpha'(t)$ to the appropriate tangent vector on the surface embedded in $\mathbb{R}^3$.

When we view a surface in $\mathbb{R}^3$, we can consider how the image projection, π^I, maps the surface to the image. If we consider the two-dimensional surface N as embedded in $\mathbb{R}^3$ by $i : N \longrightarrow \mathbb{R}^3$, with $i(N) = S$, then we can consider the map $\pi^I \circ i$. This is the map that takes points on the surface, N, to points on the image, I.[16]

---

[15] An immersion is often defined as a map from a lower dimensional space to an equal or higher dimensional space with the derivative having maximal rank everywhere; an embedding is an immersion with the additional property that distinct pieces of the immersed surface are always well separated. Technical details can be found in (Abraham, Marsden, and Ratiu, 1983).

[16] Another notation for this is $i^*\pi^I \triangleq \pi^I \circ i$ which is called the "pull-back" of π^I by i because i^* pulls the projection map $\pi^I : \mathbb{R}^3 \to I$ back to the surface N.

If we allow time-varying surface embeddings, the image of a point on the imaging surface (assuming a time invariant projection system) is $\pi^I \circ i_t(p)$, and so the motion field of the image is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}(\pi^I \circ i_t(p)) = D\pi^I \circ \left(\frac{\mathrm{d}}{\mathrm{d}t} i_t(p)\right);$$

thus, $D\pi^I$ plays the crucial role of mapping velocity vectors in space to motion field vectors in the image.

Our interest here will mostly be with still images, so that in general we can assume that the embedding i is the inclusion map: N is a surface in $\mathbb{R}^3$, and $i(p) = p$, so $N = S$ and we can interchangeably refer to $i(S)$ and S.

2.2.1 The Bounding Contour

We can use the general image projection map to define the bounding contour of a surface and image. We have a smooth map $\pi^I \circ i : N \to I$ from the two-dimensional surface to the two-dimensional image. The image consists of an interior where two-dimensional patches of the surface are connected to two-dimensional patches of the image, and a bounding contour where the relationship between the surface and the image must be more complicated. It seems reasonable that most of the interior of the image corresponds to a region of the surface where the map $\pi^I \circ i$ is a diffeomorphism; at pieces of the bounding contour (which may include curves in the interior of the image), this is not the case.

To see where $\pi^I \circ i$ is a diffeomorphism, we assume i and π^I are smooth (as we shall do for the duration). Using the inverse function theorem,[17] we know that $\pi^I \circ i$ is a diffeomorphism on a neighborhood of $p \in N$ if $D(\pi^I \circ i)$ is invertible at p.

[17] As with many results in modern differential geometry, the inverse function theorem for manifolds is analagous to the inverse function theorem for real vector functions, and can be proved by coordinatizing the manifolds and applying the real result. Knowing theorems in the real case usually gives the right intuition for the manifold case.

How likely is it that $D(\pi^I \circ i)$ is invertible? Sard's theorem says that almost all points of $\pi^I(S)$ will be regular values for $\pi^I \circ i$, meaning $D(\pi^I \circ i)$ will have full rank and so will be invertible. What can happen to make $D(\pi^I \circ i)$ not invertible? We can use the chain rule to take this derivative apart: $D(\pi^I \circ i) = D\pi^I \circ Di$. By assumption, i is an embedding so that Di always has maximal rank, i.e. rank two. Similarly, by assumption we can insist that the image projection π^I have full rank at all points of its domain in $\mathbf{IR}^3$, again rank two. The only way that the composition of the two linear maps $D\pi^I$ and Di can have rank less than two is if the range of Di intersects non-trivially the null space[18] of $D\pi^I$; i.e. if there is a non-zero vector $\mathbf{u} \in T_pN$ such that for $\mathbf{v} = Di_p(\mathbf{u})$ we have $D\pi^I{}_{i(p)}(\mathbf{v}) = 0$.[19]

We can interpret this in another way using projection fibres (Figure 2.10). A projection fibre is a set of points in space all of which project to the same image point. If we consider a curve $\alpha : I \rightarrow \mathbf{IR}^3$ running along a projection fibre, so that $\pi^I(\alpha(t)) = \pi^I(\alpha(0))$, we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}(\pi^I \circ \alpha)(t) = D\pi^I \circ \alpha'(t).$$

Since $D\pi^I$ has rank two at each point of $\mathbf{IR}^3$, we can consider a unit vector $k(\alpha(t))$ parallel to $\alpha'(t)$ as determining the one-dimensional null space for $D\pi^I$ at each point $p = \alpha(t)$. We can do this smoothly over the whole projection domain using a flow down all the fibres simultaneously. At each point $p \in \mathbf{IR}^3$, $k(p)$ essentially represents the local projection direction; in the usual case of projection rays, $k(p)$ will actually point towards (or away from) the true image point. If $\mathbf{v} = Di_p(\mathbf{u})$ is in the null space of $D\pi^I$, as required for $\pi^I \circ i$ to not be a diffeomorphism, then we must have $\mathbf{v}$ parallel to $k(i(p))$. Put another way, in this case the local projection direction is

[18] The null space of a linear map A consists of all vectors v such that $Av = 0$.

[19] One of the reasons for having tangent spaces around is to make liberal use of linear algebra results on the tangent space at each point of the manifold. The tangent bundle acts like a theoretical Connection Machine for linear algebra.

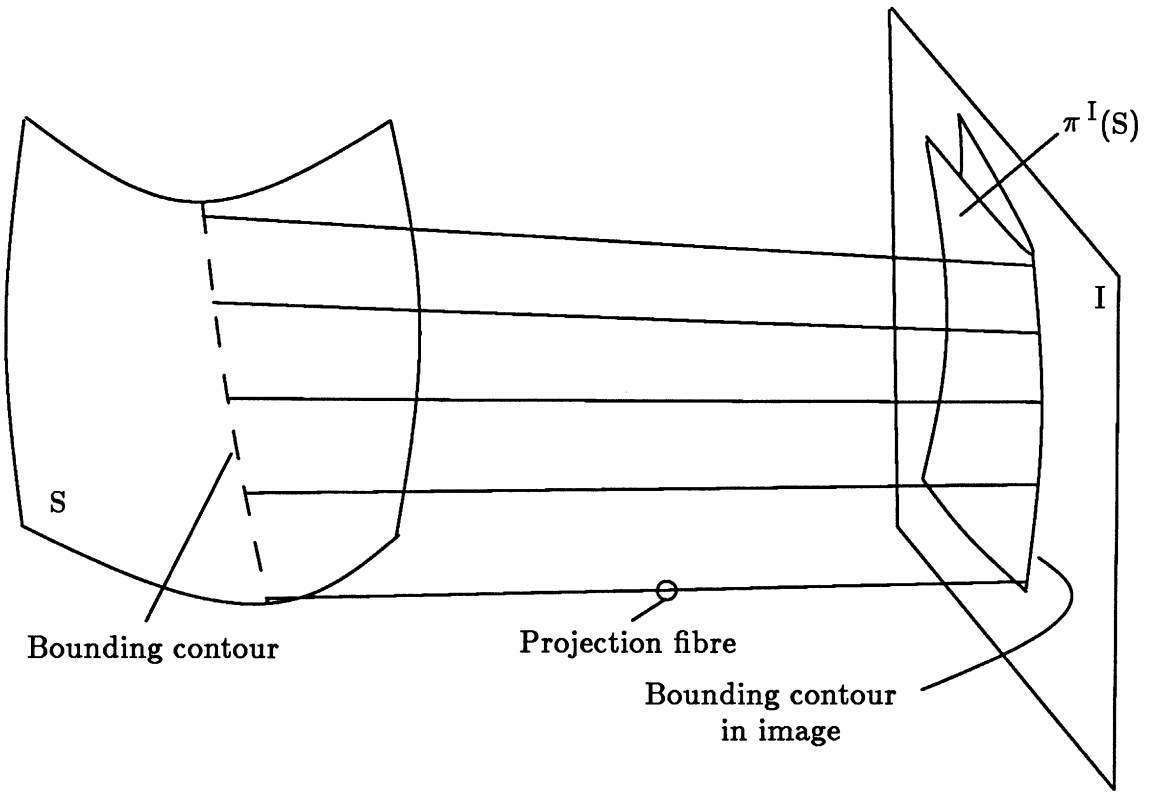


Figure 2.10. The bounding contour and projection fibres. The projection fibres are tangent to the surface at the bounding contour.

parallel to the tangent plane of the surface in $\mathbb{R}^3$; this is what is usually meant by the bounding contour, also called the self-occlusion locus, or extremal contour. At such points, the projection direction just grazes the surface in space (see Figure 2.10).

We now have invariant characterizations of the surface in $\mathbb{R}^3$, the image projection from $\mathbb{R}^3$ to the image I, and the bounding contour of the projection. In the next chapter we discuss the reflectance function and an invariant version of the image irradiance equation.

Chapter 3

The Image Irradiance Equation

In this chapter we will discuss what the image irradiance equation looks like when written in invariant language. Assuming fixed lighting conditions and uniformity of surface material, Horn (Horn, 1975) modeled the shading of an image of a surface as dependent on the locations and local orientations of points on the surface. This suggests that the surface tangent planes are important objects for understanding shape from shading.

We begin the mathematical preliminaries section with a discussion of the space $\mathcal{C}(\mathbb{R}^3, 2)$ of all two-dimensional tangent planes in $\mathbb{R}^3$ and a useful class of coordinate charts for it; $\mathcal{C}(\mathbb{R}^3, 2)$ is the space of surface orientations on which the reflectance map acts. We discuss how two-dimensional surfaces embedded in $\mathbb{R}^3$ can be "lifted" to be two-dimensional surfaces in the space $\mathcal{C}(\mathbb{R}^3, 2)$. We discuss contact 1-forms, which are linear functionals on the space $\mathcal{C}(\mathbb{R}^3, 2)$ useful for detecting when a two-dimensional surface in $\mathcal{C}(\mathbb{R}^3, 2)$ is lifted from $\mathbb{R}^3$. In Section 3.2 we discuss the reflectance map and the image formation process. We write the invariant image irradiance equation, and show that it is the same as Horn's equation given a certain set of coordinates. Finally, we state an invariant version of the shape from shading problem. In the next chapter, we tackle how this invariant description of the image irradiance equation can be used to generate an invariant vector field on $\mathcal{C}(\mathbb{R}^3, 2)$; this is essentially the classical characteristic strip method used by Horn.

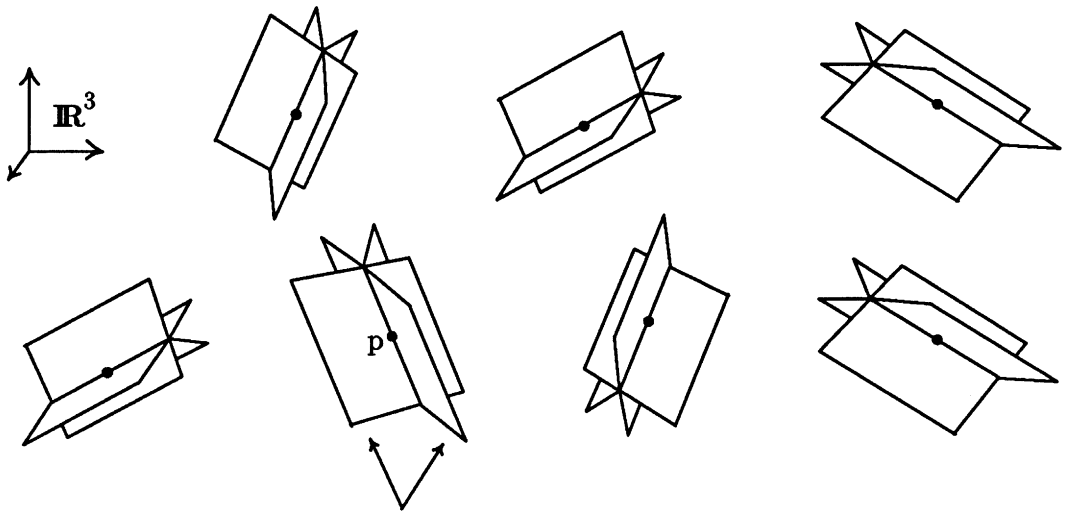


Figure 3.1. Set of all tangent planes in space, $\mathcal{C}(\mathbb{R}^3, 2)$. At each point there is a collection of 2-dimensional planes, each of which could be a tangent space to a 2-d surface.

3.1 Mathematical Preliminaries

3.1.1 $\mathcal{C}(\mathbb{R}^3, 2)$

Consider the surface, S, embedded in $\mathbb{R}^3$ by the embedding $i : S \longrightarrow \mathbb{R}^3$. As discussed in the last chapter, associated with each point p of S there is a tangent plane to S denoted T_pS. This can be considered as a two-dimensional subspace of the three-dimensional tangent space $T_p\mathbb{R}^3$. The orientation of the surface S at the point p is determined by the tangent plane.

To discuss orientations of surfaces, it is useful to consider the set $\mathcal{C}(\mathbb{R}^3, 2)$ of all possible two-dimensional tangent planes in $\mathbb{R}^3$. As discussed in the last chapter, the tangent bundle for $\mathbb{R}^3$, $T\mathbb{R}^3$, is defined as the collection of all tangent vectors to $\mathbb{R}^3$. $T\mathbb{R}^3$ is equivalent to the Cartesian product of $\mathbb{R}^3$ with itself, $\mathbb{R}^3 \times \mathbb{R}^3$, so that a tangent vector $\mathbf{v}_p \in T\mathbb{R}^3$ is equivalent to a point $(p, \mathbf{v}) \in \mathbb{R}^3 \times \mathbb{R}^3$. We can think of $p \times \mathbb{R}^3$ as a vector space in its own right isomorphic to $\mathbb{R}^3$ by doing vector operations only on the second component. A tangent plane at p is then equivalent to a two-dimensional subspace $p \times W$ of $p \times \mathbb{R}^3$, where W is a

two-dimensional subspace of $\mathbb{R}^3$. If we define $\mathcal{G}(2,3)$ as the set of two dimensional subspaces of $\mathbb{R}^3$, we have $\mathcal{C}(\mathbb{R}^3,2)$ equivalent to the Cartesian product of $\mathbb{R}^3$ with $\mathcal{G}(2,3)$, $\mathcal{C}(\mathbb{R}^3,2) \simeq \mathbb{R}^3 \times \mathcal{G}(2,3)$. $\mathcal{G}(2,3)$ is itself a manifold, where each subspace is considered a point of the set; it can be given coordinates by picking a coordinate chart for the Gaussian sphere of normal vectors, and then identifying two-dimensional subspaces of $\mathbb{R}^3$ with their normal vectors.[1]

3.1.2 $\mathcal{C}(\mathbb{R}^3,2)$ and a useful coordinate system

Since both $\mathcal{G}(2,3)$ and $\mathbb{R}^3$ are manifolds, the manifold $\mathcal{C}(\mathbb{R}^3,2) \simeq \mathbb{R}^3 \times \mathcal{G}(2,3)$ is as well. If (V_1, ϕ_1) is a chart on $\mathcal{G}(2,3)$, and (V_2, ϕ_2) is a chart on $\mathbb{R}^3$, then we can create a chart $(V_1 \times V_2, \phi_1 \times \phi_2)$ for $\mathcal{C}(\mathbb{R}^3,2) \simeq \mathbb{R}^3 \times \mathcal{G}(2,3)$. The usual (x,y,z,p,q) gradient coordinate chart based on rectilinear coordinates for $\mathbb{R}^3$ can be thought of this way: it splits into the chart (x,y,z) on $\mathbb{R}^3$ and a chart (p,q) for orientations, where the (p,q) chart does not depend on the point (x,y,z) in space to determine the coordinates for orientation.

We can generalize the (x,y,z,p,q) coordinates for $\mathcal{C}(\mathbb{R}^3,2)$ to include nonlinear coordinates on $\mathbb{R}^3$. Consider (V,x) a coordinate chart for $\mathbb{R}^3$, so that V is an open set of $\mathbb{R}^3$ and x is a chart map defined on V. Pick one of the coordinates, say x^3, as "special." We will construct a coordinate chart for $\mathcal{G}(2,3)$ at each p in V. We have tangent vectors $\frac{\partial}{\partial x^1}\big|_p$, $\frac{\partial}{\partial x^2}\big|_p$, $\frac{\partial}{\partial x^3}\big|_p$ to $\mathbb{R}^3$ at each point p defined using the coordinate chart x. These span $T_p\mathbb{R}^3$. We will define a linear map ${L_p}^a$ which connects a pair of numbers, $a = (a_1, a_2)$, to a two-dimensional subspace of $T_p\mathbb{R}^3$ by mapping elements of the subspace spanned by $\left\{ \frac{\partial}{\partial x^1}\big|_p, \frac{\partial}{\partial x^2}\big|_p \right\}$ to elements of a two-dimensional subspace uniquely determined by a. We define

$${L_p}^a : \text{span}\left\{ \frac{\partial}{\partial x^1}\bigg|_p, \frac{\partial}{\partial x^2}\bigg|_p \right\} \longrightarrow T_p\mathbb{R}^3$$

[1] A manifold made up of linear subspaces of a vector space is called a Grassman manifold. (Abraham, Marsden, and Ratiu, 1983)

in coordinates by

$$L_p{}^a \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ a_1 & a_2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix},$$

where (v_1, v_2) are the coordinates of a vector $\mathbf{v} \in \text{Span}\left\{ \frac{\partial}{\partial x^1}\Big|_p, \frac{\partial}{\partial x^2}\Big|_p \right\}$ with respect to the basis vectors $\frac{\partial}{\partial x^1}\Big|_p$ and $\frac{\partial}{\partial x^2}\Big|_p$, and the coordinates of the range vector in $T_p\mathbb{R}^3$ are given with respect to the basis $\frac{\partial}{\partial x^1}\Big|_p, \frac{\partial}{\partial x^2}\Big|_p, \frac{\partial}{\partial x^3}\Big|_p$. Each choice of $a = (a_1, a_2)$ gives a different linear map $L_p{}^a$, and the range of $L_p{}^a$ is always two dimensional.[2]

We can consider the map $L_p{}^a$ as defining a map $\mathcal{L} : \mathbb{R}^3 \times \mathbb{R}^2 \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$ given by $\mathcal{L} : (p, a) \longmapsto \text{Range}(L_p{}^a)$. If W_p is a two-dimensional subspace of $T_p\mathbb{R}^3$ such that $\partial/\partial x^3$ is not a member of W_p, then there is an $a \in \mathbb{R}^2$ such $\mathcal{L}(p, a) = W_p$. [3] If a and $\tilde{a}$ are two vectors in $\mathbb{R}^2$ with $a \neq \tilde{a}$, then $\mathcal{L}(p, a) \neq \mathcal{L}(p, \tilde{a})$, since the matrix columns of the two $L_p{}^a$ span different subspaces.

---

[2] As often happens, there is a trade-off between writing things invariantly and writing things in coordinates; to write this map invariantly requires the use of inner products with the basis vectors, etc, and would not add materially to our discussion.

[3] To see this, pick two vectors in W_p that span it. Using x coordinates, put these column vectors together into a 3×2 matrix, $\mathbf{A}$, so that

$$\mathbf{A} = \begin{pmatrix} \mathbf{C} \\ \mathbf{d} \end{pmatrix},$$

where $\mathbf{C}$ is 2×2 and $\mathbf{d}$ is 1×2. We have $\text{Range}(\mathbf{A}) = W_p$ in coordinates. $\mathbf{C}$ must have rank 2; if not, then the columns of $\mathbf{C}$ are dependent, and there is a 2-vector $\mathbf{v} = (v_1, v_2)^T$ such that

$$\mathbf{A} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ e \end{pmatrix},$$

implying that $\partial/\partial x^3$ is in the range of $\mathbf{A}$, contrary to our assumption. Since $\mathbf{C}$ has rank 2, there are independent 2-vectors $\mathbf{u}$ and $\mathbf{v}$ such that

$$\mathbf{A}(\mathbf{u}, \mathbf{v}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ a_1 & a_2 \end{pmatrix}:$$

multiplying $\mathbf{A}$ on the right by a column vector takes linear combinations of the columns of $\mathbf{A}$ and hence of the columns of $\mathbf{C}$, so we can pick $\mathbf{u}$ and $\mathbf{v}$ to give this combination of the columns of $\mathbf{C}$ since $\mathbf{C}$ is invertible. The vectors $(1, 0, a_1)^T$ and $(0, 1, a_2)^T$ span the same subspace as the columns of $\mathbf{A}$ do, namely W_p. Defining $a = (a_1, a_2)$, we have

$$\mathcal{L}(p, a) = \text{Range}(\mathbf{A}) = W_p.$$

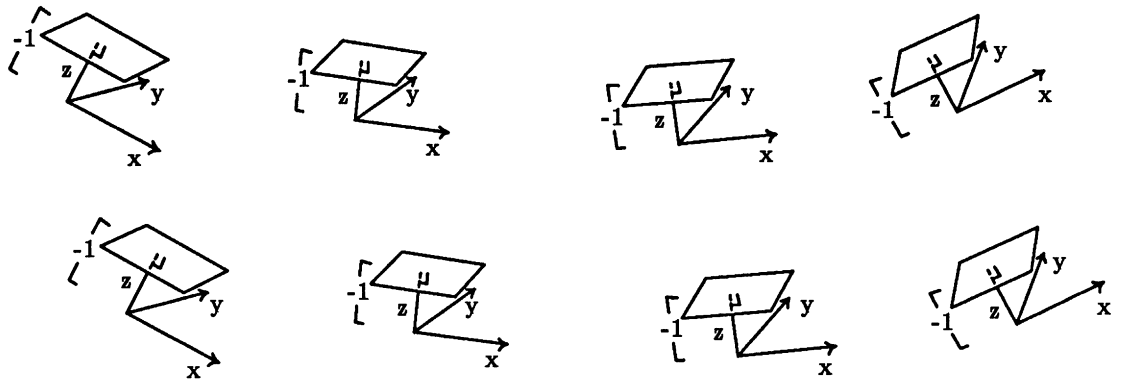


Figure 3.2. Local view of generalized (x, y, z, p, q) coordinates: planes at $(0, 0, -1)$ perpendicular to local $\partial/\partial z$ direction. The local (p, q) coordinates for the direction of a vector $\mathbf{v}$ at p is given by the coordinates of the intersection of the line parallel to $\mathbf{v}$ through p with the appropriate plane.

Given the base chart (V, x) we can define a local chart $(U, (x, a))$ on $\mathcal{C}(\mathbb{R}^3, 2)$, where U is the open set of subspaces W_p in $\mathcal{C}(\mathbb{R}^3, 2)$ such that $\partial/\partial x^3 \notin W_p$ and $p \in V$. We define the chart $(x, a) : U \subset \mathcal{C}(\mathbb{R}^3, 2) \longrightarrow \mathbb{R}^3 \times \mathbb{R}^2$ as the map $(x, a) : W_p \longmapsto \mathcal{L}^{-1}(W_p)$.[4]

If the base chart x is the standard orthonormal one, this is essentially the usual (x, y, z, p, q) chart of gradient space, where the special coordinate x^3 is z, and $(p, q) = (a_1, a_2)$. Typically, we would write a surface S in coordinates as $(x, y, z(x, y))$. By looking at velocities of the paths $(t, y, z(t, y)$ and $(x, t, z(x, t))$ in S, we see that the two-dimensional tangent subspace to S would be the subspace spanned by the vectors $(1, 0, z_x)^T = (1, 0, p)^T$ and $(0, 1, z_y)^T = (0, 1, q)^T$. This subspace is (in coordinates) the range of the matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ z_x & z_y \end{bmatrix},$$

exactly in the form of the matrix for $L_p{}^a$; taking $p = z_x$, and $q = z_y$ makes $\mathcal{L}((x, y, z), (p, q)) = T_p S$.

[4] To avoid a Greek explosion, we use the standard (admittedly confusing) convention of giving the chart the same name as the coordinates we are going to use, e.g. $x : M \to \mathbb{R}^m$ is a coordinate chart whose values are $x(p) = x = (x_1, x_2, x_3)$. This works because a chart $\phi : M \to \mathbb{R}^m$ does divide into m real functions $\phi_i : M \to \mathbb{R}$ such that $\phi(p) = (\phi_1(p), \ldots, \phi_m(p))$. This useful confusion is also the source of the proper definition of $\mathrm{d}x, \mathrm{d}y, \mathrm{d}z$, etc.: $\mathrm{d}z$ is the derivative map of the coordinate function $z : M \to \mathbb{R}$, and so is a differential one-form on the manifold.

A graphic picture of the standard orthonormal (x, y, z, p, q) coordinate system comes from looking at normal directions to the tangent planes of S at $(x, y, z(x, y)) \in S$. A normal vector to the surface is parallel to the vector $(1, 0, z_x)^T \times (0, 1, z_y)^T = (-z_x, -z_y, 1)^T$. The intersection of this direction with the plane

$$P = \{\mathbf{v} | \mathbf{v} \cdot (0, 0, -1)^T = 1\}$$

(perpendicular to $(0, 0, 1)^T$ and located a distance 1 down the z-axis) is $(z_x, z_y, -1)^T$, so the coordinates (p, q) are the first two coordinates of this point on the plane (Figure 3.2).

What we have done is to localize this construction: using the local vector space coordinates defined by $\{\partial/\partial x, \partial/\partial y, \partial/\partial z\}$, at each point $(x, y, z) \in \mathbb{R}^3$ we can construct a plane perpendicular to the $\partial/\partial z$ direction at the point $(0, 0, -1)$ (where the origin is now at (x, y, z)). A two-dimensional tangent subspace at p has a normal direction through p which intersects this plane with (local) coordinates $(p, q, -1)$.

In the Appendix to this chapter, we show that two charts (V, x) and $(\tilde{V}, \tilde{x})$ on $\mathbb{R}^3$ generate two charts $(U, (x, a))$ and $(\tilde{U}, (\tilde{x}, \tilde{a}))$ on $\mathcal{C}(\mathbb{R}^3, 2)$ which smoothly overlap as required to form an atlas for $\mathcal{C}(\mathbb{R}^3, 2)$.

3.1.3 Lifting of surfaces

We can now look at how surfaces in $\mathbb{R}^3$ become surfaces in $\mathcal{C}(\mathbb{R}^3, 2)$. We have a natural map $\Pi^C : \mathcal{C}(\mathbb{R}^3, 2) \longrightarrow \mathbb{R}^3$ defined as $\Pi^C(W_p) = p$. If $i : S \longrightarrow \mathbb{R}^3$ is an embedded two-dimensional surface in $\mathbb{R}^3$, we can "lift" i to be a related embedded 2-surface $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$ in $\mathcal{C}(\mathbb{R}^3, 2)$ so that $\Pi^C \circ \mathrm{i} = i$.

We define the lifted map $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$ by

$$\mathrm{i}(p) = T_{i(p)} i(S) \simeq T_p S.$$

We will not distinguish between $T_{i(p)} i(S)$ and $T_p S$ unless it is necessary. Essentially this new surface in $\mathcal{C}(\mathbb{R}^3, 2)$ includes explicitly the information about the surface

orientation: if $p \in i(S)$ is a point on the original embedded surface, then the related point on the lifted surface sitting in $\mathcal{C}(\mathbb{R}^3, 2)$ is defined to be i$(p) = T_pS$, the tangent plane to S at p.

We can look at the coordinate view of this map in the (x, a) chart discussed in the previous section. If we take $s = (s^1, s^2)$ as local coordinates[5] on S and x as local coordinates on $\mathbb{R}^3$, we have in local coordinates $i(s) = (x^1(s), x^2(s), x^3(s))$.[6] Since i is an embedding, we know that the rank of Di is two; without loss of generality (we may have to relabel the x coordinates), let us assume that Di has full rank on the first two coordinates of x. By the inverse function theorem, we know that $\hat{i} : (s^1, s^2) \longmapsto (x^1(s), x^2(s))$ is an isomorphism, and we can effectively change local coordinates on S so that our embedding locally looks like $i : (x^1, x^2) \longmapsto (x^1, x^2, x^3(x^1, x^2))$.

Using this coordinate system for S and the coordinate system x for $\mathbb{R}^3$, we know that

$$Di = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \partial x^3/\partial x^1 & \partial x^3/\partial x^2 \end{pmatrix}$$

in coordinates, so

$$\text{Range}(Di) = \text{Range} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \partial x^3/\partial x^1 & \partial x^3/\partial x^2 \end{pmatrix} = T_pS$$

---

[5] Meaning we have a chart $s : S \to \mathbb{R}^2$ given by $s(p) = (s^1, s^2)$

[6] The charts have become almost invisible here. We have maps arranged as follows:

$$\begin{array}{ccc} S & \xrightarrow{i} & \mathbb{R}^3 \\ \uparrow{\scriptstyle s^{-1}} & & \downarrow{\scriptstyle x} \\ (s^1, s^2) \in \mathbb{R}^2 & \longrightarrow & (x^1(s), x^2(s), x^3(s)) \in \mathbb{R}^3 \end{array}$$

Essentially, we have defined (confusingly, but in keeping with common practice) functions x^i of the coordinates (s^1, s^2) as the coordinates in $\mathbb{R}^3$ of the embedded surface S. The rationale for this naming abuse is that the function $x^i(s)$ is really defined as $x^i(i(s))$, where in the latter expression x^i is a coordinate function on $\mathbb{R}^3$.

and so from the previous section we see that in the coordinate system (x, a) for $\mathcal{C}(\mathbb{R}^3, 2)$ we have

$$\left(x^1, x^2, x^3(x^1, x^2), \frac{\partial x^3}{\partial x^1}, \frac{\partial x^3}{\partial x^2}\right)$$

as the coordinates for T_pS. Note that x does not have to be an orthogonal set of coordinates for this to be true. If we write $(x^1, x^2, x^3, a_1, a_2) = (x, y, z, p, q)$ as is conventional, then the coordinates for the tangent plane to the surface, T_pS, considered as a point in $\mathcal{C}(\mathbb{R}^3, 2)$ are just (x, y, z, z_x, z_y).

3.1.4 Contact 1-forms

We would like to be able to detect when a two dimensional manifold $\tilde{S}$ in $\mathcal{C}(\mathbb{R}^3, 2)$ is really a lifted two-dimensional surface S from $\mathbb{R}^3$. Since $\mathcal{C}(\mathbb{R}^3, 2) \simeq \mathbb{R}^3 \times \mathcal{G}(2, 3)$, we can project any two-dimensional subspace $W_p \subset T_p\mathbb{R}^3$ in $\mathcal{C}(\mathbb{R}^3, 2)$ onto its base point p just by taking the first part of the Cartesian product: we have as before[7]

$$\Pi^C : \mathcal{C}(\mathbb{R}^3, 2) \longrightarrow \mathbb{R}^3$$
$$\Pi^C : (p, W) \in \mathbb{R}^3 \times \mathcal{G}(2, 3) \longmapsto p.$$

Clearly we must have $\Pi^C(\tilde{S})$ as a two-dimensional surface in $\mathbb{R}^3$ in order for $\tilde{S}$ to be a lifted surface.

There is another constraint as well: consider the lifted surface given by i : $S \to \mathcal{C}(\mathbb{R}^3, 2)$ where we have assumed the original embedded surface is $i : S \to \mathbb{R}^3$. Let $\alpha : \mathbb{R} \to S$ be a path on the original surface; the lifted path will be i $\circ\, \alpha : \mathbb{R} \to \mathcal{C}(\mathbb{R}^3, 2)$. Taking a coordinate system (x, y, z) for $\mathbb{R}^3$ and lifting it to a coordinate system (x, y, z, p, q) for $\mathcal{C}(\mathbb{R}^3, 2)$, we can consider the surface to be given locally as $i(r, s) = (r, s, f(r, s))$ and the lifted surface to be i$(r, s) = (r, s, f(r, s), f_r(r, s), f_s(r, s))$. The curve $\alpha(t)$ is $(r(t), s(t))$ in coordinates, so using

[7] $\mathcal{C}(\mathbb{R}^3, 2)$ is a so-called trivial bundle over $\mathbb{R}^3$.

the chart (x, y, z, p, q) we can write $z(\mathrm{i} \circ \alpha(t)) = f(r(t), s(t))$. Taking the time derivative, using the chain rule, and using the chart (x, y, z, p, q), we can write

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} z(\mathrm{i} \circ \alpha(t)) &= f_r \frac{\mathrm{d}}{\mathrm{d}t} x(\mathrm{i} \circ \alpha(t)) + f_s \frac{\mathrm{d}}{\mathrm{d}t} y(\mathrm{i} \circ \alpha(t)) \\ &= p(\mathrm{i} \circ \alpha(t)) \frac{\mathrm{d}}{\mathrm{d}t} x(\mathrm{i} \circ \alpha(t)) + q(\mathrm{i} \circ \alpha(t)) \frac{\mathrm{d}}{\mathrm{d}t} y(\mathrm{i} \circ \alpha(t)).\end{aligned}$$

Using differentials[8] of the coordinate functions (x, y, z) on $\mathbb{R}^3$ we can write this as

$$\mathrm{d}z \circ (\mathrm{i} \circ \alpha)'(t) = p(\mathrm{i} \circ \alpha)[\mathrm{d}x \circ (\mathrm{i} \circ \alpha)'(t))] + q(\mathrm{i} \circ \alpha)[\mathrm{d}y \circ (\mathrm{i} \circ \alpha)'(t))].$$

Since we can generate any vector $\mathbf{v} \in T_p S$ by picking an appropriate path α such that $\alpha'(t) = \mathbf{v}$, we must have the 1-form condition[9]

$$(\mathrm{d}z - p\mathrm{d}x - q\mathrm{d}y) \circ (D\mathrm{i} \circ \mathbf{v}) = 0$$

for all $\mathbf{v} \in T_p S$. We can define a contact 1-form θ in the coordinate system (x, y, z, p, q) (i.e. on the open set $V \subset \mathcal{C}(\mathbb{R}^3, 2)$ on which the chart is defined) as

$$\theta_{(x,y,z,p,q)} = (\mathrm{d}z - p\mathrm{d}x - q\mathrm{d}y)|_{(x,y,z,p,q)}.$$

Thus, if i is the lift of i, then for all $\mathbf{u} \in T_{\mathrm{i}(p)}\mathrm{i}(S)$ we have $\theta(\mathbf{u}) = 0$.

The contact 1-forms are defined within particular coordinate systems, and as such are not defined on the whole space $\mathcal{C}(\mathbb{R}^3, 2)$. However, we can construct the differential ideal, I, generated by a set of contact forms defined for a set of charts covering the manifold. A differential ideal, I, of differential forms is a vector subspace

---

[8] A differential is the derivative of a map from a manifold to the real numbers, in this case the coordinate functions.

[9] For the moment, we are mostly interested in 1-forms, which are just linear maps from $T_p S$ to $\mathbb{R}$: derivatives of scalar maps from a manifold are 1-forms, for example. If f is a real function on the manifold M and θ is a 1-form, we define $f\theta(\mathbf{v}_p) = f(p)\theta(\mathbf{v}_p)$ so that $f\theta$ is also a 1-form on M. The differentials of the coordinate functions form a basis for the linear space of 1-forms at a point p. A differential k-form in general is the generalization of a linear idea: at each point $p \in M$ we stick a multilinear (with k arguments) alternating map acting on the vector space $T_p M$; if the choice is made smoothly, the result is a differential k-form.

of differential forms such that for any form $\theta \in I$ and any other differential form ρ not necessarily in I, $\rho \wedge \theta$ is also in I.[10] We use a partition of unity to force the contact forms to be defined over the whole manifold, and the partition of unity assures that at any point on the manifold $\mathcal{C}(\mathbb{R}^3, 2)$, some of the contact forms will be non-zero. (See Appendix Section A3.2 at the end of this chapter.) It turns out that any contact 1-form θ defined by any coordinate chart (x, y, z, p, q) on $\mathcal{C}(\mathbb{R}^3, 2)$ constructed from a chart (x, y, z) on $\mathbb{R}^3$ is a member of this ideal, so that the ideal itself is a coordinate independent object.

If $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$ is a two-dimensional embedding of S in $\mathcal{C}(\mathbb{R}^3, 2)$ with $\mathrm{Rank}(\Pi^C \circ \mathrm{i}) = 2$, then it turns out a contact 1-form θ restricted to $\mathrm{i}(S)$ is 0 if and only if i is a lifted embedding. The condition that the rank of $\Pi^C \circ \mathrm{i}$ be two is quite reasonable: if i is the lift of a two-dimensional surface in $\mathbb{R}^3$, then the associated embedding into $\mathbb{R}^3$, i, must have rank two, and from the definition of a lift we must have $i = \Pi^C \circ \mathrm{i}$.

We have already shown in coordinates that θ restricted to $\mathrm{i}(S)$ is 0 if i is a lifted embedding. We can show this in a slightly different way by using the notion of the pull-back of a differential form: if θ is a differential k-form on M (so that θ_p is a multi-linear map from T_pM to the real numbers), and we have $\mathrm{i} : N \rightarrow M$ as a map between manifolds N and M, then we can define a differential form $\mathrm{i}^*\theta$ on N as

$$(\mathrm{i}^*\theta)_p(\mathbf{v}_1, \ldots, \mathbf{v}_k) \stackrel{\triangle}{=} \theta_{(\mathrm{i}(p))}(D\mathrm{i} \circ \mathbf{v}_1, \ldots, D\mathrm{i} \circ \mathbf{v}_k)$$

where the k-form $\mathrm{i}^*\theta$ acts on vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in T_pN$. If $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2) : p \longmapsto T_pS$ is a lifted embedding of a two-dimensional surface S and $\theta \in \mathbf{I}$ is a contact 1-form, we can compute $\mathrm{i}^*\theta$ in θ's defining coordinate system. We will use the coordinate system $(U, ((x^1, x^2, x^3), (a^1, a^2))) = (U, (x, a))$ where $\partial/\partial x^3 \notin T_pS$

[10] The wedge product is an operation on differential forms directly connected to the alternating multiplication of multilinear alternating maps on a vector space. It is something like a generalized cross product. For details and definitions, see (Abraham, Marsden, and Ratiu, 1983).

for all p in the neighborhood U: using an argument similar to one used in the definition of coordinate charts for $\mathcal{C}(\mathbb{R}^3, 2)$, we can take coordinates for S so that $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$ is given by $\mathrm{i} : (x^1, x^2) \longmapsto (x^1, x^2, x^3(x^1, x^2), \partial x^3/\partial x^1, \partial x^3/\partial x^2)$ in coordinates. We have

$$\begin{aligned}
\mathrm{i}^*\theta &= \mathrm{i}^*(\mathrm{d}x^3 - a_1\mathrm{d}x^1 - a_2\mathrm{d}x^2) \\
&= \mathrm{d}(x^3 \circ \mathrm{i}) - (a_1 \circ \mathrm{i})\mathrm{d}(x^1 \circ \mathrm{i}) - (a_2 \circ \mathrm{i})\mathrm{d}(x^2 \circ \mathrm{i}) \\
&= \mathrm{d}(x^3(x^1, x^2)) - \frac{\partial x^3}{\partial x^1}\mathrm{d}x^1 - \frac{\partial x^3}{\partial x^2}\mathrm{d}x^2 \\
&= \frac{\partial x^3}{\partial x^1}\mathrm{d}x^1 + \frac{\partial x^3}{\partial x^2}\mathrm{d}x^2 - \frac{\partial x^3}{\partial x^1}\mathrm{d}x^1 - \frac{\partial x^3}{\partial x^2}\mathrm{d}x^2 \\
&= 0,
\end{aligned}$$

so on a lifted embedding a contact 1-form is pulled back to 0, i.e. the contact form restricted to a lifted embedded surface is 0.[11]

On the other hand consider if $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$ is a two-dimensional embedding with $i \triangleq \Pi^C \circ \mathrm{i}$ also of rank two and $\mathrm{i}^*\theta = 0$ for some contact 1-form θ. Because $\Pi^C \circ \mathrm{i}$ is of rank two, on θ's defining coordinate system we can put local coordinates on S so that $\mathrm{i} : (x^1, x^2) \longmapsto (x^1, x^2, x^3(x^1, x^2), a_1(x^1, x^2), a_2(x^1, x^2))$, and

$$\begin{aligned}
\left(\mathrm{i}^*\theta\right)_{(x,a)} &= 0 \\
\mathrm{i}^*\left(\mathrm{d}x^3 - a_1\mathrm{d}x^1 - a_2\mathrm{d}x^2\right) &= 0 \\
\frac{\partial x^3}{\partial x^1}\mathrm{d}x^1 + \frac{\partial x^3}{\partial x^2}\mathrm{d}x^2 - a_1\mathrm{d}x^1 - a_2\mathrm{d}x^2 &= 0.
\end{aligned}$$

Equating coefficients of the cotangent basis $\{\mathrm{d}x^1, \mathrm{d}x^2, \mathrm{d}x^3\}$ we must have

$$a_1 = \frac{\partial x^3}{\partial x^1} \text{ and } a_2 = \frac{\partial x^3}{\partial x^2},$$

indicating that i is the lift of $i = \Pi^C \circ \mathrm{i}$.

[11] For details on rules for calculating with differential forms, see (Abraham, Marsden, and Ratiu, 1983).

3.2 The Reflectance Function and Image Formation

Horn (Horn, 1975) makes use of the reflectance function to summarize the influences of light source properties, surface reflection properties, and image formation physics on the formation of an image from a surface. At each point in space, an orientation of the surface tangent plane will give rise to a particular brightness in the image defined by the reflectance function R. The reflectance map as described by Horn (Horn, 1975) can be considered to act on $\mathcal{C}(\mathbb{R}^3, 2)$: a surface patch with tangent plane W at point p in space will generate a brightness in the image given by $R(p, W)$, i.e. $R : \mathcal{C}(\mathbb{R}^3, 2) \longrightarrow \mathbb{R}$. We will often assume that the reflectance function does not depend on the location in space: essentially, this removes the dependence of image brightness on viewpoint, and is accurate in the limit of long lighting distances and long viewing distances. Full generality of the map R includes the projection of slides onto surfaces: we do not expect that a visual system is able to deal correctly with a completely arbitrary R. Understanding restrictions on comprehensible reflectance functions for human shape from shading is an open area.

An image is a combination of the reflectance function and the image projection. From the discussion in the last chapter about the image projection map, we have a map $\pi^I : \mathbb{R}^3 \rightarrow I$ which generates a map $\pi^I \circ i : S \rightarrow I$ from points on the surface to image points. The brightness at a point $\pi^I(i(p))$ in the image is given by $R(\mathrm{i}(p))$, where $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$ is the lifted embedding of the two-dimensional surface S in the space of possible tangent spaces $\mathcal{C}(\mathbb{R}^3, 2)$. If we define $E : I \longrightarrow \mathbb{R}$ as the brightness of the image at a point in the image, then we have the image irradiance equation:

$$E \circ \pi^I \circ i = R \circ \mathrm{i}.$$

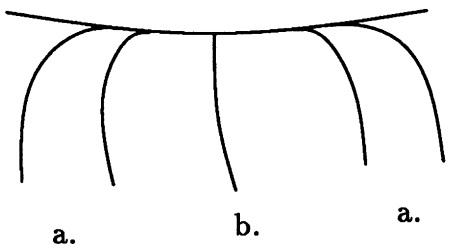


Figure 3.3. Curves on the surface through the bounding contour: a. Usual case. b. Curve parallel to projection direction at the bounding contour.

If we can use local coordinates such that

$$i : (x, y) \in S \longmapsto (x, y, z(x, y))$$

$$\mathrm{i} : (x, y) \in S \longmapsto (x, y, z(x, y), z_x(x, y), z_y(x, y)) \in C(\mathbf{IR}^3, 2)$$

$$\pi^I : (x, y, z) \in \mathbf{IR}^3 \longmapsto (x, y) \in I,$$

$$R : (x, y, z, p, q) \in C(\mathbf{IR}^3, 2) \longmapsto R(x, y, z, p, q) \in \mathbf{IR},$$

and

$$E : (x, y) \in I \longmapsto E(x, y) \in R,$$

then we can write the image irradiance equation in coordinate form as

$$E(x, y) = R(x, y, z, p, q),$$

where $p = z_x$ and $q = z_y$. This is the image irradiance equation of Horn.

3.2.1 Cut vs. Rolled Edges

We can use the reflectance function and the image irradiance equation to understand the difference between image brightnesses at a cut edge and a rolled edge in an image. As we discussed in Section 2.2 the bounding contour of an image consists of those points that are the images of points on the surface where the projection rays just graze the surface. As discussed in Section 2.2, if $i : S \to \mathbf{IR}^3$ describes how the surface is embedded in space, and $\pi^I : \mathbf{IR}^3 \to I$ gives the image projection map, the bounding contour consists of points where $D\pi^I \circ Di$ has rank one, even though both $D\pi^I$ and Di have rank two. What happens to the image constant brightness contours and the values of image brightness at such points in the image?

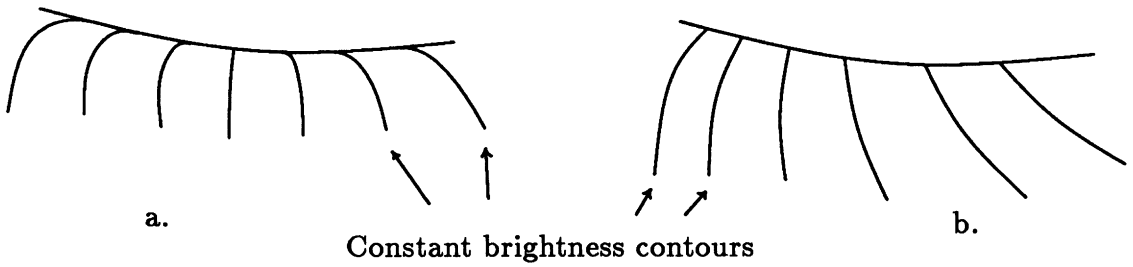


Figure 3.4. a. Rolled constant brightness contours: they are almost always tangent to the bounding contour. b. Cut constant brightness contours: they are almost always not tangent to the bounding contour.

We can consider the fate of the projection of an arbitrary smooth curve $\beta : \mathbb{R} \rightarrow S$ onto the image if $\beta(0)$ occurs on a bounding contour element. The projected curve is $\alpha(t) = \pi^I \circ i \circ \beta(t)$, so the tangent vector to the curve in the image is $\alpha'(t) = D(\pi^I \circ i) \circ \beta'(t)$. At a bounding contour element, $D(\pi^I \circ i)$ has rank one; this means the range of $D(\pi^I \circ i)$ falls in a one dimensional subspace at the image bounding contour point—this is the direction tangent to the bounding contour in the image. Thus, $\alpha'(0)$ is parallel to the bounding contour, and so the projection of β is tangent to the bounding contour (Figure 3.3a). This hides one potential complexity: consider a path $\beta : \mathbb{R} \rightarrow S$ which itself is parallel to the projection direction at the bounding contour at $t = 0$: we have $D(\pi^I \circ i) \circ \beta'(0) = 0$. In this special case, the image of the path $\beta(t)$ can approach the bounding contour at a non-zero angle because the image curve $\alpha(t)$ has a critical point at $t = 0$: the tangent to the geometrical path drawn out in the image is not constrained (Figure 3.3b).

Let us take $\beta : \mathbb{R} \rightarrow S$ as a curve on the surface S which yields a constant brightness: this means $R(\tilde{\beta}) = c$, where $\tilde{\beta} : \mathbb{R} \rightarrow \mathcal{C}(\mathbb{R}^3, 2)$ is the corresponding curve $\tilde{\beta} = \mathrm{i} \circ \beta$ on the lifted surface $\mathrm{i} : S \rightarrow \mathcal{C}(\mathbb{R}^3, 2)$. The image contour corresponding to β will be $\alpha = \pi^I \circ i \circ \beta$; this is now a constant brightness contour in the image.

The constant brightness contours on the surface, determined by the reflectance map and orientation of the surface's tangent planes, intersect the bounding contour on the surface at varying angles. As a result, almost all the image projections of the constant brightness contours will be parallel to the bounding contour if the

discontinuity in the image is indeed due to the surface rolling away from the observer (Figure 3.4a). If the image discontinuity is due to the surface being cut along some curve, the projections of the constant brightness contours will almost all hit this curve at an angle (Figure 3.4b). This provides a clue about whether a boundary of image brightnesses is due to a bounding contour or to a discontinuity in the surface.

The image brightnesses themselves behave badly near the bounding contour. Although the image brightness at the bounding contour in the image is defined by the reflectance function and the tangent plane orientation at the corresponding point on the surface, the derivatives of the image brightness explode as we approach the bounding contour. The image irradiance equation can be written as

$$E \circ (\pi^C \circ \mathrm{i}) = R \circ \mathrm{i},$$

where we define $\pi^C = \pi^I \circ \Pi^C$, where Π^C takes a subspace $W_p \in \mathcal{C}(\mathbb{R}^3, 2)$ and maps it to the base point p, and π^I is the usual image projection from $\mathbb{R}^3$ to the image. Taking the derivative of this we have

$$DE \circ D(\pi^C \circ \mathrm{i}) = DR \circ D\mathrm{i}$$

$$DE = DR \circ D\mathrm{i} \circ (D(\pi^C \circ \mathrm{i}))^{-1},$$

using the inverse function theorem and the chain rule at image points away from the bounding contour so that $D(\pi^C \circ \mathrm{i})$ is invertible. As we approach the bounding contour, $D(\pi^C \circ \mathrm{i})$ has an inverse that becomes larger and larger.[12] As a result, the magnitude of DE will also almost always increase without limit as the bounding contour is approached, even though the value of E is bounded. This is very like the behavior of $\sqrt{x}$ as x approaches 0, and in Chapter 6 we shall make this similarity quite explicit for generic surfaces.

[12] If A is an invertible matrix, we have $\det(A^{-1}) = 1/\det(A)$. As $\det(A)$ approaches zero the magnitude of A^{-1} must explode, since the determinant function is a smooth function on the space of matrices.

3.2.2 The Invariant Image Irradiance Problem

We can state the image irradiance problem in a form that makes an invariant approach to characteristic strips possible. We can define the image projection map as a map on $\mathcal{C}(\mathbb{R}^3, 2)$: as before, we define $\pi^C : \mathcal{C}(\mathbb{R}^3, 2) \longrightarrow I$ by $\pi^C = \pi^I \circ \Pi^C$ as the map $W_p \longmapsto \pi^I(p)$, where Π^C is the standard projection of $\mathcal{C}(\mathbb{R}^3, 2)$ to the base space $\mathbb{R}^3$ and π^I is the image projection. Since $\Pi^C \circ \mathrm{i} = \pi^I \circ i$, where i is the lifted version of the embedding i, we can write the generalized image irradiance equation above as

$$E \circ \pi^C \circ \mathrm{i} = R \circ \mathrm{i}.$$

Using the notation for the pullback of a function by a map, we can write this as

$$\mathrm{i}^*(E \circ \pi^C) = \mathrm{i}^* R,$$

where for a real-valued function $f : M \longrightarrow R$, and map $\phi : N \longrightarrow M$, the pullback, $\phi^* f$, of the map f is defined as the map $\phi^* f \stackrel{\triangle}{=} f \circ \phi : N \longrightarrow R$. Note that if ϕ is an embedding of N in M, the effect of pulling back the function f defined on all of M is to restrict it to the embedded N.

By subtracting, we can write

$$\mathrm{i}^*(E \circ \pi^C - R) = 0.$$

We define the image dynamical system function

$$H : \mathcal{C}(\mathbb{R}^3, 2) \longrightarrow \mathbb{R}$$

by

$$H = E \circ \pi^C - R.$$

The shape from shading problem can be stated in an invariant form as follows: given an image and the reflectance function, we are interested in all possible lifted embeddings of two-dimensional surfaces, $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$, such that

$$\mathrm{i}^* H = 0,$$

since these will correspond to two-dimensional surfaces embedded in $\mathbb{R}^3$ which satisfy Horn's image irradiance equation.

As described in section 3.1.4, this means we are interested in all two-dimensional submanifolds of $\mathcal{C}(\mathbb{R}^3, 2)$ that restrict to two-dimensional surfaces in $\mathbb{R}^3$ and that satisfy both $\mathrm{i}^* H = H \circ \mathrm{i} = 0$ and $\mathrm{i}^* \theta = 0$, where θ is a contact 1-form. This is now in a form that can be converted using differential forms to a vector field on $\mathcal{C}(\mathbb{R}^3, 2)$: solution surfaces are drawn out by curves in $\mathcal{C}(\mathbb{R}^3, 2)$ which are the characteristic strips. We pursue this in the next chapter.

Appendix to Chapter 3

A3.1 Smooth overlapping of generalized (x, y, z, p, q) charts on $\mathcal{C}(\mathbb{R}^3, 2)$

To see that two local diffeomorphisms $(U, (x, a))$ and $(\tilde{U}, (\tilde{x}, \tilde{a}))$ really are charts for $\mathcal{C}(\mathbb{R}^3, 2)$, we need to see that their overlap is smooth. We want to know that for points $(x^1, x^2, x^3, a_1, a_2) \in (x, a)(U \cap \tilde{U})$ the map $(\tilde{x}, \tilde{a})(x, a)^{-1}$ is a smooth map.

To make this calculation, we will look at the coordinates a in a different way by using the space of linear functionals on tangent vectors, also called the cotangent vector bundle, $T^*\mathbb{R}^3$. Consider the map $\alpha : \mathbb{R}^2 \longrightarrow T^*\mathbb{R}^3 : a \longmapsto \alpha^a$ given by

$$\alpha^a = a_1 \mathrm{d}x^1 + a_2 \mathrm{d}x^2 - \mathrm{d}x^3.$$

α defines a 1-form on our coordinate neighborhood. With this definition, and defining $\mathrm{Null}(\alpha^a)$ as the null space of α, we have

$$\mathrm{Null}(\alpha^a) = \mathrm{Range} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ a_1 & a_2 \end{pmatrix},$$

since the null space of α is two dimensional and the 1-form α annihilates the columns of the matrix. If we use a different chart, $(\tilde{U}, (\tilde{x}, \tilde{a}))$ for $\mathcal{C}(\mathbb{R}^3, 2)$, we have a similarly defined 1-form $\tilde{\alpha}^{\tilde{a}}$ for each $\tilde{a}$. If (x, a) and $(\tilde{x}, \tilde{a})$ refer to the same subspace $W_p \in \mathcal{C}(\mathbb{R}^3, 2)$, then we must have

$$W_p = \mathrm{Null}(\alpha^a)_p = \mathrm{Null}(\tilde{\alpha}^{\tilde{a}})_p.$$

But with 1-forms this occurs only if the forms are linearly related on each $T_p\mathbb{R}^3$; thus we must have

$$\begin{aligned}\alpha^a = a_1 \mathrm{d}x^1 + a_2 \mathrm{d}x^2 - \mathrm{d}x^3 &= \lambda(p,a)(\tilde{a}_1 \mathrm{d}\tilde{x}^1 + \tilde{a}_2 \mathrm{d}\tilde{x}^2 - \mathrm{d}\tilde{x}^3) \\ &= \lambda(p,a)\tilde{\alpha}^{\tilde{a}}\end{aligned}$$

at p, where $\lambda(p,a)$ is a real number, and hence λ is a function on $\mathcal{C}(\mathbb{R}^3, 2)$.

We can now express α^a in $(\tilde{x}, \tilde{a})$ coordinates so that we can compare the coefficients of the cotangent basis $\{\mathrm{d}\tilde{x}^1, \mathrm{d}\tilde{x}^2, \mathrm{d}\tilde{x}^3\}$. One way to do this is to use the algebraic rules for expanding the differential form α^a; another way is to observe that using the x coordinate bases for $T_p\mathbb{R}^3$, $\{\partial/\partial x^1|_p, \partial/\partial x^2|_p, \partial/\partial x^3|_p\}$, we have

$$\alpha^a(\mathbf{v}) = (a_1, a_2, -1) \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

in coordinates. If we change the basis for $T_p\mathbb{R}^3$ to $\{\partial/\partial\tilde{x}^1|_p, \partial/\partial\tilde{x}^2|_p, \partial/\partial\tilde{x}^3|_p\}$ we have

$$\alpha^a(\tilde{\mathbf{v}}) = (a_1, a_2, -1)\mathbf{X} \begin{pmatrix} \tilde{v}_1 \\ \tilde{v}_2 \\ \tilde{v}_3 \end{pmatrix},$$

where $\mathbf{X} = [\partial\mathbf{x}/\partial\tilde{\mathbf{x}}]$ is the change of basis matrix for the tangent space. Let

$$(b, c, d) = (a_1, a_2, -1)\mathbf{X},$$

then we must have

$$\alpha^a = b\mathrm{d}\tilde{x}^1 + c\mathrm{d}\tilde{x}^2 + d\mathrm{d}\tilde{x}^3 = \lambda(p,a)(\tilde{a}_1 \mathrm{d}\tilde{x}^1 + \tilde{a}_2 \mathrm{d}\tilde{x}^2 - \mathrm{d}\tilde{x}^3),$$

and, comparing coefficients of $\mathrm{d}\tilde{x}^1, \mathrm{d}\tilde{x}^2$, and $\mathrm{d}\tilde{x}^3$ we must have $\lambda(p,a) = -d \neq 0$, $\tilde{a}_1 = -b/d$, $\tilde{a}_2 = -c/d$. Since the operations used to get $\tilde{a}_1$ and $\tilde{a}_2$ in terms of x and a are locally smooth in nature, and the transition function $x \mapsto \tilde{x}$ is smooth, the whole transition function $(x,a) \mapsto (\tilde{x}, \tilde{a})$ is locally smooth as well.

A3.2 Contact 1-forms

The definition of a contact 1-form in a lifted coordinate neighborhood $(U,(x,a))$ of $\mathcal{C}(\mathbb{R}^3,2)$ is closely connected to the coordinate work in Section A3.1 above. We define a contact 1-form on this neighborhood as

$$\theta_{(x,a)} = a_1 \mathrm{d}x^1 + a_2 \mathrm{d}x^2 - \mathrm{d}x^3$$

in coordinates. θ is related to α^a defined in Appendix Section A3.1:

$$\theta_{(x,a)} = \left(\Pi^C\right)^*(\alpha^a) \triangleq \alpha^a \circ D\Pi^C,$$

where $\Pi^C : \mathcal{C}(\mathbb{R}^3,2) \longrightarrow \mathbb{R}^3 : W_p \longmapsto p$ is the standard projection of $\mathcal{C}(\mathbb{R}^3,2)$, and α^a is the 1-form defined earlier on $\mathbb{R}^3$. This is clearly dependent on the local chart used; note also that θ is a 1-form on $\mathcal{C}(\mathbb{R}^3,2)$, not on $\mathbb{R}^3$ — the coefficients of $\mathrm{d}a_1$ and $\mathrm{d}a_2$ happen to be 0.

At a point $p \in U \cap \tilde{U} \subset \mathcal{C}(\mathbb{R}^3,2)$ with coordinates (x,a) and $(\tilde{x},\tilde{a})$ in the different coordinate charts, we have from Appendix Section A3.1 that

$$\tilde{\alpha}^{\tilde{a}} = \lambda \alpha^a,$$

so

$$\tilde{\theta} = \left(\Pi^C\right)^*(\tilde{\alpha}^{\tilde{a}})) = \left(\Pi^C\right)^*(\lambda\alpha^a) = \lambda\theta,$$

where λ is a real-valued function on $\mathcal{C}(\mathbb{R}^3,2)$ as before. Thus if θ and $\tilde{\theta}$ are two contact 1-forms defined on overlapping regions of coordinate charts $(U,(x,a))$ and $(\tilde{U},(\tilde{x},\tilde{a}))$ respectively, then on the overlap, the 1-forms are linearly related in the sense that there is a function $f : U \cap \tilde{U} \subseteq \mathcal{C}(\mathbb{R}^3,2) \longrightarrow \mathbb{R}$ such that

$$\theta = f\tilde{\theta}$$

on the overlapping set.

Appendix to Chapter 3

Assume we have an atlas of charts, $\{(U_\alpha, (x,a)_\alpha)\}$ for $\mathcal{C}(\mathbb{R}^3, 2)$ with a corresponding partition of unity ρ_α. We can define the differential ideal $\mathbf{I}$ generated by all $\rho_\alpha \theta_\alpha$, where each θ_α is the contact 1-form defined on the coordinate chart $(U_\alpha, (x,a)_\alpha)$. A differential ideal I generated by differential forms $\{\omega_1, \omega_2, \ldots, \omega_k\}$ of a manifold is the graded subalgebra of differential forms

$$I = \left\{ \sum_{i=1}^{k} \rho_i \wedge \omega_i \middle| \ \rho_i \text{ are any differential forms on the manifold} \right\}.$$

I is a graded vector space of differential forms with the additional $\wedge$ operation; it also has the property that for any differential form ρ on the manifold and any differential form γ in I, $\rho \wedge \gamma$ is also in I. This is the property that makes I an ideal in the algebraic sense.[1]

If θ is another contact 1-form defined using another coordinate chart $(U, (x,a))$, then from above we know that on every non-empty $U \cap U_\alpha$ we have

$$\theta = f_\alpha \theta_\alpha,$$

where f_α is a function defined on $U \cap U_\alpha$. Multiplying by the partition of unity we generate forms $\rho_\alpha \theta$ and $\rho_\alpha f_\alpha \theta_\alpha$ that are defined (although often zero) on all of U. Summing over α we have

$$\sum_\alpha \rho_\alpha \theta = \theta = \sum_\alpha \rho_\alpha f_\alpha \theta_\alpha$$

on U, since $\sum_\alpha \rho_\alpha = 1$ as $\{\rho_\alpha\}$ is a partition of unity. Thus any contact 1-form θ defined on an open set $U \subseteq \mathcal{C}(\mathbb{R}^3, 2)$ is in the ideal $\mathbf{I}$ in the sense that there exist functions $g_\alpha : U \longrightarrow \mathbb{R}$ such that $\theta = \sum_\alpha g_\alpha \theta_\alpha$ on U.

[1] For details, try (Edelen, 1985).

Appendix to Chapter 3

Chapter 4

The Image Dynamical System

We now make the transition from the image irradiance equation to the image dynamical system. The image dynamical system is essentially the characteristic vector field defining the characteristic strips in the classical solution of first order partial differential equations. If $F(x, y, z, p, q) = 0$ defines a classical first order partial differential equation for the function $z(x, y)$ with $z_x = p$ and $z_y = q$, then the method of characteristics is to solve the system of ordinary differential equations

$$\begin{aligned}
\dot{x} &= F_p \\
\dot{y} &= F_q \\
\dot{z} &= pF_p + qF_q \\
\dot{p} &= -(F_x + pF_z) \\
\dot{q} &= -(F_z + qF_z).
\end{aligned}$$

A solution to this system starting from some point is called a characteristic strip, where $p(t)$ and $q(t)$ give the orientation of the surface at each point $(x(t), y(t), z(t))$ of the curve. Solutions to the partial differential equation are made up of these characteristic strips; for example, if one chooses an initial strip (i.e. points on a curve together with surface orientations consistent with this curve) crossing the characteristic strips, the characteristics will draw out a solution surface for the partial differential equation beginning at the initial curve.

Systems of ordinary differential equations like the characteristic strip equations can be thought of as vector fields. Solution curves for the system of ordinary differential equations are trajectories that are tangent to the vector field everywhere along their length.

The first part of the mathematical preliminaries section in this chapter shows how the characteristic vector field can be described invariantly; we can then use different coordinate systems depending on the task we are trying to accomplish. The material is more technical than what has come before, relying on ideals of differential forms and the Frobenius theorem; it is based on (Edelen, 1985). The main reason for going through the derivation this way is to allow the vector field to be easily computed using differential forms in very unusual coordinate systems. We will need this for Chapter 6, where we study the image information at the bounding contour by choosing a very special coordinate system in which to write the image irradiance equation. The results of Chapter 5 and Chapter 6 do not depend on understanding the derivation in Section 4.1.1 in detail. The important idea, that the image irradiance equation can be analyzed by integrating a vector field, was used by Horn (Horn, 1975) and is at the heart of all the work reported here.

In the second part of the mathematical preliminaries section, we discuss some of the concepts behind the study of dynamical systems, essentially the study of vector fields and the trajectories that are integral curves for them. We define the image dynamical system as the dynamical system associated with the invariant characteristic vector field of the image irradiance equation. We also show that in the case of orthographic projection with space invariant reflectance function and the usual (x, y, z, p, q) coordinate system on $\mathcal{C}(\mathbb{R}^3, 2)$, the invariant image dynamical system is the same as Horn's characteristic strip equations. In the next chapter we will begin examining the image dynamical system with a technique from dynamical systems analysis, looking at the behavior near critical points.

4.1 Mathematical Preliminaries

4.1.1 From First Order PDE to Vector Field

We begin by looking at some abstract results on ideals of differential forms and vector fields. The approach taken here is algebraic, relying on certain operations defined to act on differential forms, such as the Lie derivative, the contraction (also called the interior product), and the wedge product. For details of definition and properties of these operations which give differential form analysis its peculiar effectiveness, we refer to (Abraham, Marsden, and Ratiu, 1983) and to (Edelen, 1985) from which much of this material was adapted.

4.1.1.1 Differential Ideals and the Frobenius Theorem

In the last chapter we discussed the notion of a differential ideal of differential forms as a vector space of differential forms which is also an ideal under the wedge product of differential forms. We define a characteristic vector field $\mathbf{v}$ for an ideal I on the manifold M to be a vector field such that for all ω in I, the contraction $\mathbf{i}_{\mathbf{v}}\omega$ is also in I; this is written as $\mathbf{i}_{\mathbf{v}}I \subseteq I$.[1]

The set of all characteristic vector fields, $\mathbf{V}$, for an ideal I is a module (essentially a vector space) over the set of real-valued functions on the manifold: that is, if $\mathbf{v}$ is a characteristic vector field, so is $f\mathbf{v}$, where f is a function: this is because $\mathbf{i}_{f\mathbf{v}} = f\mathbf{i}_{\mathbf{v}}$. There is an operation called the Lie bracket on pairs of vector fields which generates another vector field: $[\mathbf{u}, \mathbf{v}]$ is defined as $\mathbf{L}_{\mathbf{u}}\mathbf{v}$, where $\mathbf{L}$ is the Lie derivative with

[1] The differential form $\mathbf{i}_{\mathbf{v}}\omega$ is the contraction of the differential form ω by the vector field $\mathbf{v}$, defined as

$$\mathbf{i}_{\mathbf{v}}\omega(\mathbf{v}_2, \ldots, \mathbf{v}_m) \triangleq \omega(\mathbf{v}, \mathbf{v}_2, \ldots, \mathbf{v}_m),$$

so that $\mathbf{i}_{\mathbf{v}}\omega$ is a differential form of one degree lower than ω. By useful convention, $\mathbf{i}_{\mathbf{v}}f$ is taken to be 0 where f is a 0-form, i.e. a function. The contraction operation takes a k-form and generates a (k-1)-form by plugging in a vector field into the first position of the k-form.

respect to the vector field $\mathbf{u}$.[2] The Lie algebra of vector fields over the real numbers is defined as the vector space of vector fields together with the Lie bracket as a product operation on vector fields. If I is a closed ideal, meaning for all $\theta \in I$ we have $\mathrm{d}\theta \in I$[3] (also written $\mathrm{d}I \subseteq I$), then the space of characteristic vector fields $\mathbf{V}$ is a Lie subalgebra of the Lie algebra of vector fields over the real numbers.[4] To see this, we first use one of Cartan's "magic formulas": for any differential form ω, and vector field $\mathbf{v}$,

$$\mathbf{L}_{\mathbf{v}}\omega = \mathbf{i}_{\mathbf{v}}\mathrm{d}\omega + \mathrm{d}\mathbf{i}_{\mathbf{v}}\omega,$$

where $\mathbf{L}_{\mathbf{v}}\omega$ is the Lie derivative of ω with respect to the vector field $\mathbf{v}$.[5] If $\mathbf{v} \in \mathbf{V}$ is a characteristic vector field for the closed ideal I and ω is in I, this says that $\mathbf{L}_{\mathbf{v}}\omega$ is

---

[2] The Lie derivative operation, $\mathbf{L}$, can be invariantly defined to operate on tensors, including vector fields, functions, and differential forms. It represents a kind of derivative along the flow of the vector field: one definition is that

$$\mathbf{L}_{\mathbf{v}}\omega = \frac{\mathrm{d}}{\mathrm{d}t}[(F_t)^*\omega]|_{t=0},$$

where F_t is the flow of the vector field $\mathbf{v}$ (flows of vector fields are discussed in more detail in Section 4.1.2): since $D(F_t)$ maps T_pM to T_qM where $q = F_t(p)$, the pullback, $(F_t)^*$ takes structures on T_qM and pulls them back to T_pM for all t. Thus, $(F_t)^*\omega$ is a curve of (linear) structures on the vector space T_pM for all time, and its time derivative will also be a linear structure on T_pM. By pulling elements along the flow line of $\mathbf{v}$ back to T_pM we can take a derivative without having to move to a new and more complicated tangent space: we take advantage of the linear structure of T_pM.

[3] If ω is a k-form, then $\mathrm{d}\omega$ is a (k+1)-form which is in some sense a derivative of ω. The operation d can be invariantly defined; however, it may be easiest to think of it in coordinates: if

$$\omega = \sum_{i_1<\ldots<i_k} f_{i_1,\ldots,i_k}\mathrm{d}x^{i_1} \wedge \ldots \wedge x^{i_k}$$

is the coordinatized version of ω in the chart $(x^1, \ldots, x^n)$, then we define

$$\mathrm{d}\omega = \sum_{i_1<\ldots<i_k} \mathrm{d}f_{i_1,\ldots,i_k} \wedge \mathrm{d}x^{i_1} \wedge \ldots \wedge x^{i_k},$$

where $\mathrm{d}g \triangleq \frac{\partial g}{\partial x^1}\mathrm{d}x^1 + \cdots + \frac{\partial g}{\partial x^n}\mathrm{d}x^n$ is defined as the differential of the function g. The operator d has various useful properties in combination with the contraction operation, the Lie derivative, and pull-backs: for example, $\mathrm{d}(\phi^*\omega) = \phi^*(\mathrm{d}\omega)$, so that d commutes with pullback. See (Abraham, Marsden, and Ratiu, 1983) and (Edelen, 1985) for more details.

[4] This means that $\mathbf{V}$ as a linear subspace of the vector fields turns out to be a Lie algebra on its own if I is closed: if $\mathbf{u}$ and $\mathbf{v}$ are vector fields in $\mathbf{V}$, then so is the Lie product $[\mathbf{u}, \mathbf{v}]$ defined as $\mathbf{L}_{\mathbf{u}}\mathbf{v}$.

[5] For this and other properties, see (Abraham, Marsden, and Ratiu, 1983).

also in the ideal I: because $\mathbf{v}$ is a characteristic vector field, $\mathbf{i_v}\omega$ is in I; because I is closed, $\mathrm{d}\omega$ and $\mathrm{d}\mathbf{i_v}\omega$ are also in I, and finally I is a vector space. If we have $\mathbf{v}$ and $\mathbf{u}$ as vector fields on the manifold, and ω a differential form, we also have the identity

$$\mathbf{L_u}(\mathrm{i_v}\omega) = \mathrm{i}_{(\mathbf{L_u v})}\omega + \mathrm{i_v}(\mathbf{L_u}\omega);$$

essentially, the Lie derivative treats $\mathbf{i_v}\omega$ as a product, and $\mathbf{i_v}\omega$ is sometimes referred to as the interior product of $\mathbf{v}$ and ω. Rearranging this identity, for $\mathbf{u}$ and $\mathbf{v}$ both characteristic vector fields of the closed ideal I, we have

$$\mathrm{i}_{(\mathbf{L_u v})}\omega = \mathrm{i}_{[\mathbf{u},\mathbf{v}]}\omega = \mathbf{L_u}(\mathrm{i_v}\omega) - \mathrm{i_v}(\mathbf{L_u}\omega) \in I,$$

and so $[\mathbf{u}, \mathbf{v}]$ is also a characteristic vector field.

The Frobenius theorem for vector fields asserts that a Lie subalgebra of vector fields is *integrable*; i.e., through each point p of M there are submanifolds $i : N \longrightarrow M$ of the same dimension as the Lie subalgebra, $\mathbf{V}$, such that the Lie subalgebra $\mathbf{V}$ spans the tangent space to $i(N)$ at each point p of N. In our case, $\mathbf{V}$ is the space spanned by the characteristic vector fields. Let us assume the characteristic vector fields for the closed ideal I span a space of dimension r at each point; then an integral manifold N of the Lie subalgebra of characteristic vectors $\mathbf{V}$ has dimension r as well.

The integral submanifold $i : N \longrightarrow M$ solves the ideal I in the sense that for all ω in I, $i^*\omega = 0$. To see this, assume ω is a k-form. If $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ are in T_pN, we have

$$\begin{aligned}(i^*\omega)(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k) &= \omega(Di(\mathbf{u}_1), Di(\mathbf{u}_2), \ldots, Di(\mathbf{u}_k)) \\ &= \omega(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k),\end{aligned}$$

where $\mathbf{v}_i = Di(\mathbf{u}_i)$ must be characteristic vectors for I since they are in the tangent space of the integral submanifold N for $\mathbf{V}$. We know that $\mathbf{i_v}\omega \in I$ for any $\omega \in I$, and since

$$\mathbf{i}_{\mathbf{v}_k} \ldots \mathbf{i}_{\mathbf{v}_2}\mathbf{i}_{\mathbf{v}_1}\omega = \omega(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k),$$

we know that this 0-form (or real-valued function) must be in I. The only 0-form in I is the 0 function, since I is assumed to be generated by forms of degree 1 or higher; hence, $i^*\omega = 0$.

If we begin with a differential ideal I that is not closed, we can augment I by adding in all the differential forms $\mathrm{d}I$ to make a new ideal $\tilde{I}$. Since $I \subset \tilde{I}$, solution surfaces for $\tilde{I}$ will be solution surfaces for I. Solution surfaces for I are also solution surfaces for $\tilde{I}$. If $i : N \to M$ is a solution suface for I, then we know $i^*\theta = 0$ for any $\theta \in I$. This means $\mathrm{d}(i^*\theta) = i^*(\mathrm{d}\theta) = 0$, and as a result, $i^*\tilde{\theta} = 0$ for any $\tilde{\theta} \in \tilde{I}$. Hence, N is a solution surface for $\tilde{I}$ as well.

In our case, the base manifold M is really $\mathcal{C}(\mathbb{R}^3, 2)$, a five-dimensional manifold. We begin with the ideal $\hat{I}$ generated by the connection 1-form ideal I and the 1-form $\mathrm{d}H$, where H is a real function, $H : \mathcal{C}(\mathbb{R}^3, 2) \to \mathbb{R}$, the image dynamical system function. We find the closure, $\tilde{\mathbf{I}}$, of this ideal by including all the forms $\mathrm{d}\theta$, where $\theta \in I$ is a connection 1-form. A solution surface for our shape from shading problem will be a submanifold that solves this ideal and on which H has value 0: solving the ideal means finding submanifolds $\mathrm{i} : N \to \mathcal{C}(\mathbb{R}^3, 2)$ such that $\mathrm{i}^*\mathrm{d}H = 0$ and $\mathrm{i}^*\theta = 0$, so that i^*H is a constant (we are interested only in those with $\mathrm{i}^*H = 0$). The submanifolds are potentially lifted surfaces from $\mathbb{R}^3$. Note that if just one point p on such a solution submanifold has $\tilde{H}(p) = 0$, then the whole surface obeys this constraint.

This ideal for the shape from shading problem has a characteristic vector space of dimension one in general. This is a result of the constraints on a characteristic vector field: if $\mathbf{v}$ is a characteristic vector for $\tilde{\mathbf{I}}$, then from the definition of a characteristic vector $\mathbf{i}_\mathbf{v}\theta$ must be a 0-form, or real-valued function, in $\tilde{\mathbf{I}}$. The only 0-form in $\tilde{\mathbf{I}}$ is the zero function, so $\mathbf{i}_\mathbf{v}\theta = 0$ for all connection 1-forms θ. We must also have $\mathbf{i}_\mathbf{v}\mathrm{d}H = 0$ for the same reason. Finally, we must have $\mathbf{i}_\mathbf{v}\mathrm{d}\theta = a\theta + b\mathrm{d}H$, since these are the only possible 1-forms in $\tilde{\mathbf{I}}$ (up to function multiplication). We can look at the tangent space $T_p\mathcal{C}(\mathbb{R}^3, 2)$ at a point p and see how these three constraints on a characteristic

vector restrict the possible space of characteristic vectors at p to a one-dimensional space. We do this in the Appendix to this chapter. Since $\mathbf{V}$ has dimension 1, we have one-dimensional submanifolds $\mathbf{i} : N \to \mathcal{C}(\mathbb{R}^3, 2)$ on which $\mathbf{i}^*\theta = 0$ and $\mathbf{i}^*\mathrm{d}H = 0$. These are the characteristic strips of the classical method.

4.1.1.2 Isovector fields and extension of solutions

Another kind of vector field associated with an ideal is the set of isovector fields of the ideal. A vector field $\mathbf{v}$ is an isovector field for an ideal I if for all θ in I the differential form $\mathbf{L}_\mathbf{v}\theta$, the Lie derivative of θ with respect to $\mathbf{v}$, is also in I. This is also written $\mathbf{L}_\mathbf{v} I \subseteq I$.

For a closed ideal I, i.e. $\mathrm{d}I \subseteq I$, the characteristic vector fields of I are also isovector fields: this was essentially shown above, since for a characteristic vector field $\mathbf{v}$

$$\mathbf{L}_\mathbf{v}\theta = \mathbf{i}_\mathbf{v}\mathrm{d}\theta + \mathrm{d}\mathbf{i}_\mathbf{v}\theta \in I$$

because I is closed and $\mathbf{i}_\mathbf{v} I \subseteq I$ by definition. Isovector fields of an ideal provide a way of extending solutions to the ideal: if $i : N \longrightarrow M$ is a solution surface (not parallel to any of the characteristic vectors along it) for the ideal I so that $i^*I = 0$, and $\mathbf{v}$ is an isovector field for I, then we have a new solution for the ideal of one higher dimension given by

$$j : N \times \mathbb{R} \longrightarrow M$$

$$j : (p, s) \longmapsto F_s \circ i(p),$$

where $F_s : M \longrightarrow M$ is the flow on M generated by the vector field $\mathbf{v}$ (see Section 4.1.2 for more on flows due to vector fields):

$$\frac{\mathrm{d}}{\mathrm{d}s} F_s(p)|_{s=0} = \mathbf{v}(p)$$

$$F_0(p) = p.$$

We show this in the Appendix to this chapter. Essentially, we use the flow lines generated by the isovector field to take points on a lower dimensional solution and extend them to a higher dimensional solution.

Given an isovector field for an ideal and a solution surface for the ideal, we can generate higher dimensional solutions for the ideal if the lower-dimensional solution does not contain the isovector field in its tangent space. In the case of characteristic vectors for the closed ideal generated by the contact 1-form θ and the function $H : \mathcal{C}(\mathbb{R}^3, 2) \rightarrow \mathbb{R}$, this is exactly the classical method of characteristics for finding solution surfaces of the first order partial differential equation $H = 0$ given a curve of initial data; we use the flow due to the characteristic vector field to extend an initial solution curve. Using differential forms the method is extended beyond the range of a single coordinate chart. [6]

4.1.2 Vector Fields and Flows as Dynamical Systems

As we saw in Chapter 2, every point p on a manifold M has a vector space associated with it, T_pM, and these vector spaces can be bundled together into a new manifold called the tangent bundle, TM.[7] If we choose an element $\mathbf{v}(p) \in T_pM$ in each vector space for each point p on the manifold we have a vector field on the

[6] There is another approach to this kind of problem: instead of focusing on the space $\mathcal{C}(\mathbb{R}^3, 2)$ of tangent planes, we can look at the space of orientation vectors for tangent planes. It turns out to be conveniant to use the dual space to the tangent bundle for this purpose, where the dual space $T^*\mathbb{R}^3$ consists of all possible linear functionals acting on the tangent spaces $T_p\mathbb{R}^3$. We can generate a Hamiltonian problem on $T^*\mathbb{R}^3$ corresponding to the characteristic problem on $\mathcal{C}(\mathbb{R}^3, 2)$ using the symplectic 2-form $\omega = \mathrm{d}x \wedge \mathrm{d}p + \mathrm{d}y \wedge \mathrm{d}q$ on $T^*\mathbb{R}^3$; the contact 1-form on $\mathcal{C}(\mathbb{R}^3, 2)$ essentially becomes the canonical 1-form on $T^*\mathbb{R}^3$. Trajectories of this Hamiltonian dynamical system are again related to characteristic trajectories; note that the problem is now set in a six dimensional space.

[7] We can generalize the idea of the tangent bundle to the notion of a vector bundle: we have two manifolds related by a map $\pi : E \rightarrow M$, such that $\pi^{-1}(p)$ is a vector space (isomorphic to some constant vector space V for all the different p). We then speak about a map $\mathbf{v} : M \rightarrow E$ as a section of E if $\pi \circ \mathbf{v}(p) = p$; in other words $\mathbf{v}$ picks out a vector in the vector space connected to p. In our case of a tangent bundle, we have the map $\pi : TM \rightarrow M$ which takes any vector $\mathbf{v} \in T_pM \subset TM$ to its base point $p \in M$, and $TpM = \pi^{-1}(p)$, so the tangent bundle is a vector bundle, and a vector field is a section of this bundle. The idea of a vector bundle can be used to construct lots of linear structures and tools for working with tangent bundles: typically tools from linear algebra form vector spaces, so we can put a copy of each such vector space "above" each point on the manifold to help work on the tangent space at each point.

manifold. Another way to think of this is as a map $\mathbf{v} : M \to TM$ taking a point p to a vector $\mathbf{v}(p) \in T_pM \subset TM$. [8]

We can pick coordinates to examine the detailed behavior of a vector field: if we take a chart (U, ϕ) for M around a point $p \in M$, then at p the intrinsic vector $\mathbf{v}(p) = [(\phi, v)]$ has a coordinate representative $v \in \mathbb{R}^n$ using the chart ϕ. As discussed in Chapter 2, one way to find this coordinate representative is by using paths: say $\mathbf{v}(p)$ can be thought of as the tangent vector at time t of the path $\alpha : \mathbb{R} \to M$. The coordinate view of this path, $\phi \circ \alpha$, will have v as its derivative at $t = 0$: using the derivative map, we have $D(\phi \circ \alpha)_t(1) \stackrel{\triangle}{=} (\phi \circ \alpha)'(t)$, so

$$D(\phi \circ \alpha)_t(1) = D\phi \circ D\alpha_t(1) = D\phi \circ \mathbf{v}(p).$$

It turns out that for a vector field $X : M \to TM$ we can find paths $\alpha(t)$ such that $\alpha'(t) = X(\alpha(t))$ for an entire interval of t's; in other words, at each point along the path, the time derivative of α is the same as the value of the vector field at the point on the path. In fact, we can find an entire family of such paths, $F(x,t) : M \times \mathbb{R} \to M$ such that $\frac{\mathrm{d}}{\mathrm{d}t}F(x,t) = X(F(x,t))$, and with the following additional properties: $F(x,0) = x$, so that the path $F(x,t)$ can be thought of as starting from x at $t = 0$, $F_t(F_s(x)) = F_{t+s}(x)$, where $F_t(x) = F(t,x)$, F is smooth, and under certain mild conditions, F is unique. F is called the flow of the vector field X, and the properties about it are proved using the fundamental existence theorem for ordinary differential equations on $\mathbb{R}^n$: essentially, we can use charts to convert the problem into problems on patches of $\mathbb{R}^n$ and tie these together.[9] If the

[8] Note that even though we are actually making statements about mappings from the manifold to a set of equivalence classes, we can quite happily ignore this formal "complication" and reason as if we were still at home in $\mathbb{R}^n$. Even cleaning up the details does not make much reference to the underlying equivalence classes; one speaks about taking coordinate charts and applying them to vectors and points.

[9] Again, for technical details see (Abraham, Marsden, and Ratiu, 1983).

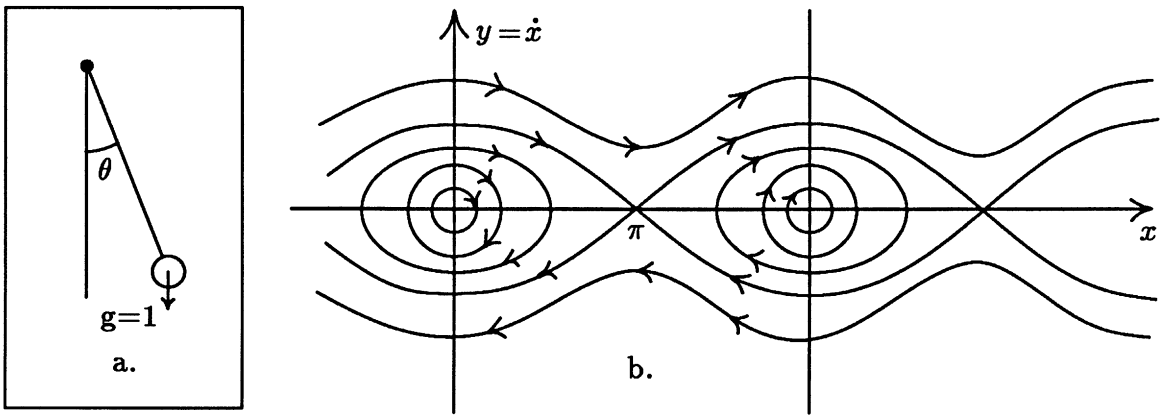


Figure 4.1. The pendulum dynamical system: a. The physical system. b. The flow lines; the vector field is tangent to the flow line at each point.

vector field is thought of as a field of velocity vectors for a fluid, the flow tells how a lightweight particle in the fluid would move over time.

A vector field on a manifold is called a dynamical system. The flow lines $F_t(x)$ considered as a function of t are the trajectories of the system. As an example of such a dynamical system, consider the simplified equation for a two-dimensional pendulum:

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\theta = -\sin\theta.$$

If we take $x = \theta$ and $y = \dot{\theta}$, we can write this second order equation as a pair of first order equations:

$$\dot{x} = y$$

$$\dot{y} = -\sin x.$$

We can package x and y together and consider the vector field

$$X(x,y) = \begin{bmatrix} y \\ -\sin x \end{bmatrix}$$

as a vector field on $\mathbf{IR}^2$.[10] A flow for this vector field will be a map $F : \mathbf{IR}^2 \times \mathbf{IR} \to \mathbf{IR}^2$

[10] We could be more careful and consider it as a vector field on the manifold TS^1, the tangent bundle of the unit circle; the angle θ is a chart for the circle, and the angle and angular velocity together give a tangent vector to the circle, i.e. a point of TS^1. The vector field tells the time derivative of the position and velocity at each point of TS^1.

such that

$$\frac{\mathrm{d}}{\mathrm{d}t}F_t(p)|_{t=0} = X(p);$$

if we consider $(x(t), y(t)) = F_t(x(0), y(0))$ to be a path along the vector field starting at $(x(0), y(0))$, then we get solutions to the first order differential equation (Figure 4.1).

Part of the task of studying a dynamical system is to connect features of the vector field with features of the flow. An important feature of a vector field is a critical point, a point $p \in M$ where the vector field is the 0 vector of T_pM. p is then a fixed point of the flow F, that is $F(p,t) = p$ for all t: the flow "line" $F_t(p) = p$ is consistent with the vector field, since $\frac{\mathrm{d}}{\mathrm{d}t}(F_t(p)) = 0$. For example, in the pendulum case, both the origin $(0,0)$ and the point $(\pi, 0)$ are critical points: $X = 0$ at these points. The behavior of the flow lines around these two points is quite different: at the origin, the flow lines form circles around the critical point describing how the pendulum oscillates back and forth around the origin, with velocity about ninety degrees out of phase with position. At $(\pi, 0)$ we have a very different picture: we seem to have two flow lines intersecting at the critical point. In fact there are four trajectories that approach infinitely close to the critical point at $(\pi, 0)$. One pair of these are trajectories for the pendulum that become infinitely slow at the top of the pendulum's arc ($\theta = \pi$), and the other pair are trajectories which begin infinitely close to the top of the arc with near zero speed and begin to move away with increasing speed. Other flow lines run near these four trajectories without touching the critical point. We shall have much more to say about critical points and their classification in the next chapter.

One of the questions that comes up in the study of a dynamical system is that of stability: what features of a particular dynamical system remain even if the system is slightly perturbed, e.g., the vector field is changed a bit? There are several ways to define what is meant by a dynamical system remaining substantially "unchanged": we shall consider two dynamical systems to be related if the flow lines are

topologically equivalent in the two systems, i.e. there is a homeomorphism from one dynamical system to the other which maps flow lines to flow lines.

Part of the difficulty in answering questions about stability is in defining what sorts of perturbations we are interested in. For example, if we are studying a particular class of vector fields (Hamiltonian vector fields, for example) we have to decide if we are interested in questions of stability in the set of all Hamiltonian vector fields, or stability in the set of all smooth vector fields: the difference can be important. In the pendulum example, if we modify the vector field by adding an arbitrary small smooth vector field to all the points, it turns out that the circular closed orbits around the origin will in all likelihood disappear and become trajectories that spiral in or spiral out from the critical point (friction will make the trajectories spiral inwards, for example); the character of the trajectories near $(0, \pi)$ will be unaffected. On the other hand, if we restrict ourselves to perturbations that give energy conserving Hamiltonian dynamical systems "near" the original, it turns out the closed orbits near the origin are a stable feature of the vector field.

A perturbation approach can be useful in understanding the computational feasibility of a particular dynamical system. If one actually tries to implement a dynamical system on a computer, one has to face the difficulties of discretizing a smooth theoretical construction. One way to model this is as a (presumably small) perturbation from the original problem to the discretized one. Unless special care is taken, this discretization perturbation will not be very forgiving of restrictions on the class of dynamical systems in which a feature is stable: it may be quite difficult to accurately study features that are unstable under relatively general small perturbations.

The computational problem is one kind of perturbation stability to be concerned about. Another comes from the lack of exact knowledge about the dynamical system itself. In the image dynamical system we will construct, we assume we know the reflectance function. It is of interest to know how features of the dynamical system

change if we have made a small mistake in this regard. This is different from the computational perturbation in that the perturbation of the reflectance function may generate a quite restricted perturbation of the vector field due to the special form of the vector field. This sort of perturbation study is an open question for the image dynamical system.

4.2 The Image Irradiance Dynamical System

As indicated last chapter, we can summarize the image solution problem as the question of finding lifted solution surfaces $\mathrm{i} : S \to \mathcal{C}(\mathbb{R}^3, 2)$ such that

$$\mathrm{i}^* H = 0.$$

We can consider this problem to be that of finding all two-dimensional submanifolds projecting to two-dimensional manifolds in $\mathbb{R}^3$ and such that $\mathrm{i}^* H = 0$ and $\mathrm{i}^* \theta = 0$, where θ is any contact 1-form.

The function $H = E \circ \pi^C - R$ defined in Chapter 3 essentially determines a first order partial differential equation. As described above, we can use contact 1-forms and the calculus of differential forms together with the Frobenius theorem to convert the above problem into a vector field integration problem. The characteristic vector field derived in the last section is the same as the characteristic vector field of Horn, but is not tied to a particular coordinate representation.

We can show that the characteristic vector field is the same as Horn's characteristic strip equations in a particular coordinate system. We will work within the following framework: we assume orthogonal projection and global rectilinear coordinates for $\mathbb{R}^3$ and the image plane such that in coordinates we have

$$\pi^I(x, y, z) = (x, y).$$

We also assume a space invariant reflectance function so that $R(x, y, z, p, q) = R(p, q)$. We want to find the components of a characteristic vector field X. We have the

constraint $\mathbf{i}_X \mathrm{d}\theta = a\theta + b\mathrm{d}H$. In coordinates, we have $\theta = \mathrm{d}z - p\mathrm{d}x - q\mathrm{d}y$, $\mathrm{d}\theta = \mathrm{d}x \wedge \mathrm{d}p + \mathrm{d}y \wedge \mathrm{d}q$, and $\mathrm{d}H = E_x\mathrm{d}x + E_y\mathrm{d}y - R_p\mathrm{d}p - R_q\mathrm{d}q$. Using the definition of contraction acting on the wedge product,[11] the constraint $\mathbf{i}_X \mathrm{d}\theta = a\theta + b\mathrm{d}H$ on the characteristic vector field $X = (X_x, X_y, X_z, X_p, X_q)$ in coordinates becomes:

$$X_x\mathrm{d}p - X_p\mathrm{d}x + X_y\mathrm{d}q - X_q\mathrm{d}y$$
$$= a(\mathrm{d}z - p\mathrm{d}x - q\mathrm{d}y) + b(E_x\mathrm{d}x + E_y\mathrm{d}y - R_p\mathrm{d}p - R_q\mathrm{d}q).$$

Gathering together coefficients of the basis 1-forms $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$, $\mathrm{d}p$, $\mathrm{d}q$, we have

$$a = 0$$
$$X_x = -bR_p$$
$$X_y = -bR_q$$
$$X_p = -bE_x$$
$$X_q = -bR_y.$$

We also have $\mathbf{i}_X\theta = 0$; evaluated in coordinates this contraction becomes $X_z - pX_x - qX_y = 0$. Thus, in coordinates our vector field X looks like

$$X = -b \begin{bmatrix} R_p \\ R_q \\ pR_p + qR_q \\ E_x \\ E_y \end{bmatrix}$$

where b is an arbitrary constant. An integral path for such a vector field with $b = -1$ would have

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} x \\ y \\ z \\ p \\ q \end{bmatrix} = \begin{bmatrix} R_p \\ R_q \\ pR_p + qR_q \\ E_x \\ E_y \end{bmatrix};$$

these are the characteristic strip equations as derived by Horn (Horn, 1975).

---

[11] It can be shown that

$$\mathbf{i}_\mathbf{v}(\theta \wedge \omega) = (\mathbf{i}_\mathbf{v}\theta) \wedge \omega + (-1)^k \theta \wedge (\mathbf{i}_\mathbf{v}\omega),$$

where θ is a k-form: see (Abraham, Marsden, and Ratiu, 1983).

We can view the shape from shading problem as a dynamics problem on $\mathcal{C}(\mathbb{R}^3, 2)$, where the dynamics are defined by the characteristic vector field. The dynamical trajectories of the system essentially draw out characteristic strips on possible solution surfaces. A possible solution surface is therefore an invariant surface of the characteristic flow, in the sense that points on the surface move to other points on the surface under the characteristic flow.

As Horn recognized, some way to choose a subset of the characteristic trajectory strips is needed: they fill a volume of $\mathcal{C}(\mathbb{R}^3, 2)$. One way might be the specification of an initial non-characteristic trajectory from which to draw the rest of the surface using characteristic trajectories: this is the classical Cauchy problem of first order partial differential equations. Unfortunately, we do not usually have such an initial strip given. How can we reduce the ambiguity in the choice of characteristic trajectories that will make up a solution surface?

One way to look at the ambiguity of solution surfaces is to consider how many solution surface patches are consistent with a small image patch and the given reflectance function. We can examine patches without critical points or bounding contour points, patches with just critical points, or patches with just bounding contour points, for example.

If the image patch does not contain critical points of the image intensity or any bounding contour elements, then we can take a curve $(x(t), y(t))$ in the image patch and specify depth values $z(t)$ along it. The image irradiance equation $E(x, y) = R(p, q)$ and the contact 1-form integrability condition $z' = px' + qy'$ will almost always lift this curve in $\mathbb{R}^3$ to a curve in $\mathcal{C}(\mathbb{R}^3, 2)$ by determining p and q. If this curve is not tangent to any of the characteristic curves (this is one reason critical points have to be excluded), then a solution patch will be drawn out by the characteristic curves from this initial curve in $\mathcal{C}(\mathbb{R}^3, 2)$. The ambiguity of smooth solution surfaces for such an image patch can be summed up by the smooth specification of depth values along a given path in the image.

There may be problems with such solutions as they are extended beyond the bounds of a small patch. One kind of problem that may arise is the folding of the surface "unnaturally" in $\mathcal{C}(\mathbb{R}^3, 2)$ in such a way that no possible space surface could have generated it without self-intersections or other strange behavior we do not expect for a surface consistent with a smooth image. This is an open area for exploration.

We look at ambiguities of an image patch which contains a critical point in the next chapter. The use of critical points to help determine solution surfaces was first explored by Horn (Horn 1975) who used the critical point to generate an approximate initial contour on the surface from which to draw the rest of the surface with characteristics. Bruss (Bruss 1980) made further theoretical progress for a particular class of reflectance functions.

Critical points in the image due solely to critical points in the reflectance function determine critical points of the characteristic vector field on $\mathcal{C}(\mathbb{R}^3, 2)$. The key idea exploited in the next chapter is that a possible solution surface for a smooth region of an image containing a critical point should be a smooth, invariant manifold of the image dynamical system containing the corresponding critical point of the image dynamical system. As we shall see, in general this provides strong constraints on what the behavior of the surface can be, since most characteristic curves near a critical point will not actually approach the critical point.

In Chapter 6 we look at ambiguities in solution surfaces for an image patch containing a piece of bounding contour. As indicated in Chapter 2, the bounding contour is a curve in the image for which we actually know the surface normal direction at each point. We will suggest that local patches of bounding contour image data provide more of a constraint on the reflectance function than on the local behavior of the surface near the bounding contour: given the correct reflectance function, we hypothesize (and show to third order) that the surface is determined only up to a choice of depth values along the bounding contour in the image, very similar to an image patch in the interior without critical points.

By examining the task of extracting shape information from shading in an image with geometric tools, one hopes to make clear the contributions of different sources of information to potential solutions. Hopefully, one can see more precisely why certain assumptions are required, are useful, or are reasonable in choosing an interpretation of an image.

Appendix to Chapter 4

A4.1 The dimension of the characteristic subspace

We show here that the characteristic subspace of the closed ideal generated by the contact ideal and the function $H : \mathcal{C}(\mathbb{R}^3, 2) \to \mathbb{R}$ is in general one dimensional.

One way to do this is to look at the coordinate representation of these constraints using matrix notation. Pick a coordinate neighborhood $(U, (x, a))$ for $\mathcal{C}(\mathbb{R}^3, 2)$ used to define a connection 1-form θ. In this coordinate system we can write the 1-forms θ and $\mathrm{d}H$ as

$$\theta = \mathrm{d}x^3 - a_1 \mathrm{d}x^1 - a_2 \mathrm{d}x^2$$

$$\mathrm{d}H = H_1 \mathrm{d}x^1 + H_2 \mathrm{d}x^2 + H_3 \mathrm{d}x^3 + H_4 \mathrm{d}a_1 + H_5 \mathrm{d}a_2.$$

The 2-form $\mathrm{d}\theta$ is an antisymmetric 2-tensor at p, and so can be represented as an antisymmetric 5×5 matrix, $\mathbf{A}$. In this coordinate system we have $\mathrm{d}\theta = \mathrm{d}x^1 \wedge \mathrm{d}a_1 + \mathrm{d}x^2 \wedge \mathrm{d}a_2$, so in matrix form we can write

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \end{bmatrix},$$

where

$$\mathrm{d}\theta(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T \mathbf{A} \mathbf{v}$$

using the coordinate vector representations for $\mathbf{u}$ and $\mathbf{v}$. As row vectors in this coordinate system we can write

$$\theta = [-a_1, -a_2, 1, 0, 0]$$
$$\mathrm{d}H = [H_1, H_2, H_3, H_4, H_5].$$

Two conditions for a characteristic vector $\mathbf{v}$, $\theta(\mathbf{v}) = 0$ and $\mathrm{d}H(\mathbf{v}) = 0$, can be combined and rewritten in matrix form as

$$\mathbf{W}\mathbf{v} \triangleq \begin{bmatrix} -a_1 & -a_2, & 1 & 0 & 0 \\ H_1 & H_2 & H_3 & H_4 & H_5 \end{bmatrix} \begin{bmatrix} v^1 \\ v^2 \\ v^3 \\ v^4 \\ v^5 \end{bmatrix} = 0.$$

The condition $\mathbf{i_v}\mathrm{d}\theta = a\theta + b\mathrm{d}H$ can also be written in vector form in the following way: there exists a 2-vector $\mathbf{u} \in \mathbb{R}^2$ such that

$$\mathbf{v}^T\mathbf{A} = \mathbf{u}^T \begin{bmatrix} -a_1 & -a_2, & 1 & 0 & 0 \\ H_1 & H_2 & H_3 & H_4 & H_5 \end{bmatrix} = \mathbf{u}^T\mathbf{W}.$$

Notice that the rows of the matrix on the right are the representations of θ and $\mathrm{d}H$. The assertion of the matrix statement is that the 1-form $\mathbf{i_v}\mathrm{d}\theta$ (represented by the row vector $\mathbf{v}^T\mathbf{A}$) is a linear combination of θ and $\mathrm{d}H$ (represented as row vectors) with coefficients given by the 2-vector $\mathbf{u}$. By making an augmented row vector $[\mathbf{v}^T, \mathbf{u}^T]$, we can write this condition as

$$[\mathbf{v}^T, \mathbf{u}^T] \begin{bmatrix} \mathbf{A} \\ -\mathbf{W} \end{bmatrix} = 0.$$

We can add the two 1-form conditions to this expression by augmenting the matrix:

$$[\mathbf{v}^T, \mathbf{u}^T] \begin{bmatrix} \mathbf{A} & \mathbf{W}^T \\ -\mathbf{W} & 0 \end{bmatrix} = [0, 0].$$

If we find an augmented non-zero vector $[\mathbf{v}^T, \mathbf{u}^T]$ such that this condition is obeyed, then the $\mathbf{v}$ part obeys the conditions for a characteristic vector.

Appendix to Chapter 4

Since the augmented 7×7 matrix

$$\tilde{\mathbf{A}} \triangleq \begin{bmatrix} \mathbf{A} & \mathbf{W}^T \\ -\mathbf{W} & 0 \end{bmatrix}$$

$$= \left[\begin{array}{ccccc|cc} 0 & 0 & 0 & 1 & 0 & -a_1 & H_1 \\ 0 & 0 & 0 & 0 & 1 & -a_2 & H_2 \\ 0 & 0 & 0 & 0 & 0 & 1 & H_3 \\ -1 & 0 & 0 & 0 & 0 & 0 & H_4 \\ 0 & -1 & 0 & 0 & 0 & 0 & H_5 \\ \hline a_1 & a_2 & -1 & 0 & 0 & 0 & 0 \\ -H_1 & -H_2 & -H_3 & -H_4 & -H_5 & 0 & 0 \end{array}\right]$$

is antisymmetric, it has maximum possible rank 6 (an antisymmetric matrix always has even rank); by examining the columns of the matrix, one can see that the matrix always has rank 6. There is always a one dimensional subspace of augmented vectors $[\mathbf{v}^T, \mathbf{u}^T]$ such that $[\mathbf{v}^T, \mathbf{u}^T]\tilde{\mathbf{A}} = 0$. If the $\mathbf{v}$ part of an augmented vector in this subspace is non-zero, then it is a characteristic vector, since by the above construction $\mathbf{i_v}\theta = 0 \in \tilde{\mathbf{I}}$ for connection 1-forms θ, $\mathbf{i_v}\mathrm{d}H = 0 \in \tilde{\mathbf{I}}$, and $\mathbf{i_v}\mathrm{d}\theta = a\theta + b\mathrm{d}H \in \tilde{\mathbf{I}}$, so $\mathbf{i_v}\tilde{\mathbf{I}} \subseteq \tilde{\mathbf{I}}$.

When will the null space of $\tilde{\mathbf{A}}$ have a $\mathbf{v}$ component that is zero? For this to happen we must have $\mathbf{u}^T\mathbf{W} = 0$, which means the two 1-forms θ and $\mathrm{d}H$ are a dependent set. This could happen either because they are linear multiples of each other, or when $\mathrm{d}H = 0$. Assuming the coordinate representation used previously for image projection $\pi^I : \mathbb{R}^3 \longrightarrow \mathcal{I} : (x^1, x^2, x^3) \longmapsto (x^1, x^2)$, and using the definition of H, $\mathrm{d}H = 0$ describes a kind of picture function extremum which occurs when

$$E_{x^1} = R_{x^1}$$

$$E_{x^2} = R_{x^2}$$

$$R_{x^3} = R_{a_1} = R_{a_2} = 0.$$

In the space invariant reflectance function case where $R_{x^1} = R_{x^2} = R_{x^3} = 0$, this means that an extremum has occured in both the image brightnesses and the image reflectance function. The case where $\mathrm{d}H$ and θ are dependent also reduces to this case if the reflectance function is space-invariant.

For the most part, the $\mathbf{v}$ component of the augmented vector will be non-zero and therefore there will be a one-dimensional set of characteristic vectors for the ideal $\tilde{\mathbf{I}}$.

A4.2 Extension of solutions by isovector field flow

We show here that if F_s is the flow of an isovector field $\mathbf{v}$ of a differential ideal I, then the embedding $j : N \times \mathbb{R} \to M$ given by $j(p,s) = F_s(i(p)$ extends a solution surface $i : N \to M$ for the ideal I by one dimension.

To see that j is a solution to the ideal, we observe that

$$j^*\theta = i^*{F_s}^*\theta.$$

It turns out that ${F_s}^*I \subseteq I$: one way to see this is to observe that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}({F_s}^*\theta)|_{s=s'} &= \frac{\mathrm{d}}{\mathrm{d}t}({F_{s'+t}}^*\theta)|_{t=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}t}{F_t}^*({F_{s'}}^*\theta)|_{t=0} \\ &= \mathbf{L_v}({F_{s'}}^*\theta),\end{aligned}$$

since $(\mathrm{d}/\mathrm{d}t){F_t}^*\omega|_{t=0} = \mathbf{L_v}\omega$, the Lie derivative of ω with respect to $\mathbf{v}$, for any differential form ω. At $s = 0$ we have

$${F_s}^*\theta|_{s=0} = \theta.$$

If we consider these two as a first order differential equation on the vector space (here thought of as over the real numbers) of differential forms, i.e.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\rho(s) &= \mathbf{L_v}\rho(s) \\ \rho(0) &= \theta,\end{aligned}$$

where $\mathbf{L_v}$ is now a linear operator on the real vector space of differential forms, we have the operator series solution given as

$$\rho(s) = \exp(s\mathbf{L_v}) \circ \theta \triangleq \left(\sum_{i=0}^{\infty} s^i \mathbf{L_v}^i\right)\theta.$$

By uniqueness of the solution $\rho(s)$ to the differential equation with initial condition, we must have

$$F_s{}^*\theta = \exp(s\mathbf{L_v}) \circ \theta.$$

We have assumed that $\mathbf{L_v} I \subseteq I$ since $\mathbf{v}$ is an isovector field; thus, the series expression

$$\left(\sum_{i=0}^{\infty} s^i \mathbf{L}_\mathbf{v}^i\right)\theta$$

must also remain in I for all θ in I, and so $F_s{}^* I \subseteq I$.[1]

Since i is a solution for I we have $i^* I = 0$; since $F_s{}^* I \subseteq I$, we have $i^* F_s{}^* I = 0$, and hence $j(s,p) \stackrel{\Delta}{=} F_s(i(p))$ is a new solution for the ideal I. In order for j to be an embedding Dj must have full rank. To calculate Dj we observe that if we consider the flow F_s of the vector field $\mathbf{v}$ as a map $F : \mathbb{R} \times M \longrightarrow M : (s,p) \longmapsto F_s(p)$, then

$$DF : TR \times TM \longrightarrow TM$$
$$DF : \begin{bmatrix} r \\ \mathbf{u} \end{bmatrix} \longmapsto [D_1 F, D_2 F] \begin{bmatrix} r \\ \mathbf{u} \end{bmatrix},$$

---

[1] Another way to see this is to consider the space of all differential k-forms on M as an infinite dimensional manifold, $\mathcal{M}$, with individual forms θ being points on the manifold. It so happens that this space is a vector space: we can add points in the space, and multiply by scalars (real numbers here). The tangent space, $T_\theta \mathcal{M}$, to a point θ in this manifold is therefore equivalent to the original space itself, $T_\theta \mathcal{M} \simeq \mathcal{M}$, and at each point $T_\theta \mathcal{M}$ will also be infinite dimensional. A tangent vector can be considered as an ordered pair of k-forms, e.g. $(\theta, \rho) \in T_\theta \mathcal{M}$, with θ being the base point, and ρ giving the tangent vector representative in $\mathcal{M}$. A vector field on this new manifold consists of a choice of tangent vector for each base point in $\mathcal{M}$. We can consider $\mathbf{L_v}$ to be a vector field: for each point θ in $\mathcal{M}$, we consider $\mathbf{L_v}$ to pick out the tangent vector $(\theta, \mathbf{L_v}\theta)$. We can now properly consider the flow of this vector field on the manifold $\mathcal{M}$, $\mathcal{F}_s : \mathcal{M} \longrightarrow \mathcal{M}$ where

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathcal{F}_s(\theta)|_{s=0} = (\mathbf{L_v})(\theta) = (\theta, \mathbf{L_v}\theta)$$
$$\mathcal{F}_0(\theta) = \theta$$

defines the flow for the vector field $\mathbf{L_v}$ on the infinite dimensional manifold $\mathcal{M}$. If the vector field $\mathbf{L_v}$ restricted to a submanifold I of $\mathcal{M}$ lies completely in the tangent space to I, then we know that the induced flow of a point on the submanifold I must also lie in this submanifold. In our case, I is the set of k-forms in the ideal $\mathbf{I}$, and is in fact a subspace of $\mathcal{M}$. The vector field $\mathbf{L_v}$ at a point $\theta \in I$ is also in I when $\mathbf{v}$ is an isovector field, since $\mathbf{L_v}\theta \in I$. The flow is the flow $\mathcal{F}_s$, and so $\mathcal{F}_s(\theta) \in I$. By uniqueness of solution, we know $\mathcal{F}_s(\theta) = F_s{}^*\theta$, and so $F_s{}^* I \subseteq I$.

where we have split the derivative DF into two components, the first with respect to flow time and the second with respect to location on the manifold. From the flow equation we know that

$$D_1 F_{(s,p)} = \mathbf{v}(F(s,p)),$$

where $(s,p) \in \mathbb{R} \times M$. At (s,p) we have

$$\mathbf{v}(F_s(p)) = \frac{\mathrm{d}}{\mathrm{d}t}(F_t \circ F_s)|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}(F_s \circ F_t)|_{t=0} = D_2 F_{(s,p)} \circ \mathbf{v}(p).$$

Since $j(s,p) = F \circ (s, i(p))$, we have

$$\begin{aligned} Dj_{(s,p)} \begin{bmatrix} r \\ \mathbf{u} \end{bmatrix} &= \begin{bmatrix} D_1 F & D_2 F \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Di \end{bmatrix} \begin{bmatrix} r \\ \mathbf{u} \end{bmatrix} \\ &= \begin{bmatrix} D_1 F & D_2 F \circ Di \end{bmatrix} \begin{bmatrix} r \\ \mathbf{u} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{v}(F(s,i(p))) & D_2 F \circ Di_p \end{bmatrix} \begin{bmatrix} r \\ \mathbf{u} \end{bmatrix} \\ &= \begin{bmatrix} D_2 F \circ \mathbf{v}(i(p)) & D_2 F \circ Di_p \end{bmatrix} \begin{bmatrix} r \\ \mathbf{u} \end{bmatrix}, \end{aligned}$$

and so the rank of j is full if and only if $\mathbf{v}(i(p))$ and Di_p span independent spaces for all $p \in N$. Note that this condition needs only to be verified along the lower dimensional submanifold $i : N \longrightarrow M$, not along the flow lines: the flow of $\mathbf{v}$ is a diffeomorphism so that $D_2 F$ is full rank, and this makes j an embedding throughout the range of definition of the flow if it is an embedding along the initial submanifold.

Appendix to Chapter 4

Chapter 5

Critical Points of the Image Dynamical System

In this chapter we discuss the role of critical points in the image in determining solution surfaces for the shape from shading problem. The main result is that in general (i.e. generically, in the absence of special symmetries) near certain kinds of critical points in the image there are at least two and at most four possible solution surfaces. These critical points are due to critical points in the reflectance function, and are either the result of local maxima (the usual case) or minima in the reflectance function. More complicated behavior may occur with saddles in the reflectance function; more work is needed to understand this case completely.

We use the image dynamical system developed in the last chapter and apply a technique from dynamical systems theory called linearization to study the behavior of the dynamical system near the critical points. Linearization essentially involves looking at the first order behavior of the characteristic equations around the critical point to get a linear dynamical system with behavior similar to the nonlinear system. A linear dynamical system $X = \mathbf{A} \cdot \mathbf{x}$ can be analyzed by looking at eigenvalues and eigenvectors of the matrix $\mathbf{A}$, and in general invariant subspaces of linear dynamical systems are vector spaces spanned by sets of eigenvectors. If the eigenvalues of the nonlinear system do not have zero real parts, the invariant manifolds of the nonlinear system are topologically isomorphic to those of the linear system on a neighborhood

of the critical point. We analyze the image dynamical system in this way to get results.

We look to see when critical points in the image are "good" critical points to which this analysis can be applied, and we look at the connection between the reflectance function critical point type, the surface type at the critical point, and the type of the two-dimensional image dynamical system restricted to the correct solution surface.

Two of the invariant manifolds that are possible solution surfaces are the stable and unstable manifold. Simply reversing time interchanges the labels, so it is of interest to have ways to find, say, unstable manifolds for the image dynamical system. In Section 5.3 of this chapter, we show a method based on a mathematical theorem called the Lambda Lemma: we take an initial surface that cuts the stable manifold, and allow it to flow forward using the image dynamical system. The Lambda Lemma says that as t goes to infinity, the deformed surface will C^1 approach the unstable manifold (Palis and de Melo, 1982). We show some experiments using the Connection Machine, a highly parallel computer, on implementing this idea. The resulting methods seem to have good noise tolerance and robustness in the face of wrong information about reflectance functions.

5.1 Mathematical Preliminaries

Modern dynamics has emphasized the study of the local behavior of trajectories near critical elements of a system. There are three types of critical elements to consider: one type is the closed orbit with a finite period; another type is a critical point of the vector field, i.e. a point p where $X(p) = 0$; this can be considered as a trajectory with infinitely short period. Finally, there are other critical elements which can be described as chaotic. We will concern ourselves exclusively with critical points; existence and properties of the other critical elements for an image dynamical system is an open area for research.

Dynamicists have been very interested in the stable and unstable manifolds, S^+ and S^-, associated with a critical element in a vector field as a way of classifying the critical point and studying its interaction with other critical points. The stable manifold contains those trajectories that wind towards the critical element as time along the trajectory proceeds in the positive direction; the unstable manifold consists of the trajectories that wind towards the critical element with decreasing time. "Wind toward" means that the critical elements are approached in the limit as time runs to infinity, positive or negative in the stable and unstable cases respectively. In our situation, "time" is just a parameter along the trajectories of interest, so the stable and unstable manifolds are quite similar in character: they are sets that are invariant under the flow of the vector field, they include the critical element, and they contain trajectories that all approach the critical element asymptotically.

5.1.1 Critical Points and Invariant Manifolds

Much of the methodology comes from (Abraham and Marsden, 1985). To examine in more detail the trajectories near a critical point, we can use the linearized version of the vector field. If p is a critical point for the vector field X on a manifold M, we define the linearized vector field X' on the vector space T_pM around p as a linear map

$$X' : T_pM \longrightarrow T_pM$$
$$X'(v) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(DG_\lambda(v)\right)|_{\lambda=0},$$

where G_λ is the flow for X. This definition makes sense, as $G_\lambda(p) = p$ since p is a critical point for X, so

$$DG_\lambda : T_pM \longrightarrow T_pM,$$

for all λ. The curve $\lambda \mapsto DG_\lambda(v)$ for fixed v is a curve in the fixed vector space T_pM and so we can sensibly take the derivative with respect to λ and still get a value in the vector space T_pM.

In a coordinate system x on M, it can be shown that the matrix of X' is simply given by

$$[X'] = \left[\frac{\partial X^i}{\partial x^j}\right]$$

(Abraham and Marsden, 1985). We consider T_pM to be an approximation to the manifold near p: a small vector v in T_pM represents a point near p. $X'(v)$ is the first order linear approximation to the vector field at this approximate location. This is also suggested by a Taylor series argument in $\mathbb{R}^k$: we have $X(x+h) = X(x) + DX_x(h) + \cdots$; our X' in coordinates is essentially the derivative DX. At a critical point x_c, $X(x_c) = 0$, so the derivative approximates the vector field near x_c.

What is the behavior of X' on the tangent space to an invariant manifold through the critical point? Let S be the invariant manifold containing the critical point p. We must have

$$G_\lambda(S) \subseteq S$$

by definition, and since G_λ is a diffeomorphism, $G_\lambda(S)$ is also two-dimensional. Thus,

$$D(G_\lambda)_p\,(T_pS) = T_pS$$

for all λ and hence $X'(T_pS) \subseteq T_pS$. Thus, the tangent space T_pS to the invariant manifold of the vector field X at the critical point is an invariant linear subspace of X'. If $X'(T_pS)$ is one-dimensional, then T_pS contains a zero-eigenvalue eigenvector for X'.

If we have a linear operator on a vector space, $A : V \longrightarrow V$, what are the invariant linear subspaces of it? Linear algebra provides some information about this: if A is diagonalizable, every two-dimensional invariant subspace is the span of two eigenvectors of A.[1] If A has eigenvalues with multiplicity 1, the number of

[1] Concepts from (Hoffman and Kunze, 1971): if W is an invariant subspace of A so that $AW \subseteq W$, we can define $A|_W : W \longrightarrow W$ as the restriction of A to W. The minimal and characteristic polynomials for $A|_W$ divide the minimal and characteristic polynomials of A. The characteristic polynomial is defined as $\det(A - \lambda I)$, while the minimal polynomial is the unique polynomial $g(\lambda)$ of lowest degree which annhilates A (i.e. $g(A) = 0$) with 1 as the coefficient of the highest power of λ. The Cayley-Hamilton theorem says

possible W is quite constrained: the finite number of pairs of different eigenvectors. If A has an eigenvalue with multiplicity greater than 1, and hence an eigenspace of dimension 2 or more, then any combination of an eigenvector from this eigenspace and another eigenvector from another eigenspace can give rise to an invariant subspace: the number of possible invariant subspaces is now infinite (but there is still constraint on the orientations.)

Returning to the dynamical system, if X' is diagonalizable with eigenvalues of multiplicity 1, then there are a finite number of invariant linear subspace in T_pM. There are only a finite number of possible solution surface tangent planes through the critical point that are consistent with the dynamical system.

The local stable manifold theorem (Abraham and Marsden, 1985) indicates that the invariant subspace spanned by the set of eigenvectors of X' with negative eigenvalues gives rise to a unique invariant manifold of X near the critical point, called the stable manifold, S^+; the set associated with positive eigenvalues gives rise to the unique unstable manifold, S^-. Points on S^+ will move asymptotically close to p in the limit of increasing time; points on S^- will do the same in the limit of decreasing time. The dynamical system restricted to the stable/unstable manifold looks like a sink or a source near the critical point.

5.1.2 Hyperbolic Critical Points

To be able to easily examine other invariant manifolds of the flow near a critical point, we restrict our attention to a certain class of critical points. If a critical point has no eigenvalue with real part equal to zero, it is called a hyperbolic critical

---

that the characteristic polynomial of A annihilates A, so the minimal polynomial divides the characteristic polynomial. The roots of the characteristic polynomial give the eigenvalues of A, and the multiplicity of the roots gives the necessary dimensions of the eigenspaces if A is to have a basis of characteristic vectors, i.e., if A is to be diagonalizable. It turns out that A is diagonalizable if and only if the minimal polynomial has linear factors. Assuming A is diagonalizable, then the minimal polynomial of $A|_W$ must also be the product of linear factors, since it divides the minimal polynomial of A. Hence, $A|_W$ is diagonalizable as well. The eigenvalues of $A|_W$ are a subset of the eigenvalues of A. A characteristic vector of $A|_W$ in W is also a characteristic vector of A in V, so W is spanned by a pair of characteristic vectors of A.

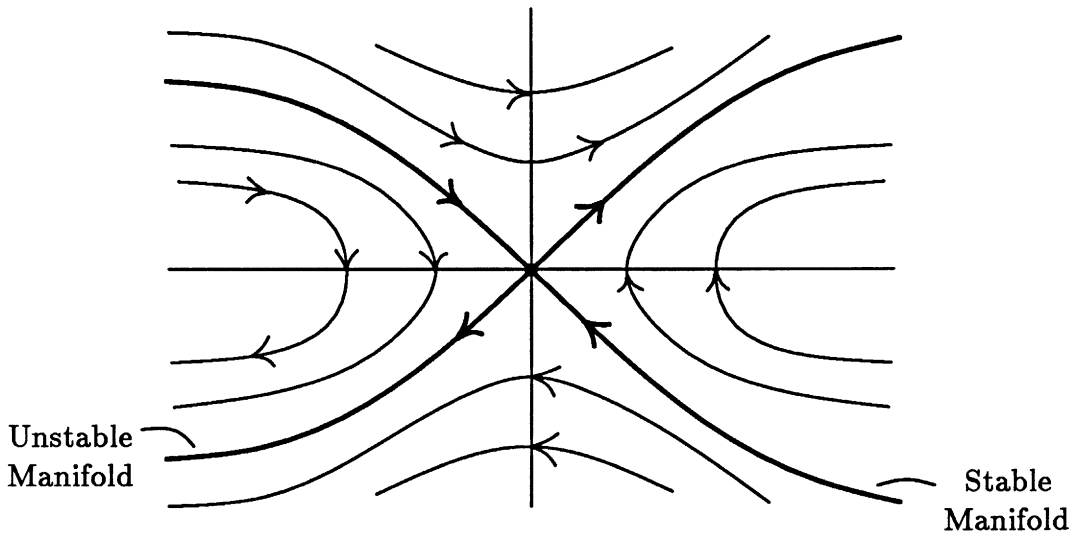


Figure 5.1. Pendulum hyperbolic critical point at $(\pi, 0)$.

point. Such a critical point is stable under generic perturbations of the vector field: that is, nearby vector fields in the space of all vector fields also have a hyperbolic critical point near this one, with the same number of eigenvalues with positive and negative real parts (Abraham and Marsden, 1985). This reflects the generic nature of matrices (corresponding to various possible X') with eigenvalues that are non-zero and not purely imaginary. In addition, the Grobman-Hartman theorem (Palis and de Melo, 1982) asserts that there is a neighborhood of a hyperbolic critical point in which the nonlinear flow lines are homeomorphic to the flow lines of the linearized dynamical system X' on T_pM. Thus, in a neighborhood of a hyperbolic critical point, the invariant (linear) subspaces of the dynamical system X' generated by pairs of eigenvectors of X' are homeomorphic to invariant manifolds of the nonlinear dynamical system X on M. Not only are we assured of the existence of the stable and unstable manifolds in this case, but any pair of eigenvectors, e.g. one associated with a positive real part eigenvalue and one associated with a negative real part eigenvalue, generate an invariant manifold in the neighborhood of the critical point. In the latter case, the flow restricted to the two-dimensional invariant manifold will be a saddle flow.

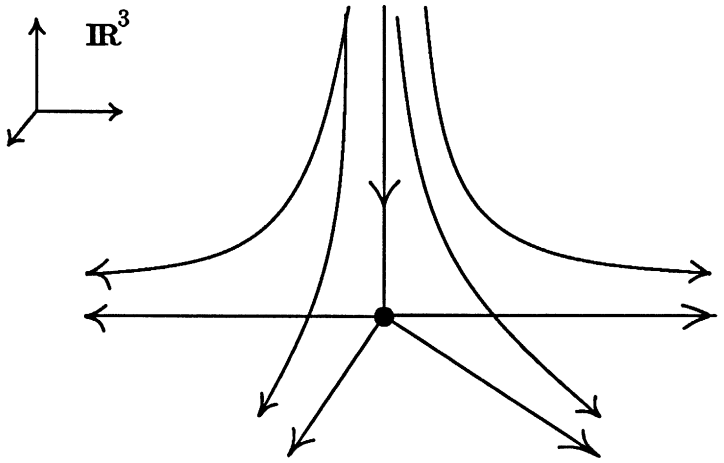


Figure 5.2. Hyperbolic critical point of a dynamical system in $\mathbb{R}^3$.

A simple example of a hyperbolic critical point is given by the pendulum example from Chapter 4 at $(\pi, 0)$, the top of the pendulum's arc. The stable and unstable manifolds are indicated in Figure 5.1—they are the unique manifolds containing only trajectories that all approach the critical point, either with positive time (unstable manifold) or negative time (stable manifold).

We can see this behavior in $\mathbb{R}^3$ as well. In Fig. 5.2 a hyperbolic critical point of a dynamical system (reminiscent of the wash from helicopter blades) is shown with two-dimensional unstable manifold (the ground) and one-dimensional stable manifold (vertical axis). The dynamical system restricted to the stable manifold is a source. There is another invariant manifold: a "saddle" invariant manifold consisting of the plane spanned by one eigenvector with a negative eigenvalue and one with a positive eigenvalue.

As we shall discuss in Section 5.2.3, critical points in $\mathcal{C}(\mathbb{R}^3, 2)$ in the space-invariant reflectance function case are not isolated: there will be a one parameter family of critical points as one moves along the projection direction due to the space invariant symmetry of the problem. This is a reflection of the well-known depth ambiguity. It makes sense in this case to reduce the dimensionality of the problem, effectively by ignoring the depth coordinate, to make the critical point an isolated hyperbolic point.

5.2 Image Critical Points

Both Horn (Horn, 1975) and Bruss (Bruss, 1980) recognized the potential importance of critical points in the image intensities. As Horn pointed out, isolated image intensity critical points are likely to be due to critical points in the reflectance function; if these are also isolated, then the potential surface normal directions are limited to just these critical points on the reflectance function. As he indicated, one cannot draw out a solution surface from this known point in $\mathcal{C}(\mathbb{R}^3, 2)$ with characteristic trajectories because the vector field is stationary there. His solution was to make a spherical cap through the critical point with an arbitrarily (large) chosen radius and consistent with the presumed normal. Solution curves could then be extended from this cap. He showed empirical evidence that these solution curves did not change much as the radius was changed, suggesting a certain stability of the solution trajectories.

Bruss, although working towards uniqueness of solutions in a very restricted domain of reflectance functions, those with elliptical constant brightness contours in p, q space, made mention and use of the ideas of the stable and unstable manifolds associated with a critical point to demonstrate existence of solutions for her particular problem. We expand upon this.

5.2.1 Location of Critical Points

When are critical points of the image actually due to critical points in the reflectance function? In the case of a space invariant reflectance function, R depends only on the orientation part of $\mathrm{i} : S \longrightarrow \mathcal{C}(\mathbb{R}^3, 2)$. The orientation part of i is effectively the map from points on the surface to the normal directions in space and so is related to the Gauss map, N.

In the study of surface shape, one examines the Gauss map $N : i(S) \longrightarrow S^2 \subset \mathbb{R}^3$ which maps points on the surface embedded in $\mathbb{R}^3$ to unit normals representing the orientation of the surface (Figure 5.3). The Weingarten map or shape operator

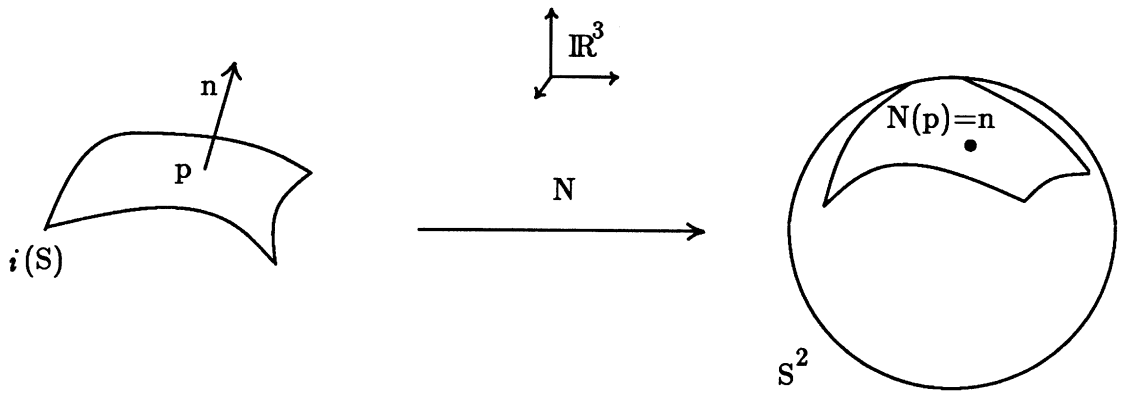


Figure 5.3. Gauss map. It maps surface normals of $i(S)$ to $S^2 \subset \mathbb{R}^3$.

is the derivative DN of this map. Since both the tangent plane to the embedded surface and the tangent plane to the embedded sphere are perpendicular to the surface normal, we can consider $DN : Ti(S) \longrightarrow Ti(S)$.[2] If we examine DN at a point on the surface in $\mathbb{R}^3$, it tells us how the orientation of the surface begins to change as we begin to move in different directions on the surface; for example the eigenvalues and eigenvectors of DN are the principal curvatures and principal directions along the surface. Here again, the tangent plane to S at x (in $\mathbb{R}^3$) is used as an approximation for the surface, and the value of DN tells approximately how the orientation has changed as one moves approximately to a nearby point on the surface. Thus we expect the characteristic vector field X on $\mathcal{C}(\mathbb{R}^3, 2)$ to give us information consistent with DN in the direction of the characteristic trajectories: the orientation component of X is related to $DN(X_{sp})$, where X_{sp} is the space component of X.

Let us assume we have the embedding $i : S \to \mathbb{R}^3$. We also have the lifted embedding $\mathrm{i} : S \to \mathcal{C}(\mathbb{R}^3, 2)$ which takes $p \in S$ to the tangent plane $T_pS \subset T_p\mathbb{R}^3$. Since $\mathcal{C}(\mathbb{R}^3, 2) \simeq \mathbb{R}^3 \times G(2,3)$, where $G(2,3)$ is the set of all two-dimensional subspaces of $\mathbb{R}^3$, we can divide i into two components: $\mathrm{i}(p) = (i(p), n(p))$, where $i(p) \in \mathbb{R}^3$ is the space component of the embedding and $n(p) \in G(2,3)$ is the orientation component. We have a natural map $\chi : S^2 \to G(2,3)$ given by mapping

[2] A modern view of the shape map can be found in (Thorpe, 1979).

a normal vector to the plane perpendicular to it. We can now write $\chi \circ N \circ i(p) = n(p)$, and if we define $\tilde{R} \triangleq R \circ \chi$, we can write the image irradiance equation as

$$E \circ \pi^I \circ i = \tilde{R} \circ N \circ i.$$

Taking the derivative and looking for critical points in the image, we have

$$DE \circ D(\pi^I \circ i) = D\tilde{R} \circ DN \circ Di,$$

where DN is the Weingarten map. If we assume we are on the interior of the image, then $D(\pi^I \circ i)$ is an invertible linear map, and so the image will have a critical point, that is $DE = 0$, if and only if

$$0 = D\tilde{R} \circ DN,$$

since i is a diffeomorphism between S and $i(S)$. When does this condition occur?

This condition certainly occurs when $DN = 0$ at a point. This is a locally flat point, a point where both principle curvatures are zero.

The condition can also occur when DN has rank one, i.e. when DN has just one non-zero eigenvalue. This occurs at parabolic points that are not locally flat. We get a critical point in the brightness if the principle direction, spanning the one-dimensional range of DN, is in the null space of $D\tilde{R}$; otherwise phrased, the principle direction must be perpendicular to the gradient of $\tilde{R}$.

If this latter situation occurs, is it a stable occurrence? Could a small perturbation of the reflectance function remove this occurrence? Since in general parabolic points form curves on the surface, there will be a one dimensional curve of non-zero eigenvalue principle directions on S parameterized by the parabolic curve, one at each point of the curve. There is also a one dimensional set of null directions for $D\tilde{R}$ parameterized by the parabolic curve on S. At each point on the parabolic curve, there is only a two-dimensional vector space of possible tangent vectors since T_pS has dimension two, and hence only a one-dimensional space of tangent orientations (parameterized, for example, by angle). We can locally model this one dimensional

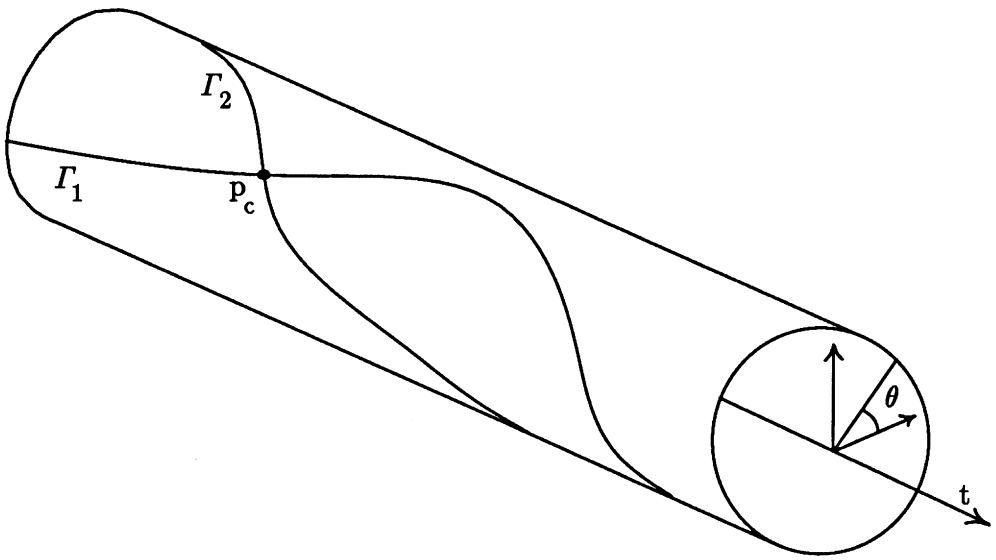


Figure 5.4. Tube of possible orientations of principle direction for DN (Γ_1) and null direction for $D\tilde{R}$ (Γ_2) along a parabolic line, $\alpha(t)$. An orientation at time t is defined as a point on the unit circle with t as its center; this picks an angle, θ, for the orientation of the direction vector. The intersection, p_c, of the two curves gives a critical point of the image on the parabolic line.

set of orientations as a circle, and consider the parabolic line together with the possible orientation directions as a two dimensional tube $\mathbb{R} \times S^1$: one coordinate gives distance along the parabolic curve, the other coordinate (around the circle) gives an orientation (Figure 5.4). The null directions of $D\tilde{R}$ and the non-zero eigenvalue principle directions form curves that sit on the tube: at each point of the parabolic curve, they pick out two orientations on the circle. If there is a point on the surface where these two coincide, we can consider the curves on the tube to have intersected. This will be a point where the principle direction falls in the null space of the gradient of R, and hence is the critical point of interest. Since this is an intersection of one dimensional curves on a two dimensional surface, the intersection at a point is transversal, i.e. stable to small smooth changes in the parabolic line location or the principal directions or the reflectance function.

Koenderink and Van Doorn (Koenderink and Van Doorn, 1979) have also described this kind of point on the surface where the image has critical points. If we consider the Gauss map, N, as laying the surface $i(S)$ onto the unit sphere, then the

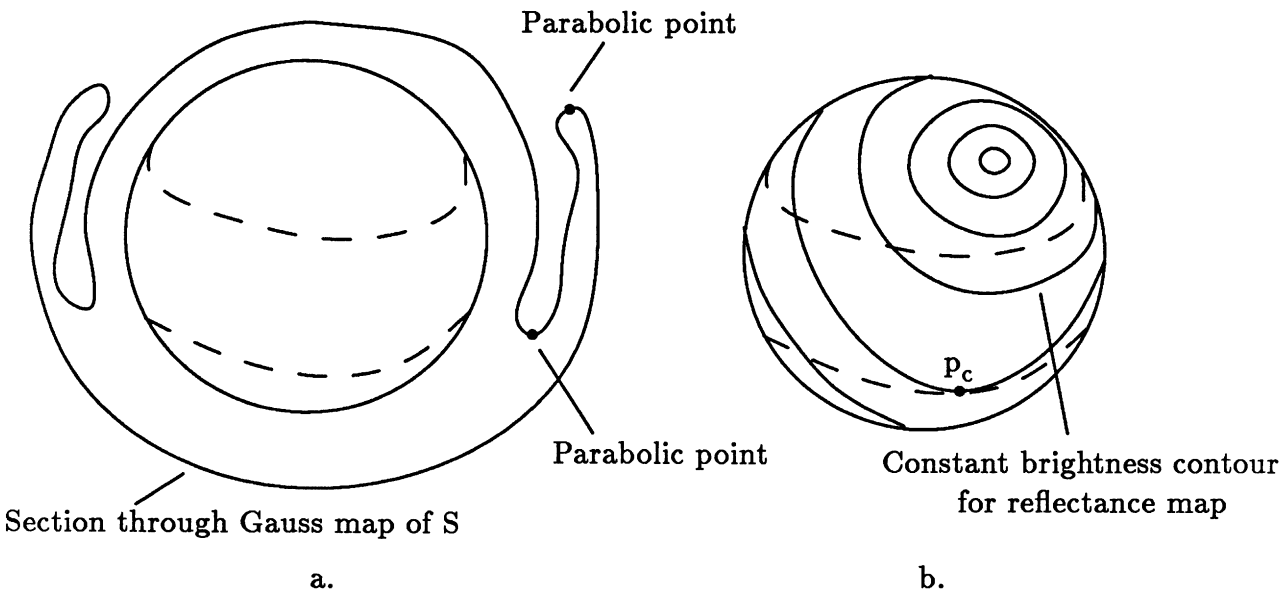


Figure 5.5. The surface S mapped to the Gaussian sphere: parabolic lines (dashed lines) are folds in this map. a. An exploded section through the Gauss map of a surface onto the sphere—the fold lines are lines of parabolic points, and there is a triple thickness of surface mapped between the folds. b. Relationship between the Gauss map and reflectance function constant brightness contours (solid lines on the sphere). The dashed lines are folds with a triple thickness of surface mapped between them; p_c is a parabolic point with maximum brightness on a fold, and so will be a saddle critical point for the image.

parabolic lines on the surface are the lines of the folds for the image of the surface (Figure 5.5 a). If we overlay the reflectance function on top of the Gaussian sphere, then a maximum in brightness along the parabolic line, or fold, will correspond to a critical point in the image even though the local maxima of the reflectance function will be elsewhere: effectively, the fold has trapped the brightness values (Figure 5.5 b). If we consider a curve $\alpha : \mathbb{R} \longrightarrow S$ on the surface crossing a parabolic curve and consider the corresponding path on the Gaussian sphere which connects the surface normals along α, this path on the Gaussian sphere touches the parabolic fold and does not cross it. If the path runs in a certain direction, the path will actually double back on itself, providing one critical direction for the brightness values. This will be true at any parabolic point. If there is a maximum or minimum of the brightness along the parabolic fold, this will provide the other critical direction to make a critical point for brightness on the fold. Depending on the relationship between

the constant brightness contours on the Gaussian sphere and the parabolic line, this critical point in the image can be a maximum, a minimum, or a saddle.

Since the characteristic vector field in the usual coordinates has the form

$$X = \begin{bmatrix} R_p \\ R_q \\ pR_p + qR_q \\ E_x \\ E_y \end{bmatrix}$$

these parabolic image critical points will not be critical points in the image dynamical system, since DR is not zero and so the vector field is not zero. From the form of the vector field X, the space part of X is non-zero at these points while the orientation part goes through a zero; there is only a minor influence of these points on the flow picture in space.

Finally, if DN is of full rank, then necessarily a critical point in the image must be due to a critical point in $\tilde{R}$, since DN spans the effective domain of $D\tilde{R}$, and so $D\tilde{R} \circ DN = 0$ implies $D\tilde{R} = 0$.

In summary, we have several kinds of critical points: 1) due to $DN = 0$, i.e. locally flat points; 2) at certain locations on parabolic lines; and 3) at points where the reflectance function has a critical point. Critical points in the image that are due solely to critical points in the reflectance function will be called *good* critical points; all others will be called *bad.* Note also that the bad critical points in the image are almost always not due to reflectance function critical points, and so the brightness of these image critical points will not be the same as those due to reflectance function critical points. In the simplest case of a reflectance function with single maximum (e.g. a Lambertian reflectance function) generically only the good critical points have the brightest value of the reflectance function.

We also ignore in this analysis the role of shadows and shadow boundaries. The most egregious omission is the role of the self-shadow boundary which can be considered to be due to the reflectance function: usually there is an extended set of

orientations for which the reflectance is at a minimum because patches of the surface are turned completely away from the light source. We will focus exclusively on images (or portions of images) without self-shadows.[3]

5.2.2 Good Critical Points

If we are at a good critical point, what kind of critical point can it be? We can look at the second derivative of the image irradiance equation to get some ideas about this.

We will implicitly assume a coordinate system for the calculations–this will allow expressions like $D^2 f$ to make sense on their own, as the matrix of second derivatives with respect to the coordinates.[4]

For conveniance, we define $\tilde{N} = N \circ i$. Taking derivatives of the image irradiance equation as originally expressed, we have

$$E \circ \pi^I \circ i(p) = \tilde{R} \circ \tilde{N}(p)$$

$$DE_{\pi^I \circ i(p)} \circ D(\pi^I \circ i)_p(\mathbf{v}) = D\tilde{R}_{\tilde{N}(p)} \circ D\tilde{N}_p(\mathbf{v})$$

$$D^2E(D(\pi^I \circ i)(\mathbf{u}), D(\pi^I \circ i)(\mathbf{v})) + DE \circ D^2(\pi^I \circ i)(\mathbf{u}, \mathbf{v}) =$$

$$D^2\tilde{R}(D\tilde{N}(\mathbf{u}), D\tilde{N}(\mathbf{v})) + D\tilde{R} \circ D^2\tilde{N}(\mathbf{u}, \mathbf{v}).$$

At a good critical point both DE and $D\tilde{R}$ are 0, so we have

$$D^2E(D(\pi^I \circ i)(\mathbf{u}), D(\pi^I \circ i)(\mathbf{v})) = D^2\tilde{R}(D\tilde{N}(\mathbf{u}), D\tilde{N}(\mathbf{v}))$$

---

[3] Note that the self-shadow line is a constant brightness contour; it is the "dark side" of the self-shadow line that is not included here.

[4] One could define second derivatives independently of coordinates in two ways: one way is to consider derivative maps of real functions as maps from the tangent bundle TM to $\mathbb{R}$. The second derivative is essentially the derivative of this map. In coordinates, the most interesting component ends up being a linear combination of the second derivative in coordinates and the first derivative in coordinates. Another way to deal with second derivatives on surfaces is to specify a particular way of tying the tangent planes together: this is called a connection, and effectively tells how to take derivatives of arbitrary tensors on the surface. For our purposes we do not need either construction.

In the interior of the image, $\pi^I \circ i$ will be a local diffeomorphism; if we are not on a parabolic line, $D\tilde{N} = D(N \circ i)$ will also be a local diffeomorphism, and hence D^2E and $D^2\tilde{R}$ will be of the same rank and type (i.e. same numbers of positive and negative eigenvalues).

To see this, we can consider the different derivatives in this last expression as matrices. "D^2E" is a bilinear function of its arguments, so if we use a matrix representation and let $A = D^2E$, $B = D^2R$, $P = D(\pi^I \circ i)$, and $Q = D\tilde{N}$, then we have for all $\mathbf{u}$ and $\mathbf{v}$ in the tangent space to S at p:

$$\mathbf{u}^T P^T A P \mathbf{v} = \mathbf{u}^T Q^T B Q \mathbf{v},$$

and hence

$$P^T A P = Q^T B Q.$$

We can write

$$A = (QP^{-1})^T B (QP^{-1}).$$

In Appendix A5.1 to this chapter, we show that this means A and B have the same number of positive and negative eigenvalues, so that D^2E is positive (negative) definite if and only if D^2R is. We will assume from here on that D^2R is full rank, the usual and generic case.

If we are at a good critical point which is a maximum of the reflectance function so that D^2R is negative definite, D^2E will also be negative definite, and hence the critical point in the image will be a local maximum. Such a critical point is surrounded by closed constant brightness contours, and the presence of such a critical point in a region can be strongly suspected if there are closed constant brightness contours in the region which are increasing in value towards the inside. In a sense, a maximum brightness point has an image "signature" that extends around the point. We suspect that such a signature without the actual presence of the critical point (e.g. the critical point is obscured) may be sufficient to get strong limits on possible

solution surfaces. (Brooks (Brooks, 1982) discusses an example of an annular image of a hemisphere lit from the viewing direction with the hole of the annulus placed over the critical point.)

If the reflectance function has a single maximum point (e.g. the Lambertian case) this will be the only critical point in the reflectance function, and will generate the only good critical points in the image. In this case, any saddles in the image will be due to parabolic points. With more general reflectance functions, saddles in the reflectance function can lead to saddles in the image at good critical points.

5.2.3 The Linearized and Reduced Image Dynamical System

The good critical points generate critical points of the image dynamical system. A solution surface consistent with a smooth image patch containing a good critical point should be a smooth surface made up of characteristic trajectories; this means we are interested in locally invariant manifolds of the image dynamical system which include the critical point. As indicated in Section 5.1.1, an important technique for studying the behavior of a dynamical system around a critical point is to look at the linearization of the dynamical system at the critical point.

We can examine in detail the behavior of an image dynamical system under the conditions used to derive a coordinate expression for X in Chapter 4: we assume orthographic projection and a space-invariant reflectance function. As derived in Chapter 4, this leads to a characteristic vector field on the image interior of the form

$$X = \begin{bmatrix} R_p \\ R_q \\ pR_p + qR_q \\ E_x \\ E_y \end{bmatrix}$$

using the standard rectilinear (x, y, z, p, q) coordinate system on $\mathbb{R}^3$ and $\mathcal{C}(\mathbb{R}^3, 2)$, with image projection given by $\pi^I(x, y, z) = (x, y)$. We can calculate the linearization

of this vector field at a critical point p_c where $0 = R_p = R_q = E_x = E_y$ as

$$X' = \begin{bmatrix} 0 & 0 & 0 & R_{pp} & R_{pq} \\ 0 & 0 & 0 & R_{qp} & R_{qq} \\ 0 & 0 & 0 & pR_{pp} + qR_{qp} & pR_{pq} + qR_{qq} \\ E_{xx} & E_{xy} & 0 & 0 & 0 \\ E_{yx} & E_{yy} & 0 & 0 & 0 \end{bmatrix},$$

where we use the notation E_{xx} to mean $(E \circ \pi^C)_{xx}$, where π^C is the image projection from $\mathcal{C}(\mathbb{R}^3, 2)$,[5] and we have substituted the values $R_p = R_q = 0$ after taking the derivative.

By inspection we can see that the matrix X' is singular, meaning X' has at least one zero eigenvector. It is the vector $(0, 0, 1, 0, 0)^T$ and reflects the translation invariance along the projection direction. This means we are not at a hyperbolic critical point, which would require non-zero eigenvalues for the linearized system. Although there are some results on the existence and uniqueness of invariant manifolds for non-hyperbolic critical points, the Grobman-Hartman theorem mentioned in Section 5.1.2 suggests it would be convenient to make the problem into a dynamical system with a hyperbolic critical point. Among other difficulties, a non-hyperbolic critical point is not generic: a small perturbation can radically change the character of the dynamical system.

We can convert our image dynamical system by focusing on the (x, y, p, q) coordinates and leaving out the z coordinate. Effectively, the ordinary differential equation defining the dynamical system,

$$\dot{\mathbf{x}} = X,$$

[5] This small confusion in notation turns out to be correct here because the image projection, $\pi^I(x, y, z) = (x, y)$, in coordinates makes $E_{xx} = (E \circ \Pi^C)_{xx}$, where the first set of derivatives are with respect to the image coordinates and the second set are with respect to coordinates on $\mathcal{C}(\mathbb{R}^3, 2)$.

does not depend on z on the right hand side; we can solve for (x, y, p, q) first, and then integrate the $\dot{z}$ equation to find the solution for z.[6] In coordinates the reduced dynamical system is

$$\tilde{X} = \begin{bmatrix} R_p \\ R_q \\ E_x \\ E_y \end{bmatrix},$$

with

$$\tilde{X}' = \begin{bmatrix} 0 & 0 & R_{pp} & R_{pq} \\ 0 & 0 & R_{qp} & R_{qq} \\ E_{xx} & E_{xy} & 0 & 0 \\ E_{yx} & E_{yy} & 0 & 0 \end{bmatrix}.$$

We will call the new space on which $\tilde{X}$ is a vector field $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$.

5.2.4 Eigenvalues of $\tilde{X}'$

What are the eigenvalues and eigenvectors of $\tilde{X}'$? Examining the matrix form of $\tilde{X}'$ we have:

$$\tilde{X}' = \begin{bmatrix} \mathbf{0} & D^2R \\ D^2E & \mathbf{0} \end{bmatrix},$$

where D^2E (shorthand here for $D^2(E \circ \pi^C)$) and D^2R are both 2×2 matrices representing second derivatives of $E \circ \pi^C$ and R with respect to space and orientation coordinates respectively. We can consider the eigenvalues of

$$\tilde{X}'\tilde{X}' = (\tilde{X}')^2 = \begin{bmatrix} D^2R \circ D^2E & 0 \\ 0 & D^2E \circ D^2R \end{bmatrix}:$$

we have

$$\begin{aligned} 0 &= \det((\tilde{X}')^2 - \gamma I) = \det(D^2R \circ D^2E - \gamma I)\det(D^2E \circ D^2R - \gamma I) \\ &= \left(\det(D^2E \circ D^2R - \gamma I)\right)^2, \end{aligned}$$

[6] It is also possible to do this in an invariant manner by using the symmetry group corresponding to flow down the projection direction: the dynamical system is invariant to this symmetry, and so the order of the system can be reduced by one. We have a new dynamical system defined on the space $I \times G(2,3)$ where $G(2,3)$ is the space of two-dimensional subspaces of a three-dimensional vector space, and I can be considered either as the space of projection fibres or as the image.

so the eigenvalues of $(\tilde{X}')^2$ are the same as the eigenvalues of $D^2E \circ D^2R$. We also have

$$\det((\tilde{X}')^2 - \gamma I) = \det(\tilde{X}' - \lambda I)\det(\tilde{X}' + \lambda I),$$

where $\lambda^2 = \gamma$. If γ is a characteristic value for $(\tilde{X}')^2$, then $\pm\lambda$ are possible characteristic values for $\tilde{X}'$; one of them must be a characteristic value. It turns out (see below) that both actually are characteristic values of $\tilde{X}'$. The 2×2 submatrix $A = D^2R \circ D^2E$ has a quadratic characteristic polynomial, and the characteristic equation for $(\tilde{X}')^2$ is

$$0 = (\gamma^2 + \text{trace}(A)\gamma + \det(A))^2.$$

This means the characteristic values for $\tilde{X}'$ must be the four roots of

$$0 = \lambda^4 + \text{trace}(A)\lambda^2 + \det(A),$$

and so come in pairs as the two square roots of each characteristic eigenvalue for A.[7]

As discussed in Section 5.2.2, D^2R and D^2E have the same number of positive and negative eigenvalues. In the case where both matrices are definite, we show in Appendix A 5.3 that $A = D^2R \circ D^2E$ will have positive real eigenvalues, and hence $\tilde{X}'$ will have real eigenvalues, generically distinct. In the case where D^2R is indefinite, we may generically have negative roots for A, yielding purely imaginary eigenvalues for $\tilde{X}'$ and the critical point may therefore not be hyperbolic. We will focus our attention on *very good* critical points, those due only to definite critical points of the reflectance function: the usual case of the maximum of reflectance function is a very good critical point. The analysis of invariant manifolds of the image dynamical system at indefinite critical points of the reflectance function is an area of future research.

[7] The coordinate form of $\tilde{X}$ suggests a four-dimensional Hamiltonian dynamical system with (x, y, p, q) coordinates as Darboux coordinates for the 2-form $\omega = \mathrm{d}x \wedge \mathrm{d}p + \mathrm{d}y \wedge \mathrm{d}q$ and the image function H as Hamiltonian (Abraham and Marsden, 1985). The distribution of eigenvalues is the result of the matrix $\tilde{X}'$ being infinitessimally symplectic.

In general (i.e. generically in some sense) at a very good critical point there will be four distinct eigenvalues of $\tilde{X}'$ with four associated independent eigenvectors. Using the results os Section 5.11, pairs of these eigenvectors generate possible invariant tangent planes, six in all. If there are symmetries in the system (for example the image of an object symmetric around the viewing axis lit from the viewing direction) we can get eigenvalues with multiplicity two, which potentially means a two dimensional space of eigenvectors. From the stable manifold theorem there will still be unique stable and unstable manifolds, but now there may be an infinite number of possible saddle manifolds and hence solution surfaces because there may be an infinite number of subspaces spanned by pairs of eigenvectors. We will concentrate on the generic case without symmetries.

Not all possible pairs of eigenvectors will generate tangent planes in $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$ that could possibly correspond to a real two dimensional surface embedded in three dimensions. We know that $\tilde{X}'$ has the form

$$\tilde{X}' = \begin{bmatrix} 0 & D^2R \\ D^2E & 0 \end{bmatrix},$$

where D^2R and D^2E are essentially the second derivative matrices of the reflectance and image at the critical points. If $\mathbf{u}$ is an eigenvector for $\tilde{X}'$, we have

$$\begin{aligned} \tilde{X}'\mathbf{u} &= \lambda\mathbf{u} \\ \begin{bmatrix} 0 & D^2R \\ D^2E & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}_{sp} \\ \mathbf{u}_{or} \end{bmatrix} &= \\ \begin{bmatrix} D^2R\mathbf{u}_{or} \\ D^2E\mathbf{u}_{sp} \end{bmatrix} &= \begin{bmatrix} \lambda\mathbf{u}_{sp} \\ \lambda\mathbf{u}_{or} \end{bmatrix}. \end{aligned}$$

Now let $\mathbf{u}' = (-\mathbf{u}_{sp}, \mathbf{u}_{or})^T$:

$$\begin{aligned} \tilde{X}'\mathbf{u}' &= \begin{bmatrix} D^2R\mathbf{u}_{or} \\ -D^2E\mathbf{u}_{sp} \end{bmatrix} \\ &= \begin{bmatrix} \lambda\mathbf{u}_{sp} \\ -\lambda\mathbf{u}_{or} \end{bmatrix} \\ &= -\lambda\mathbf{u}', \end{aligned}$$

so that $-\lambda$ is also an eigenvalue, and $\mathbf{u}'$ is its associated eigenvector.

Although $\mathbf{u}$ and $\mathbf{u}'$ may be independent as eigenvectors associated with distinct eigenvalues in the four-dimensional reduced space $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$, the space parts of the eigenvectors $\mathbf{u}$ and $\mathbf{u}'$ do not project to independent vectors in $\mathbb{R}^2$ and hence the subspace spanned by $\mathbf{u}$ and $\mathbf{u}'$ does not correspond to a two dimensional surface in $\mathbb{R}^3$. This eliminates two of the six pairings of eigenvectors.

By the Grobman–Hartman theorem, the remaining four possible invariant tangent spaces correspond to four possible invariant manifolds of $\tilde{X}$ through the critical point that could correspond to solution surfaces. One will be the stable manifold with sink flow type when restricted to the invariant surface; another will be the unstable manifold, with source flow type on the invariant manifold; the two remaining invariant manifolds will have flows that are of saddle type.

There is one other possible constraint that may eliminate the saddle invariant manifolds as candidates. For a solution surface $S \subset \tilde{\mathcal{C}}(\mathbb{R}^3, 2)$ to solve the shape from shading problem at hand, we must have $R(p, q) = E(x, y)$ everywhere on it; this can be written using the image dynamical system function as $H(p) = 0$ for all $p \in S$. The characteristic trajectories are curves on which H is constant, but not necessarily zero. In the case of the stable and unstable manifolds, all the trajectories in them approach the critical point p_c asymptotically, and $H(p_c) = 0$ if we have matched the critical point of the image to the correct critical point of the reflectance function. This means $H = 0$ automatically on the unstable and stable manifolds. However, there are only four trajectories in the saddle case (referring to the image dynamical system type restricted to the invariant manifold) that actually approach the critical point; it is conceivable that the other characteristic trajectories making up the saddle invariant solution surface have H non-zero along their lengths. If this were the case, these invariant solution surface candidates could be rejected as possible solutions to the shape from shading problem.

However, dynamical saddle invariant solutions do occur and cannot be discarded out of hand as possible solutions. As we shall see in the next section, for very

good critical points at which the true surface is saddle-shaped in space, the image dynamical system restricted to the true surface is of saddle type. If, as in this case, the image dynamical system restricted to the true solution surface is of saddle type at the critical point, then H must be 0 on all the constituent characteristic trajectories even though they do not all approach the critical point.

5.2.5 Solution Surfaces Near a Good Critical Point

Given a solution surface with a particular Weingarten map DN describing the local surface behavior near a critical point, what is the relationship between the surface shape, the reflectance function critical point type, and the type of characteristic flow generated on the surface? The characteristic vector field restricted to the two-dimensional invariant solution surface defines a two-dimensional dynamical system on the surface. We are interested in the two-dimensional flow on the solution surface near critical points of the system: is it a source, sink, or saddle?

As in Section 5.2.4, we can decompose the vector field $\tilde{X}$ into two pieces,

$$\tilde{X} = \begin{bmatrix} X_{sp} \\ X_{or} \end{bmatrix},$$

where X_{sp} is the space component and X_{or} is the orientation component. We can similarly divide $\tilde{X}'$, the linear approximation to the vector field at a critical point into space and orientation components

$$\tilde{X}' = \begin{bmatrix} X_{sp}{}' \\ X_{or}{}' \end{bmatrix}.$$

From the form of $\tilde{X}'$ derived in Section 5.2.1, we get

$$X_{sp}{}'(\mathbf{u}_{sp}, \mathbf{u}_{or}) = D^2R \circ \mathbf{u}_{or}.$$

This represents the linear approximation to the space part of the vector field $\tilde{X}$ at an approximate displacement $(\mathbf{u}_{sp}, \mathbf{u}_{or})$ from the critical point p in $\tilde{C}(\mathbb{R}^3, 2)$.

In Section 5.2.1 we looked at the relationship between the Gauss map N and the orientation component of the lifted embedding $\mathrm{i} : S \to \mathcal{C}(\mathbb{R}^3, 2)$: we saw that $\mathrm{i}(p) = (i(p), n(p)) \in \mathbb{R}^3 \times G(2,3)$ and $\chi(N(i(p))) = n(p)$ is the map relating the surface normal $N(i(p))$ to the tangent plane $n(p)$ at p. We must have $D\chi \circ DN \circ Di = Dn$. If we assume the embedding $i : S \to \mathbb{R}^3$ is the inclusion and let $\mathrm{i} : S \to \tilde{\mathcal{C}}(\mathbb{R}^3, 2)$ instead of $\mathcal{C}(\mathbb{R}^3, 2)$ (effectively we drop the depth coordinate), then we can write $D\mathrm{i}(\mathbf{u}_{sp}) = (\mathbf{u}_{sp}, Dn(\mathbf{u}_{sp})) = (\mathbf{u}_{sp}, \mathbf{u}_{or})$, so

$$\mathbf{u}_{or} = Dn(\mathbf{u}_{sp}) = D\chi \circ DN(\mathbf{u}_{sp}).$$

The $\mathbf{u}_{or}$ component tells how the orientation direction changes on the surface while moving on the surface in the $\mathbf{u}_{sp}$ direction—this is very nearly what the Weingarten map tells us too. Combining this with the previous expression for ${X_{sp}}'$ we get

$${X_{sp}}'(\mathbf{u}_{sp}, \mathbf{u}_{or}) = D^2R \circ D\chi \circ DN \circ \mathbf{u}_{sp}.$$

Considering $\mathbf{u}_{or}$ as essentially determined by the Gauss map N for the fixed embedded surface S and $\mathbf{u}_{sp}$, as a vector field on the surface S we have

$${X_{sp}}'(\mathbf{u}_{sp}) = D^2R \circ D\chi \circ DN \circ \mathbf{u}_{sp}.$$

This tells us that the approximate space component of the characteristic vector field restricted to a solution surface near a critical point is determined by the derivative map $D^2R \circ D\chi \circ DN$. We are interested in finding the type of the two-dimensional dynamical system on S determined by ${X_{sp}}'$.

By picking local coordinates on the Grassman manifold of two-dimensional planes in three dimensions, $G(2,3)$, to match coordinates on the two-dimensional sphere S^2 (e.g. both using central projection onto the same plane $z = -1$ to give standard (x, y, z, p, q) coordinates), we can make $D\chi$ become the identity matrix in these coordinates; we will let B be the matrix for D^2R, C be the matrix for DN, and

$A = BC$ be the matrix for $D^2R \circ D\chi \circ DN$. The eigenvalues for A will tell us the type of the dynamical system on S.

A is the product of two symmetric matrices but itself need not be symmetric. For an asymmetric matrix, eigenvalues may be complex; when does this happen to A? In Appendix A5.2 to this chapter we show that if either B or C is definite, then A has real eigenvalues and has a full set of eigenvectors. The only case in which $A = D^2R \circ DN$ will not always have real eigenvalues is the case where we have a saddle in the reflectance function and a saddle in the surface.

Assume A has real eigenvalues. If γ_1 and γ_2 are the eigenvalues of $B = D^2R$, and κ_1 and κ_2 are the eigenvalues (curvatures) of $C = DN$, then

$$\det(A) = \gamma_1\gamma_2\kappa_1\kappa_2 = \alpha_1\alpha_2,$$

where α_1 and α_2 are eigenvalues for A. If $\gamma_1\gamma_2 > 0$, then the sign of $\det(A)$ equals the sign of $\kappa_1\kappa_2$. Thus, if R has a maximum at the critical point so that D^2R is negative definite, and the shape of the solution surface is either convex or concave, giving same sign principle curvatures, then the space part of the characteristic flow restricted to the two dimensional spatial solution surface is either a source or sink since the eigenvalues of A will have the same sign. Similarly if the solution surface is a saddle at the critical point, then the space part of the characteristic flow will be a saddle flow.

We also have

$$\text{trace}(A) = \text{trace}(BC) = \text{trace}(BP\Lambda P^T) = \text{trace}(P^TBP\Lambda),$$

where P is an orthonormal matrix of eigenvectors for $C = DN$ and Λ is a diagonal matrix of eigenvalues for C, so P diagonalizes C. Let us assume the diagonal entries of P^TBP are d_{11} and d_{22}; the eigenvalues of C are the principle curvatures κ_1 and κ_2. Then

$$\text{trace}(A) = \alpha_1 + \alpha_2 = \kappa_1 d_{11} + \kappa_2 d_{22}.$$

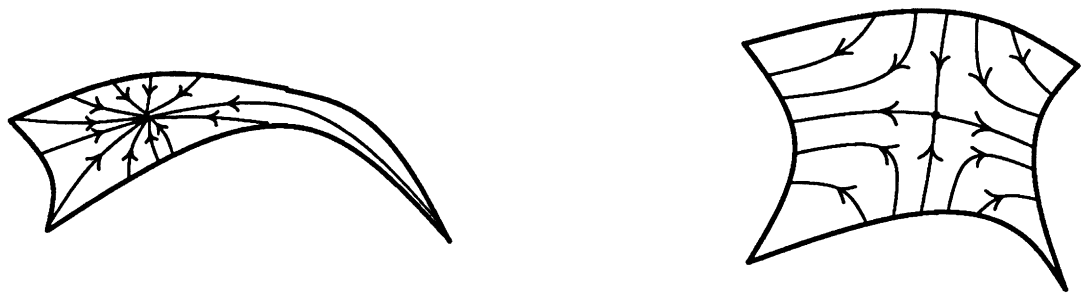


Figure 5.6. Patches of surfaces lit from above. In the middle of each patch we have a critical point due to a maximum in the reflectance function; the space part of the dynamical system restricted to the (invariant) surface patch is shown. a. Convex surface patch. b. Saddle shaped surface patch.

If $D^2R = B$ is negative definite so is P^TBP, and hence d_{11} and d_{22} are both less than zero: we have

$$(1 \quad 0)P^TBP\begin{pmatrix}1\\0\end{pmatrix} = d_{11} < 0,$$

and similarly for d_{22}. In this case, if $DN = C$ is negative (positive) definite, then we know that trace(A) is greater than (less than) zero. The eigenvalues α_1 and α_2, having the same sign by the previous argument, are positive (negative) and the characteristic flow must be a source (sink). We get similar results if D^2R is positive definite at the critical point, suggesting a minimum in the reflectance function.

Assuming A has real eigenvalues (for example, if DN is definite), a saddle in the reflectance function determines a saddle flow on the invariant manifold by the sign of the product of the eigenvalues of A. However, in the case where DN is not definite we are not assured that the eigenvalues of A are real, and the flow type is undetermined by these simple arguments.

Figure 5.6 gives two examples for surfaces lit from above with very good critical points due to reflectance function maxima: Figure 5.6a shows a convex surface, on which the image dynamical system restricts to a sink, and Figure 5.6b shows a saddle shaped surface, on which the image dynamical system restricts to a saddle dynamical system.

We can summarize the results in the following table. Entries in the table give the flow type of the two-dimensional characteristic system restricted to the two-dimensional surface.

		Form of DN		
		Positive Definite	Negative Definite	Saddle
Form of	Positive Definite	Source	Sink	Saddle
D^2R	Negative Definite	Sink	Source	Saddle
	Saddle	Saddle	Saddle	Undetermined

Given an image due to a generic real surface and a known very good critical point (i.e. due to a definite (positive or negative) reflectance function critical point), there are at least two and at most four smooth surfaces consistent with an image patch containing a good critical point. In the usual case of a very good critical point due to a maximum in the reflectance function, the invariant solution surface will be the stable (sink type flow) or unstable (source type flow) manifold of the dynamical system if the surface is convex or concave; the invariant solution will be a saddle-flow invariant solution if the surface is saddle shaped.

5.2.6 The Fundamental Instability of a True Image Irradiance Equation

A true solution surface should consist of a single two-dimensional invariant manifold which is the unstable manifold for some critical points, the stable manifold for others, and a saddle invariant manifold for the rest. It turns out that this is very unusual behavior for a dynamical system. Generically, two two-dimensional invariant manifolds will only intersect (if they intersect) along a one-dimensional curve rather than merging (Abraham and Marsden, 1985). An intuitive explanation for this is that two generically transverse two-dimensional surfaces in a four dimensional space have a 0-dimensional intersection, i.e. a point; if both the surfaces are invariant under a flow, however, the flow of this point forward and backward in time will

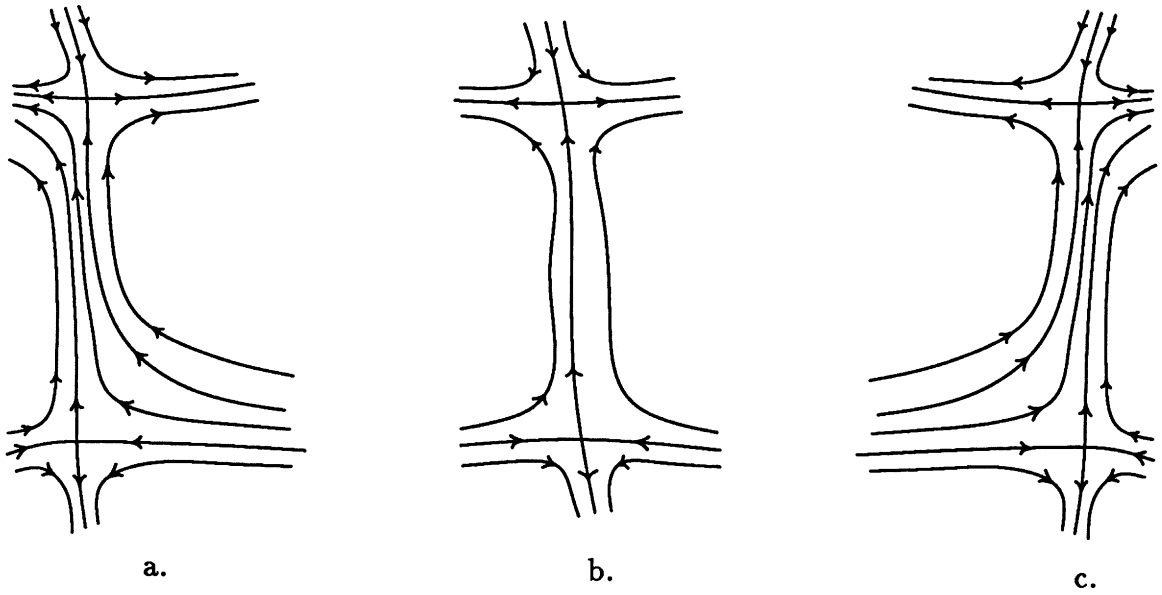


Figure 5.7. Two critical points of a two dimensional dynamical system: a. Non-intersecting stable and unstable manifolds. b. Intersecting stable and unstable manifolds. c. Non-intersecting stable and unstable manifolds.

also be in the intersection of the two invariant surfaces, leading to a one-dimensional intersection.

We can gain some insight into the situation by looking at two dimensions. In Figure 5.7, we show a piece of a two-dimensional dynamical system with two critical points, each of which has both a stable and an unstable manifold. In Figure 5.7 a and 5.7 c, these manifolds do not intersect; the unstable manifold for one critical point is just another trajectory for the other critical point. The case where they intersect is shown in Figure 5.7 b. If Figure 5.7 b is perturbed generically, it will fall apart to be either like Figure 5.7 a or 5.7 c.

The image dynamical system due to a real surface is therefore a very delicate thing seen from a generic perspective. This has two consequences, one bad, one potentially good. The bad news is that as we numerically try to find the stable or unstable manifold of a critical point (by directly drawing out trajectories or by the methods in the next section), errors will occur as if we had randomly perturbed the dynamical system. It is very unlikely that the perturbed unstable manifold will

merge with the perturbed invariant manifold coming from another critical point and create a single, smooth, two-dimensional surface. The good news is that that this may provide a way to tell a bad reflectance function choice from a good one: if the invariant manifolds do not "nearly" match up, we must be on the wrong track. Either the reflectance function is bad, or there is no surface corresponding to the image.

Note one other theoretical hope for optimism: although generically the two-dimensional invariant manifolds intersect in a manifold of dimension at most one (and may not intersect at all), we may get better results if we look at a one- (or many-) parameter family of vector fields, e.g. by examining a one-parameter family of reflectance functions. A result with a catastrophe theoretic flavor about generic vectors fields (Abraham and Marsden, 1985) indicates that if we have a one-parameter family of vector fields X^c on a four dimensional space controlled by the real parameter c such that at $c = 0$ a pair of two-dimensional invariant manifolds actually completely merge, then perturbing this entire family does not destroy this occurrence: in the perturbed family, there will still be a value of the new control parameter such that the matching occurs. For example, if we know that the reflectance function comes from a one-parameter family of reflectance functions and the image of the real surface contains two critical points, then the perturbations of the system due to numerical errors or slight errors in the reflectance function will not remove the theoretical occurrence of a parameter value such that the invariant manifolds merge.

In our case, we may need to match more than two invariant manifolds, and we may have more parameters in the reflectance function, so the task is harder than the above. The generic theory for more than two parameters of control is not well worked out, but it may be that adding more parameters of control allows one generically to match more invariant surfaces together and thereby determine more parameters of the reflectance function.

5.3 Invariant Manifold Shape from Shading Algorithms

From Section 5.2, we can conclude that good critical points of the image dynamical system provide constraints on reasonable solutions. Given a smooth image interior, we expect corresponding solution surfaces also to be smooth. In a neighborhood of a critical point in the image caused by a critical point of the reflectance function, a correct solution surface must be a smooth surface drawn out by characteristic curves (and can be considered locally invariant to flow along these curves), and must contain the critial point. Thus, a consistent solution surface passing through all the critical points of a smooth image interior should be an invariant manifold containing all these critical points. In some sense, this is quite special behavior for an invariant manifold, as we mention in Section 5.2.6.

In the space invariant case, we can consider the image dynamical system near the critical point to live in the four dimensional space $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$ which can be given the coordinates (x, y, p, q), where x and y are image coordinates, and p and q are the "gradient" coordinates for orientation of the surface. The typical (i.e. generic) image critical point due to a maximum in the reflectance function will be an isolated, hyperbolic critical point of the dynamical system. The Grobman-Hartman theorem lets us conclude that the nonlinear image dynamical system near the critical point will be topologocally isomorphic to the linearized dynamical system near the critical point; as a result, there are in general at least two possible invariant manifolds through the usual maximum critical point, the stable and unstable manifolds of the critical point, and in the generic case (e.g. no symmetries in the dynamical system) at most four.

These theoretical results indicate that one can find possible solution surfaces for an image in the neighborhood of a critical point by finding invariant manifolds of the critical point. One technique is to pick a curve in $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$ (an initial strip) that lies approximately close to the invariant manifold of interest, and then use the flow to

stretch this curve out to cover a larger region. This is the essence of Horn's original attempt (Horn, 1975) to find solution surfaces around a critical point. As he pointed out, there are integration problems as one gets far from the original curve due to the build-up of quantization and integration errors. As a result Horn and Brooks (Horn and Brooks, 1986) examined various regularization techniques designed to use more of the image data at once to generate a solution surface, but did not use the characteristic curves.

There is another method for finding some of the invariant solution surfaces corresponding to solutions of the shape from shading problem. In this section we discuss the intuition behind a theorem called the Lambda Lemma, or the Inclination Theorem. If we take an initial manifold of the same dimension as the unstable manifold and transverse to the stable manifold of a critical point and deform the initial manifold using the flow of the dynamical system, then in the limit as time goes to infinity the Lambda Lemma asserts that the deformed surface will approach the unstable manifold in a C^1 manner (Palis and de Melo, 1982). We discuss two different implementations of this intuition to find the unstable manifold of an image dynamical system near a critical point, and show how these methods are affected by noise in the image and errors in the reflectance function: they turn out to be fairly robust.

5.3.1 The Intuition

The unstable manifold of the critical point of a dynamical system consists of trajectories that seem to emerge from the critical point at $t = -\infty$ and continue away from the critical point. Figure 5.8 gives an example of a hyperbolic critical point (i.e. a critical point whose linearization has only real eigenvalues) in two dimensions; the unstable manifold consists of the two trajectories that approach the critical point with decreasing time.

Other trajectories starting near the unstable manifold and near the critical point approach the unstable manifold for some time within some neighborhood of the critical point. If we take an initial manifold S^0 which crosses the stable manifold

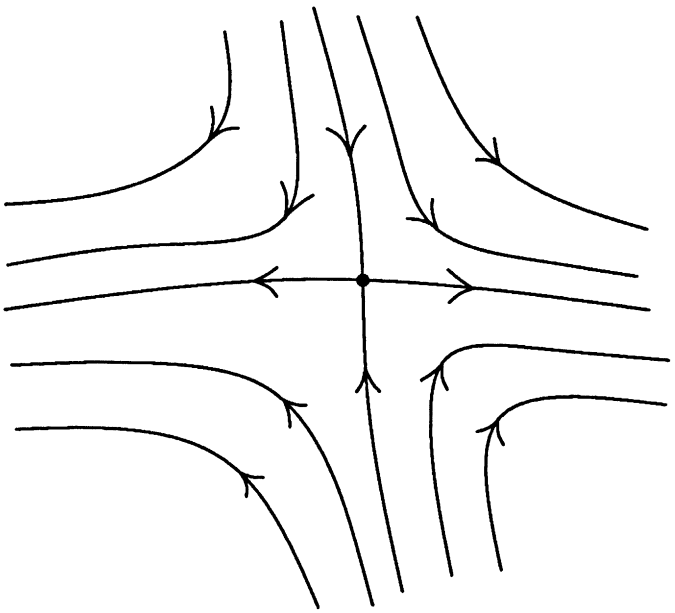


Figure 5.8. Two dimensional hyperbolic critical point.

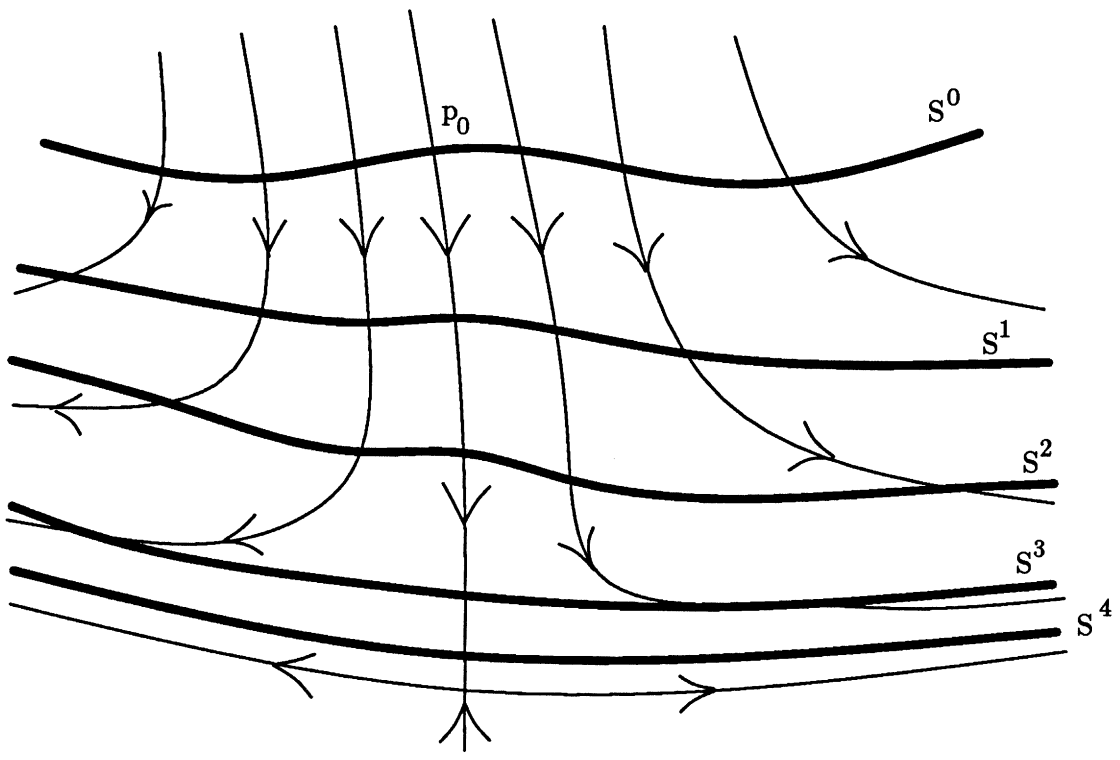


Figure 5.9. Initial manifold is S^0; S^i are the deformations of S^0 by the flow of the dynamical system.

(Figure 5.9), we can watch what happens as the initial surface is deformed by moving

each point of the surface along the flow of the dynamical system, as a stream of water might move particles suspended in it. Since the surface intersects the stable manifold, at least one point on the deformed surface begins to approach the critical point. In addition, the behavior of trajectories near the unstable manifold is an exponential approach to the unstable manifold within a certain neighborhood of the critical point; as the initial surface is stretched and deformed by the flow, more and more of it will be stretched and pushed down towards the unstable manifold by the flow lines. In the limit, this deformed surface will become the unstable manifold. Although the example shown is in two dimensions, the intuition holds more generally. A proof of the Lambda Lemma comes by looking at the deformation of an initial surface carefully and taking limits (Palis and de Melo, 1982).

The basic intuition behind the example and the theorem is to begin with an initial surface that cuts the stable manifold transversely, and then deform it through time using the flow of the dynamical system. We have implemented this as an algorithm on a highly parallel computer called the Connection Machine in two different ways: in the first case, we look mathematically at how the surface is deformed by the flow without following points on the surface; in the second case, we actually follow a grid of points on the initial surface as the flow moves them around.

5.3.2 Fixed Grid Algorithm

The Connection Machine is a highly parallel computer consisting of thousands of simple processors and a very fast disk drive system which can emulate a grid of processors. We used a 16k CM-1, which has 4k of memory for each of 16k processors. The kind of algorithms we will propose are extremely well suited to this kind of architecture with a SIMD language called *LISP as interface, since each processor can be assigned to a pixel and operates in a neighborhood of that pixel in parallel with all the other processors.

We configured the processors as a 128×128 grid. For the fixed grid algorithm, each processor can be thought of as stuck to an unmoving (x, y) coordinate plane

parallel to the image plane. A two-dimensional surface in the four-dimensional space $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$ is defined above the plane by specifying (p, q) coordinates at each (x, y).

5.3.2.1 Theory

In determining how the initial surface is deformed by the image dynamical flow, we are interested in how the surface is deformed as a set, not necessarily in where each point is mapped by the flow. Knowing how the deformation process affects the heights of the surface above each mesh point is sufficient: it is a two-dimensional surface in a four-dimensional space, so we have both $p(x, y)$ and $q(x, y)$ as heights above the point (x, y). Assuming the initial surface and the flow are smooth, we can theoretically compute the small change in surface height above a fixed point due to a short flow along the vector field. We can use this to iterate the surface height above each mesh point and find how the surface deforms.

To do this, we consider a slightly more general situation: say we have a time-dependent deforming surface S^t defined by

$$S^t = \{p | G_t(p) = 0\}$$

for some $G(t, p)$ where $G : \mathbb{R} \times \tilde{\mathcal{C}}(\mathbb{R}^3, 2) \longrightarrow \mathbb{R}^2$ is smooth, with $G_t(p) \triangleq G(t, p)$ having the maximal rank of two as a function of p for all t: the surface S^t is the level set at 0 of the function G_t. Assume we also have a surface $P = \{p | F(p) = 0\}$, where F is also smooth and of maximal rank. In our case, $P = \{(a, b, p, q) | p, q \in \mathbb{R}\}$ is the plane of possible (p, q) values for a fixed $(x, y) = (a, b)$ in the usual coordinate chart. We are interested in the intersection of S^t with P as time proceeds; this will give us (p, q) as a function of t above a fixed $(x, y) = (a, b)$.

What happens to points in $S^t \cap P$ as a function of time? Let us consider a path $\alpha : \mathbb{R} \longrightarrow S^t \cap P$ that stays in the intersection of the two surfaces for some time interval around $t = 0$, with $\alpha(0) = p_0$. This means that for all valid t, we have

$$F(\alpha(t)) = 0$$

$$G_t(\alpha(t)) = 0.$$

We can look at the time derivatives to get some constraint on the time derivative of α, $\alpha'(0) = \mathbf{v}$, at $t = 0$:

$$DF(\mathbf{v}) = 0 \tag{†}$$

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}G\right)_{(0,p)} + D(G_0)(\mathbf{v}) = 0. \tag{‡}$$

In our case, we can write

$$G_t(p) = g \circ \phi_{-t}(p),$$

where ϕ_t is the flow due to the characteristic vector field $\tilde{X}$, and $S^0 = \{p|g(p) = 0\}$ is the initial surface we started with.[8] We have, using the chain rule and the fact that ϕ_t is the flow for $\tilde{X}$,

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}G\right)_{(0,p)} = -Dg \circ \tilde{X}(p)$$

and

$$D(G_0)(\mathbf{v}) = Dg(\mathbf{v}),$$

since $G_0 = g \circ \phi_0(p) = g(p)$. Substituting in to the second constraint (‡) on $\mathbf{v}$ we have

$$Dg(\tilde{X}) = Dg(\mathbf{v}).$$

In our case, F, G_t, and g all have rank 2. If we assume that F and G_t are maximally independent (i.e. the respective surfaces are transversal) at intersection points, then we should only have isolated intersection points at each t: we are intersecting two two-dimensional surfaces in a four-dimensional space. This means the curve $\alpha(t)$ in the intersection of the two surfaces is, in fact, a unique path, and it has tangent vector at $t = 0$ given by $\mathbf{v} = (v_x, v_y, v_p, v_q)$, where $\mathbf{v} \in \mathrm{Null}(\mathrm{d}F)$ and $Dg(\tilde{X}) = Dg(\mathbf{v})$.

[8] So p is in S^t if and only if $G_t(p) = 0$, i.e., if and only if $g \circ \phi_{-t}(p) = 0$, i.e., the point at $t = 0$ that flows to p at time t is in $S^0 = \{p|g(p) = 0\}$.

With our particular coordinate choice, this is an easy system to solve: we have $F(x,y,p,q) = (x,y)$, so the first condition (†) on $\mathbf{v}$, $DF(\mathbf{v}) = 0$, means that $v_x = v_y = 0$. If we write

$$g(x,y,p,q) = (b(x,y) - p, c(x,y) - q)$$

which lets us think of the initial surface S^0 in $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$ as parameterized by

$$(r,s) \longmapsto (r,s,b(r,s),c(r,s)),$$

then we have

$$Dg = \begin{bmatrix} b_x & b_y & -1 & 0 \\ c_x & c_y & 0 & -1 \end{bmatrix},$$

so the second condition on $\mathbf{v}$, $Dg(\tilde{X}) = Dg(\mathbf{v})$, can be written as

$$b_x X_x + b_y X_y - X_p = -v_p$$

$$c_x X_x + c_y X_y - X_q = -v_q.$$

In the usual (x,y,p,q) coordinate system used for the critical point analysis, we can replace in for the values of the characteristic vector field components to get:

$$v_p = E_x - b_x R_p - b_y R_q$$

$$v_q = E_y - c_x R_p - c_y R_q.$$

v_p and v_q are the infinitessimal displacements at $t = 0$ of the surface above a fixed (x,y) due to the dynamical flow deformation. We recalculate the partial derivatives b_x, b_y, c_x, c_y where $b(x,y) = p(x,y)$, $c(x,y) = q(x,y)$ at each step of the iteration.

We can perhaps view v_p and v_q as defining a vector field on the space of smooth two-dimensional surfaces in $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$: at each t, we have a smooth surface which defines the derivatives of p and q (the derivatives of b and c respectively) with respect to the coordinate chart we have picked. These derivatives are combined together to give a smooth vector field which is, in some sense, the infinitessimal displacement of the entire surface.

5.3.2.2 Experiments

We have implemented this as a discrete iterative method on the Connection Machine. We essentially tie the grid of processors to the (x, y) coordinate plane, with a processor at each integer crossing. At each processor we can keep values for the image gradients E_x and E_y as well as current estimates of the deformed surface heights p and q for each (x, y). We use an initial surface given by $p(x, y) = q(x, y) = 0$.

We need to compute the "surface vector field" components v_p and v_q at each iteration:

$$v_p = E_x - p_x R_p - p_y R_q$$

$$v_q = E_y - q_x R_p - q_y R_q,$$

where p_x, p_y, q_x, and q_y are derivatives of the current p and q iterates.

For both image intensities and p and q values, we use a simple first order method to compute the derivatives on the interior of the 128×128 square of processors with integer (x, y) coordinates for the processors: for example, if f is the function for which we want a partial derivative, we take $f_x(x, y) \approx (1/2)(f(x+1, y) - f(x-1, y))$. At the four boundaries of the square grid, we cannot use this; we use the unbalanced estimates given by subtracting nearest neighbors where we have to. For example, at the left edge we take $f_x(0, y) \approx f(1, y) - f(0, y)$; at the top we take $f_y(x, 0) \approx f(x, 1) - f(x, 0)$; at the corners of the grid, both f_x and f_y are approximated this way.[9] We use this to estimate p_x, p_y, q_x, q_y, E_x, and E_y.

For the reflectance derivatives, we use an analytic model to give us exact values for the derivatives. In our case, we assume a Lambertian reflectance function,

$$R(p, q) = \alpha \frac{pl_1 + ql_2 - l_3}{\sqrt{p^2 + q^2 + 1}\sqrt{l_1^2 + l_2^2 + l_3^2}}$$

[9] When we discuss the deforming grid algorithm in the next section, we must treat the border differently: we effectively interpolate the values found for the derivative from the interior of the image to the borders. Using the interpolating method in the fixed-grid algorithm does not change the performance substantially; noise immunity seems slightly worse.

where (l_1, l_2, l_3) gives the orientation of the light source and α is an albedo factor; for our purposes, we take $\alpha = 1.0$. From this, we have

$$R_p(p,q) = \frac{\alpha}{\sqrt{l_1^2 + l_2^2 + l_3^2}} \frac{l_1 q^2 + l_1 - (q l_2 - l_3)p}{(p^2 + q^2 + 1)^{3/2}}$$

$$R_q(p,q) = \frac{\alpha}{\sqrt{l_1^2 + l_2^2 + l_3^2}} \frac{l_2 p^2 + l_2 - (p l_1 - l_3)q}{(p^2 + q^2 + 1)^{3/2}},$$

and we compute R_p and R_q at each processor using the current values of p and q.

To approximately deform the surface to match the dynamical system flow, we update the p and q values at each processor to p_{new} and q_{new} as follows:

$$p_{new} = p + hv_p$$

$$q_{new} = q + hv_q,$$

where h is a kind of integration step size if we consider the procedure as a parallel integration of a vector field. In the simple examples with which we have worked, we have left h constant over all processors.

If we run just this algorithm we find that it does not converge. We have found it necessary to do a small amount of smoothing of the values of p and q to avoid "checkerboard" instabilities: small amounts of noise can explode into large alternating sections of values if the p's and q's are not smoothed between iterations. This was also found by Ikeuchi and Horn (Ikeuchi and Horn, 1981), and we find good results by performing the simple averaging operation

$$f_{new}(x,y) = (f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1))/4$$

twice in succession on the p and q arrays.

This internal smoothing operation can be a source of convergence errors, however. If we run the algorithm on a sphere centered in the image with radius 100 and light source located in direction $(0, .5, -1)$, we can see what happens to the value of q at a particular point (64, 104) near the maximum brightness of the image. (Figure

5.10 shows constant brightness contours for such an image.) The correct q is $-.4365$ and we take $h = 1.0$ as step size. Over the first 500 iterations, we see the value of q decreasing from its initial value of 0 towards the correct value; as we expect, the value of v_q is negative, but with decreasing absolute value during this approach. At around iteration 600, however, the value of q overshoots the correct value: we have $q = -.4390$, and, as we expect, v_q turns slightly positive, trying to push q towards the correct value. However, the value of q continues to move negatively, converging to a value of about $-.4505$ at iteration 1600 with $v_q = 1.21 \times 10^{-4}$ still pointing towards the correct solution.

If at this stage we change the algorithm to have just one internal averaging operation instead of two, we see q begin to move towards the correct solution again, but by iteration 800 with this new method, q has stalled at $q = -.444$ with $v_q = .626 \times 10^{-4}$; we have made up some of the distance to the true solution, but not all.

Another way to see the effects of the internal smoothing operation is to begin with the correct solution surface and see where the algorithm takes it: by definition, the dynamical flow should leave an invariant manifold completely unmoved. If we do this with two internal smoothing operations, we see convergence to very nearly the same surface as for the $(p, q) = (0, 0)$ initial surface: for example, after 800 iterations, q at $(64, 104)$ appears to be converging to $-.450$ rather than staying at the original $q = -.4365$. If we do this with one internal smoothing operation, the solution starting from the invariant manifold is again consistent with that for the $(p, q) = (0, 0)$ initial surface: again the value of q converged to at $(64, 104)$ is $q = -.444$ rather than the original $q = -.4365$.

From this we can conclude that the internal smoothing operation, although numerically useful, does slightly distort the final solution away from the true invariant manifold. With one internal smoothing operation, the errors in the converged p and q are less than 5% almost everywhere, and less than 2% in the image interior. Even with two internal smoothing operations, the errors are less than 4% through almost

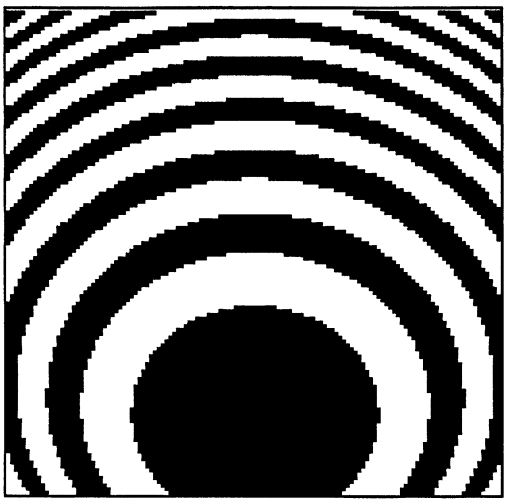


Figure 5.10. Constant brightness contours for sphere. Each stripe is 15 grey-levels wide.

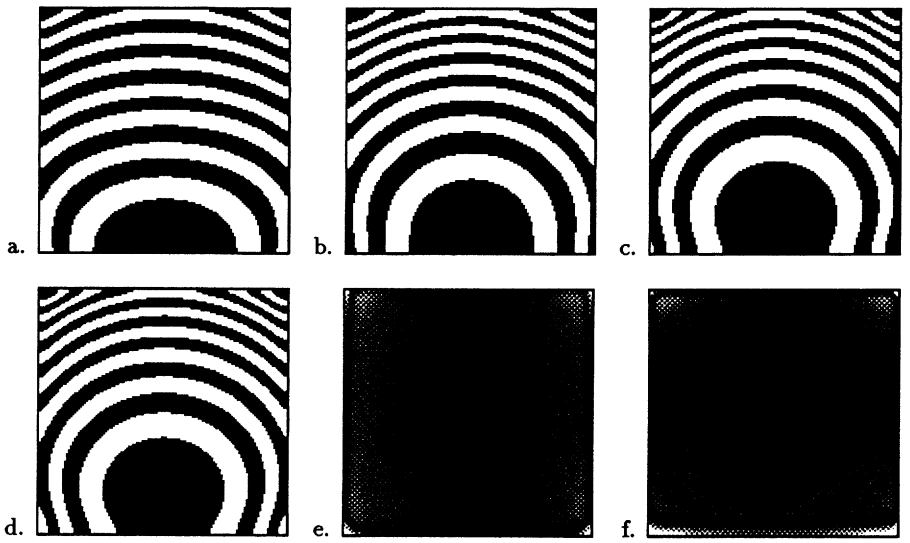


Figure 5.11. Iterates for fixed grid shape from shading algorithm: a. 100 iterations. b. 200 iterations. c. 400 iterations; d. 800 iterations. e. p error image ($255 * 10 * |p - p_{\text{true}}|$). For 800 iterations: f. q error image ($255 * 10 * |q - q_{\text{true}}|$).

all the image—where true p and q values are very near 0, of course, the relative errors are higher, and right at the borders of the image the relative errors are also higher (but never more than about 15%). We will stick with the double smoothing because of the noise immunity it seems to give later on.

In Figure 5.10, we show the constant brightness contours for a noise-free 128×128 simulated image of a sphere with radius 100, Lambertian reflectance function, and

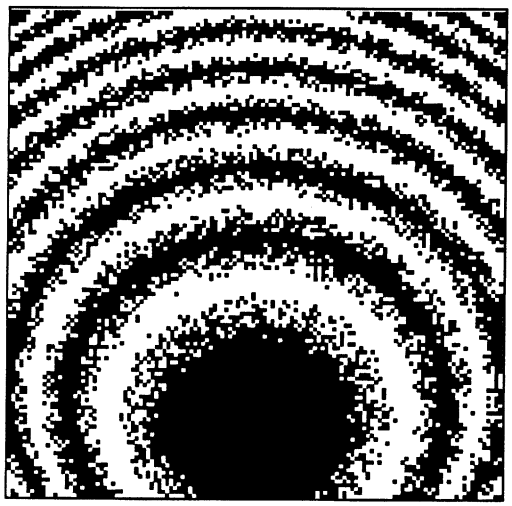


Figure 5.12. Noisy sphere image: Noise maximum is $\pm.02$.

light source located at $(0, .5, -1)$. In Figure 5.11 we show images from the approximate solutions as the iterations proceed (step size $h = 1.0$): after 100 iterations, we use the current p and q values and the reflectance function to generate an image. As one can see, the images generated by the p and q estimates do show convergence to the original image through almost all of the interior of the image as expected. We have also generated images representing the errors in the p and q values: we take $255 * 10 * |p - p_{\text{true}}|$ and $255 * 10 * |q - q_{\text{true}}|$ and treat these as grey levels for an error image. A region that is just barely completely white (a grey level just reaching 255) would represent errors of .10 in p or q, corresponding to an error of about 6° in the surface normal at $p_{\text{true}} = 0$, $q_{\text{true}} = 0$, and less as p and q increase.

If we add noise to the original image, we can degrade the performance of the algorithm. With the reflectance function we are using, the maximum brightness of the image is 1.0. If we add uncorrelated uniformly distributed noise to the image with maximum value .02 (meaning the noise is uniformly distributed between $\pm.02$), we usually still get convergence; if we add noise with maximum value .12, we do not get convergence to a possible solution. Figure 5.12 shows a noisy image (via constant brightness contours) with .02 noise maximum; Figure 5.13a shows the image formed from the p and q arrays after 800 iterations with .02 noise, and Figure 5.13b and c

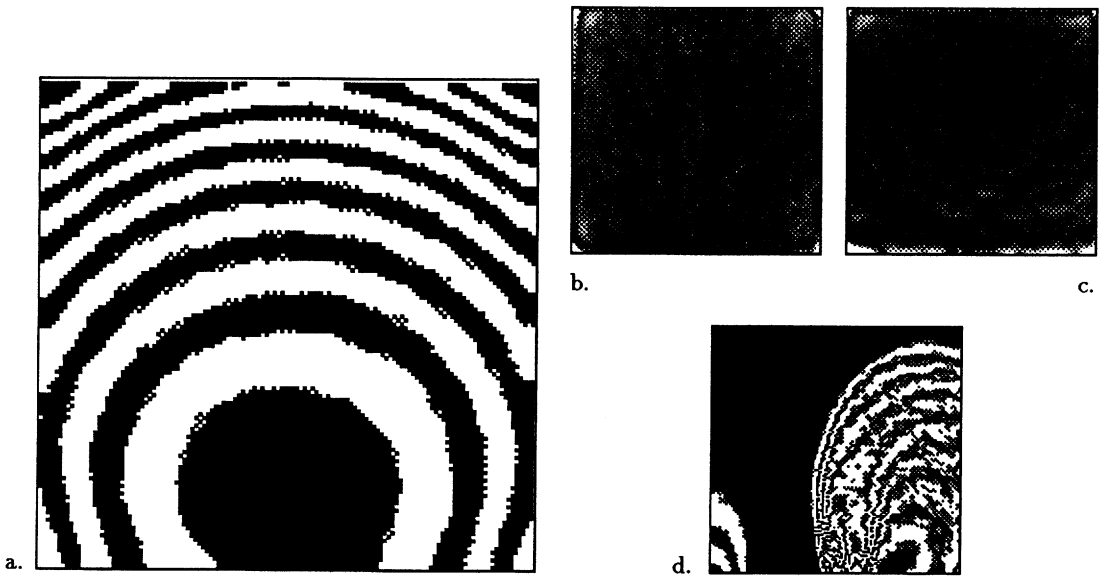


Figure 5.13. a. Estimated image from noisy images (±.02) after 800 iterations. b. p error image. c. q error image. d. Unconverged image for noise maximum .12 after 800 iterations.

show the p and q error images. Figure 5.13d shows a failed effort with noise maximum .12.

We can deal with more noise by reducing the step size, h. For example, with noise maximum .04, a step size of $h = 1.0$ converges only about half the time to a solution,[10] while a step size of $h = .25$ almost always converges. This is consistent with the view of this procedure as a parallel Euler integration of sorts: decreasing the step size makes the algorithm more likely to converge.

Another way to deal with the uncorrelated noise we have added to the image is to pre-filter the image to remove some of the discontinuous influence of the noise. For example, with $h = 1.0$ we get convergent solutions about half the time with noise maximum .04; however, if we smooth the image four times with the simple averaging operation used to smooth the p and q arrays, we get convergence almost always.

If we combine filtering and a smaller step size we can deal with even more noise. If we set the step size to $h = .25$, and pre-filter with one averaging operation, the

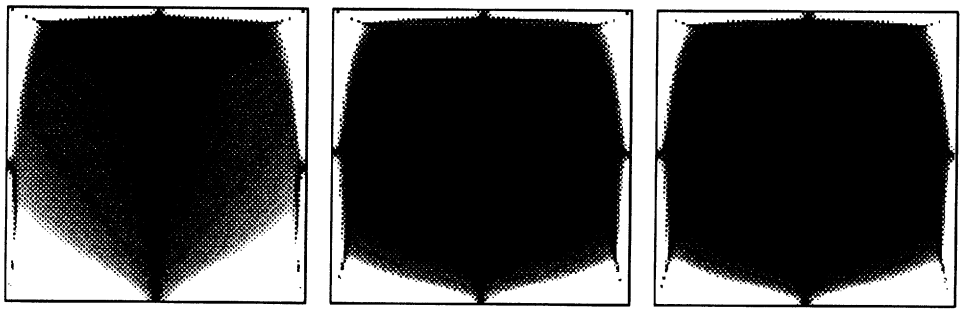
[10] That is, if we run the algorithm twenty times it converges on about ten of the trials.

brightness image on the left looks very chaotic because the noise is large compared to the constant brightness intervals even after pre-filtering. To the eye, the original image itself looks fuzzy, but is clearly interpretable as a smoothly curving object.

Note that in general the algorithm does not converge to p and q values that could have given rise to the pre-filtered noisy image, e.g. some complicated faceted figure whose noise-free image would be the pre-filtered noisy image we are working with. Instead, due to the internal smoothing operations done on the p and q arrays within the algorithm, the algorithm appears to converge to a surface that is smooth compared to the image data.

With this much noise, the converged p and q arrays do show considerable differences from the original noise-free p and q values used to generate the images. For example, although the image in Figure 5.15a appears to be an adequate match for the original image, the p's and q's that generate that image vary randomly from the original image by a fair bit, as seen in the image of p and q errors in Figure 5.15 b and c. Indeed, one would be rightly suspicious that the p and q values converged to for noisy images are not an integrable set of normal vectors. We would need some kind of constrained method like Horn and Ikeuchi's or Frankot and Chellapa's to reconstruct the depth values from the more or less noisy p and q values. Note that the amount of noise in the p and q values is directly related to the amount of noise in the image.

In fact, one of the interesting aspects of this algorithm is the lack of dependence on an explicit integrability constraint to find a solution surface. If we actually converge to the theoretical unstable manifold of a smooth image dynamical system, that unstable manifold is an integrable collection of normal vectors because it is made up of characteristic trajectories. When we add noise to the image and include numerical effects, we are effectively adding noise to the dynamical system, and the unstable manifold of the new dynamical system is no longer exactly integrable; however, if the noise is small, the normals will be very nearly integrable.



a. b. c.

Figure 5.16. Integrability pictures for iterations of the noise-free image of a sphere. a. 200 iterations. b. 400 iterations. c. 800 iterations.

We can use an integrability measure to monitor the progress of the algorithm. In order for the collection of normal vectors to be normal vectors for a real surface in space, we must have $p_y = q_x$; we can use $|p_y - q_x|$ as a measure of the integrability of the iterated surface and therefore as a measure of how close we might be to the correct unstable manifold of the image dynamical system.

In Figure 5.16 we have tried to give a pictoral representation of this integrability measure (IM) in the noise-free case (Figure 5.11). The brightness values run from 0 to 255; these images are images of $10^6|p_y - q_x|$ with truncation at values greater than 255, so that the white area area represents IM values of larger than 2.5×10^{-4}. The nearly black region (represented here as having very few white dots) have $p_y - q_x$ values a small fraction ($< 1\%$) of most of the p_y values in the region. As iterations proceed, the regions of good integrability increase, although the edges of the image,where the converged constant brightness contours do not match the original (compare Figure 5.10 with Figure 5.11d), remain with relatively high IM values.

Adding noise to the image degrades the integrability measure of the iterated solutions. In Figure 5.17 we show some examples of integrability measure pictures under noisy conditions: we look at the IM pictures for a particular image with .02 maximum uniformly distributed noise added, with step size of $h = 1.0$ and four pre-filterings of the image: as the iterations proceed, the results become more and

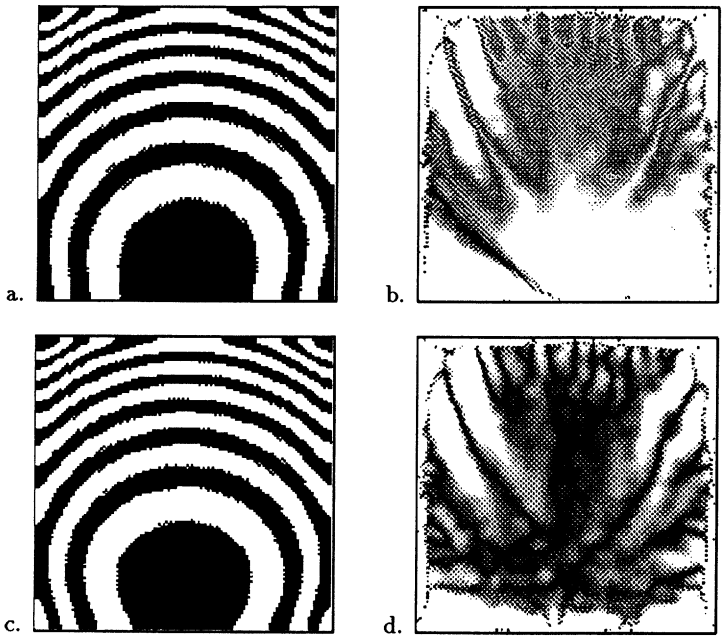
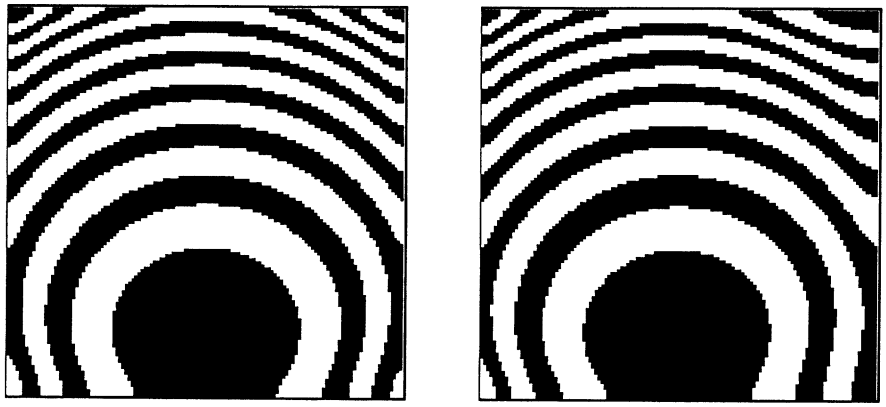


Figure 5.17. Integrability measures in the case of noise: noise maximum .02, h= 1.0, 4 smoothings of the image. a. 800 iterations. b. IM image after 800 iterations. c. 1600 iterations. d. IM image after 1600 iterations.

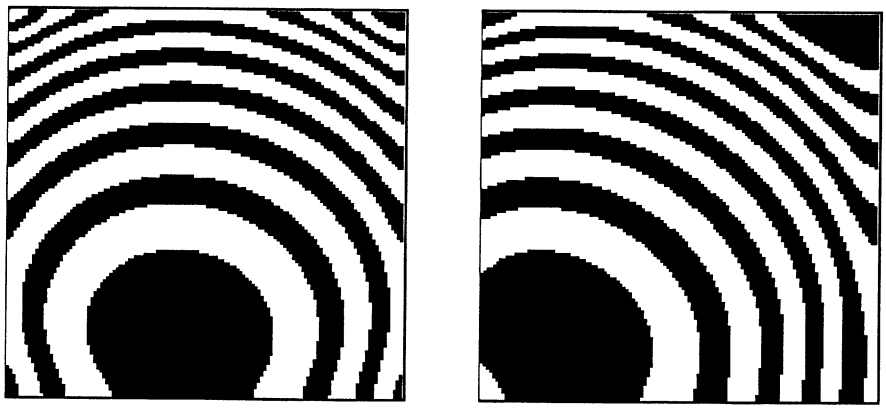
more integrable. This is consistent with the theoretical view that we are approaching the unstable manifold of the image dynamical system which should be perfectly integrable; in fact, however, progress is essentially halted after 1600 iterations: numerically, almost no more changes in the solutions or in the IM image occur. Figure 5.15d shows the integrability picture for noise maximum of .25; clearly, more noise leads to worse integrability.

We can also examine the sensitivity of the algorithm to changes in the reflectance function. In Figures 5.18 to 5.20, we explore what happens if we make errors in the direction of the light source assumed to have generated the image. We assume different values for l_1, where $(l_1, l_2, l_3) = (0, .5, -1)$ is the original light source direction. As the error gets worse, the convergence of the algorithm after 800 iterations is not as complete; this can also be seen in the integrability pictures in Figure 5.20. Nonetheless, the arrays of p and q reached do generate images that correspond to the original over large areas with the assumed reflectance function (Figure 5.18); however, the surfaces themselves are increasingly different from the correct surface,



a. b.

Figure 5.18. Incorrect reflectance function: images of p and q arrays with assumed reflectance function. Noise-free, 800 iterations, $h = 1.0$, correct $l_1 = 0.0$. a. $l_1 = .10$. b. $l_1 = .5$.

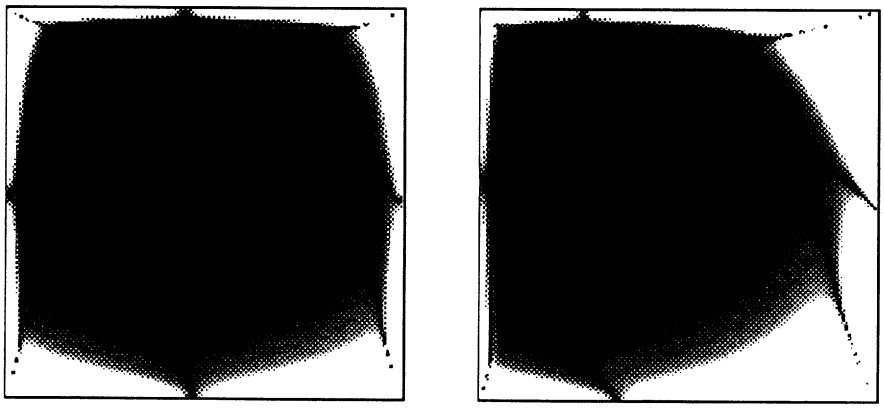


a. b.

Figure 5.19. Incorrect reflectance function: images of p and q arrays with correct reflectance function. Noise-free, 800 iterations, $h = 1.0$, correct $l_1 = 0.0$. a. $l_1 = .10$. b. $l_1 = .5$. Compare with original image, Figure 5.10.

as can be seen from the images generated by the converged surface and the correct reflectance function (Figure 5.19).

The fixed grid method appears to generate good solutions even in the presence of noise and seems to degrade gracefully if assumptions about the light source direction are incorrect. This is consistent with the theoretical underpinnings of the method: noise in the image or a poor choice of reflectance function represent perturbations of the original image dynamical system. Since the usual good critical point we deal



a. b.

Figure 5.20. Incorrect reflectance function: Integrability measure pictures. Noise-free, 800 iterations, $h = 1.0$, correct $l_1 = 0.0$. a. $l_1 = .10$. b. $l_1 = .5$.

with is hyperbolic and therefore generic, perturbing the dynamical system a small amount moves the critical point a small amount and changes the invariant manifolds a small amount: the unstable manifold of the critical point is a stable feature (in this sense) of the dynamical system.

This particular implementation shares difficulties with the simple Euler method of integrating a vector field, which sequentially adds a scalar multiple of the vector field to a point on the developing trajectory to generate the next point. For example, if the step size here is too big, the method may bounce around and not converge at all. Smaller step sizes avoid this problem, but too small a step makes for slow progress; some kind of adaptive step size setting would be useful. Note that because of the smoothing of the p and q arrays as part of each iteration step, taking an exceptionally small step size effectively allows more smoothing and less deforming along the dynamical flow. There may be analogs to the Runge–Kutta methods that make efficient use of several evaluations of the vector field to improve the estimate of the next point on the trajectory.

A feature that may be exploited is the fact that we are really interested in the characteristic paths, not the trajectories. This suggests we could use a different (positive) $h(x, y)$ to multiply the vector components v_p and v_q for each point (x, y);

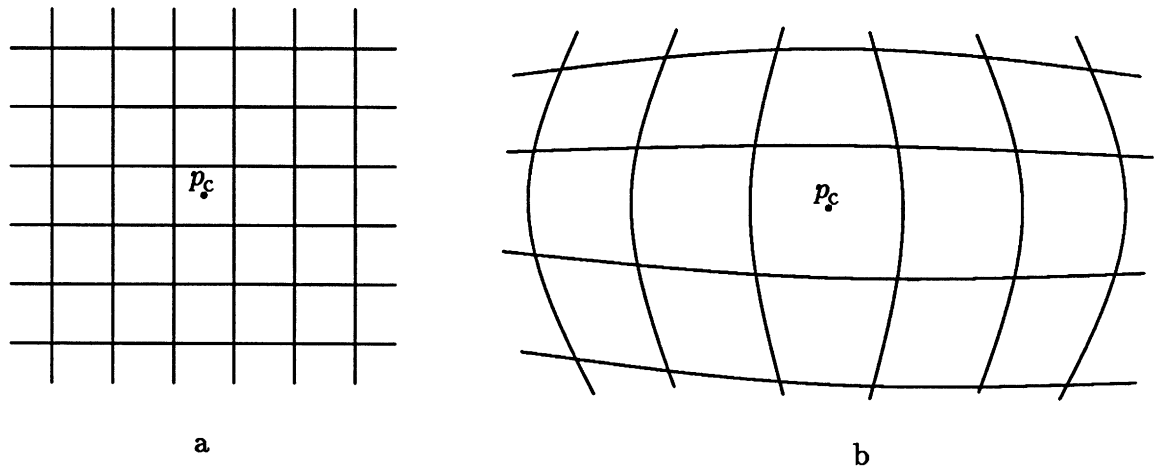


Figure 5.21. Stretching of the mesh around a critical point p_c. a. Before flowing. b. After flowing.

this would change how the initial surface is deformed, but would not change the surface that is converged to in the limit. The limit is determined by the characteristic paths and the directions they are traversed, not by how fast the paths are traversed through time; changing the magnitudes of v_p and v_q by a (positive) scalar at each point of the image should not change the theoretical limiting behavior, and may be useful in dealing with difficult regions.

The integrability measure provides a way to judge the progress of the convergence; changing the step size within regions that are having difficulty may improve convergence; and pre-filtering the image with appropriately sized filters can reduce the effects of noise. These all contribute to finding the unstable manifold of an image dynamical system critical point in a noisy environment. More work needs to be done to integrate these different features of the algorithm into a system for finding the unstable manifolds of image critical points.

5.3.3 Deforming Grid Algorithm

5.3.3.1 Theory

We have also implemented a more direct version of the theoretical idea. The basic idea is to place the mesh of Connection Machine processors on the estimated

surface in $\mathcal{C}(\mathbb{R}^3, 2)$, and move each mesh point with the flow of the image dynamical system. As can be seen from Figure 5.21, we will have an increasingly less dense concentration of processors near the critical point as time progresses. We need to continually fill in this region in some manner to maintain accuracy.

We have chosen to use a very short flow duration so that the distortion of the mesh is fairly minor. After this flow, we reset the mesh on the new surface: the original mesh is located at integer (x, y) coordinate pairs (the p and q values are not restricted). After the short flow, the mesh is located at various slightly altered non-integer (x, y) positions. We compute affine approximations to the surface near integer (x, y) positions using this distorted mesh, and then fill in the $p(x, y)$ and $q(x, y)$ values in the new integer mesh. This new mesh sits on the deformed surface and is uniformly distributed around the critical point; we can now flow this mesh and repeat.

To accomplish the flow, we use the image dynamical system vector field:

$$v_x = R_p$$

$$v_y = R_q$$

$$v_p = E_x$$

$$v_q = E_y,$$

and update the values of x,y, p, and q in each processor as follows:

$$x_{new} = x + hv_x$$

$$y_{new} = y + hv_y$$

$$p_{new} = p + hv_p$$

$$q_{new} = q + hv_q,$$

where again h is a kind of integration step size for the flow. We compute R_p, R_q analytically as in the fixed grid algorithm; E_x, and E_y we compute in the interior of

the image using a slightly different gradient operator than in the fixed grid case: as before, on the interior we take

$$f_x(x,y) = (1/2)(f(x+1,y) - f(x-1,y))$$

and similarly for f_y; however, the previous method does not yield satisfactory results at the borders. Instead, we interpolate the values of the gradients at the border from the values in the interior:

$$\begin{aligned} f_x(0,y) &= 2f_x(1,y) - f_x(2,y) \\ f_x(x,0) &= 2f_x(x,1) - f_x(x,2) \\ f_x(127,y) &= 2f_x(126,y) - f_x(125,y) \\ f_x(x,127) &= 2f_x(x,126) - f_x(x,125). \end{aligned}$$

To find the new mesh values and reset the grid on the deformed surface, we use a nine-member neighborhood of each deformed point to estimate a tangent plane through that point. To do this, we use a standard least squares approximation technique. We normalize the neighborhood to have 0 average x, y, p and q values, so that effectively we are working a least squares approximation near the origin. We solve for p's and q's separately. Letting f represent either p or q, we set

$$\begin{aligned} x_i^* &= x_i - \bar{x} \\ y_i^* &= y_i - \bar{y} \\ f_i^* &= x_i - \bar{f}, \end{aligned}$$

where x_i, y_i, and f_i are the coordinates of the nine different data points in the neighborhood, and $\bar{x}$, $\bar{y}$, $\bar{f}$ are the average values of those coordinates over the neighborhood. We can define

$$\begin{aligned} \mathbf{x} &= (x_1^*, \ldots, x_9^*)^T \\ \mathbf{y} &= (y_1^*, \ldots, y_9^*)^T \\ \mathbf{f} &= (f_1^*, \ldots, f_9^*)^T \end{aligned}$$

to be column vectors of the data coordinates. We seek a fixed column vector of coefficients $\beta = (\beta_1, \beta_2)^T$ such that we minimize

$$|| [\mathbf{x}, \mathbf{y}] \beta - \mathbf{f}||^2,$$

where $[\mathbf{x}, \mathbf{y}]$ is the 9×2 matrix of x_i and y_j values. This is the least squares problem of finding β_1 and β_2 to approximately fit the linear function $\hat{f}(x, y) = \beta_1 x + \beta_2 y$ to the (normalized) data we have. The usual vector derivative condition on β for finding the minimum of this expression is

$$0 = 2 [\mathbf{x}, \mathbf{y}]^T [\mathbf{x}, \mathbf{y}] \beta - 2 [\mathbf{x}, \mathbf{y}]^T \mathbf{f};$$

this can be rearranged to give

$$\beta = \begin{bmatrix} \mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbf{y} \\ \mathbf{y}^T\mathbf{x} & \mathbf{y}^T\mathbf{y} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x}^T\mathbf{f} \\ \mathbf{y}^T\mathbf{f} \end{bmatrix}$$

as the least squares estimate of the coefficients. Writing this in the usual summation notation, we have

$$\beta = \begin{bmatrix} \sum_i x_i^{*2} & \sum_i x_i^* y_i^* \\ \sum_i y_i^* x_i^* & \sum y_i^{*2} \end{bmatrix}^{-1} \begin{bmatrix} \sum_i x_i^* f_i^* \\ \sum_i y_i^* f_i^* \end{bmatrix}.$$

For each nine-point neighborhood of deformed points, we find $\beta = (\beta_1, \beta_2)^T$ this way. We then find the closest (x, y) integer coordinates to $(\bar{x}, \bar{y})$; say these are (m, n). We set

$$f(m, n) = \beta_1(m - \bar{x}) + \beta_2(n - \bar{y}) + \bar{f}.$$

If there is more than one neighborhood with the same (m, n) coordinate, we take the average of these values.

If we do not flow by very much, we will have found values for most of the integer coordinates in the new mesh in this way. There may still be some isolated curves of integer coordinates without new values: these are filled in using another least squares fit over the neighboring filled-in values. As we shall see, this gives us some difficulties in the integrability picture.

Again the edges of the grid pose special problems: in this case, we constrain the neighborhoods around a point to be just those points of the theoretical nine-point neighborhood that still actually lie on the grid. This gives us at worst four points (in the corners) from which to determine the tangent plane to the surface.

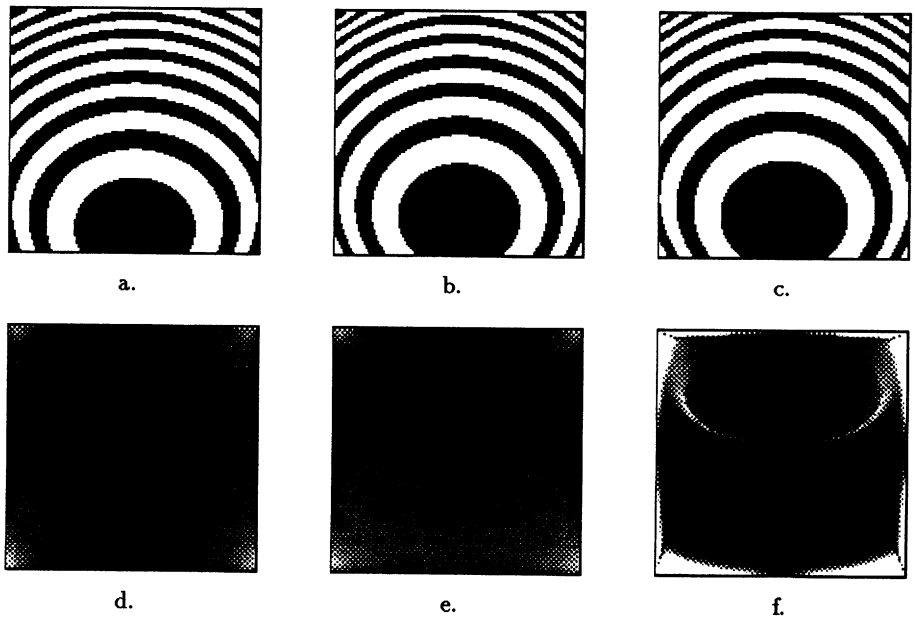


Figure 5.22. Deforming grid algorithm, no noise: a. 200 iterations. b. 400 iterations. c. 800 iterations. d. p error image after 800 iterations. e. q error image after 800 iterations. f. IM picture for 800 iterations.

5.3.3.2 Experiments

Because of the intensive least squares computations, this method is considerably slower than the fixed-grid method: it requires about 13 minutes on a 16K CM-1 to perform 100 iterations; this is in contrast with around 30 seconds for 100 iterations of the fixed-grid algorithm. In the noise-free case, the results are improved. In Figure 5.22 we see the results of the deforming grid algorithm over 400 iterations: the edges are much cleaner than with the fixed grid algorithm. We also see the IM picture at 400 iterations: most of the interior has very low $|p_y - q_x|$ values (relative values around 1% or less) except for the two moustaches of white in the upper half of the image; these correspond to slight "cracks" in the solution (smoothed somewhat by the least squares filtering) due to the method of computing the integer coordinates: on either side of these cracks, the neighborhoods used to compute the tangent to the surface are very different, e.g. on one side of the crack the integer coordinates might be towards the upper left of the patch while on the other side of the crack the integer coordinates might be towards the lower right.

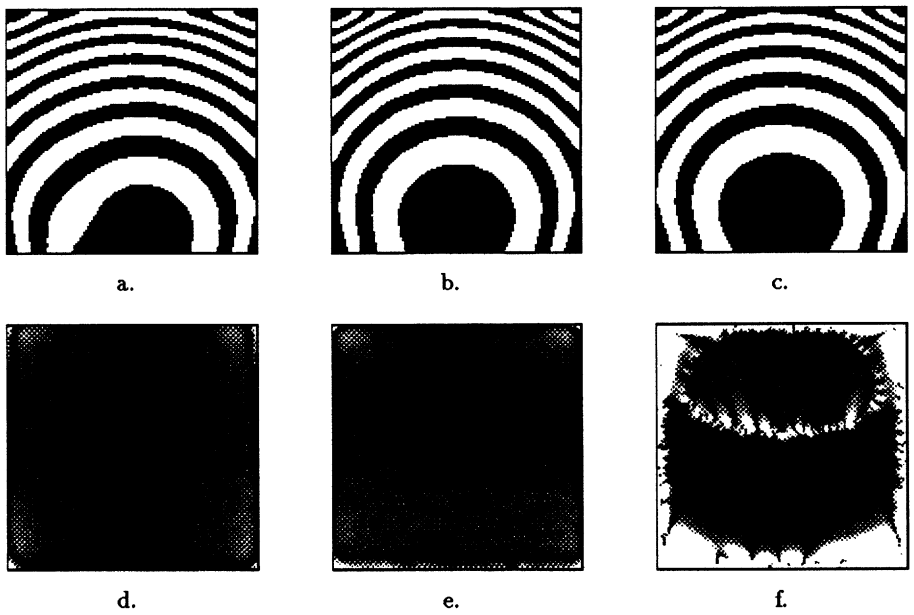
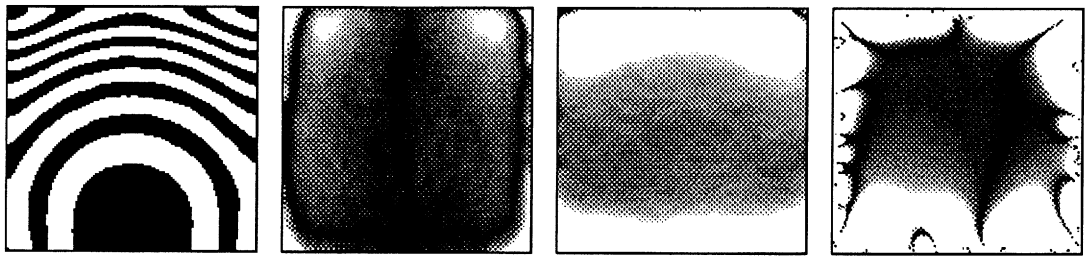


Figure 5.23. Deforming grid algorithm, noise with .01 max: a. 200 iterations. b. 400 iterations. c. 800 iterations. d. p error image after 800 iterations. e. q error image after 800 iterations. f. IM picture for 800 iterations.

Note that this result is achieved without separate internal smoothing steps: the least squares method itself provides some smoothing. If we add internal smoothing in the form of two averaging operations on the p and q values as we did in the fixed grid case, we get the slightly distorted border edges that we saw in the fixed-grid case (e.g. Figure 5.11), and the moustaches in the integrability picture fade but do not disappear.

If we add noise to the image without smoothing the image, we do not get good convergence. However, if we add internal smoothing of the p and q arrays at each iterations, we do get convergence: Figure 5.23 shows such a result, and the rest of the trials were done with the internal smoothing in place. With noise maximum values of .01, the deforming grid algorithm converges very well about 80% of the time;[11] results look quite acceptable after 600-800 iterations. The errors in the p and q values on the interior of the image seem to be running less than 5% for the p values

[11] That is, after ten attempted runs, about eight will have converged to a solution.

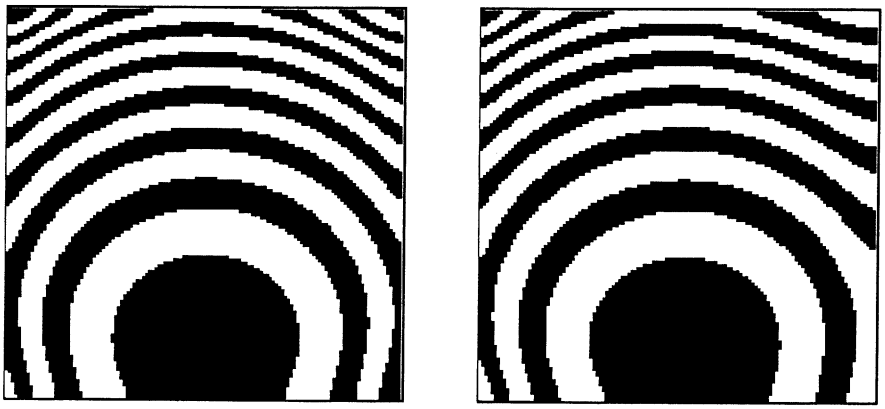


a. b. c. d.

Figure 5.24. Deforming grid algorithm, noise with .04 max, h=.25: a. Generated image after 1200 iterations. b. p error image. c. q error image. d. IM picture.

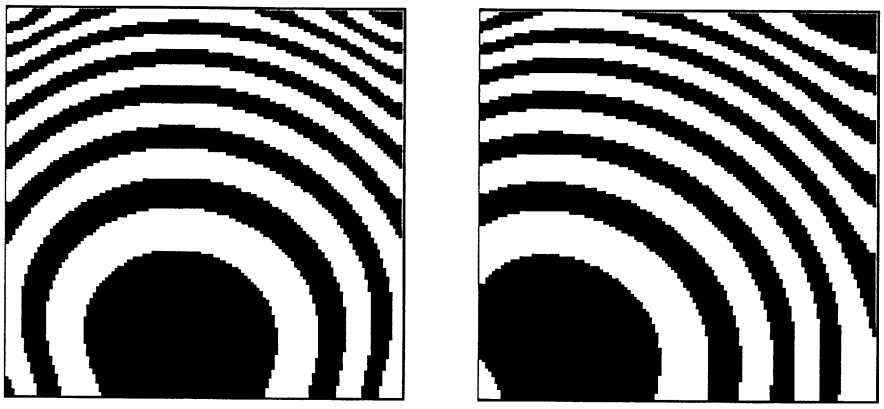
and less than 15% for the q values on the interior of the image, and as we move away from the center of the image where small p_y and q_x values magnify errors (and the moustaches make their presence felt), $|p_y - q_x|$ stays mostly below 10% of the value of p_y. To properly find a depth surface consistent with the p and q values we will have to use a method that can deal with the lack of integrability, but the values are not badly non-integrable. Note that the IM picture shares some resemblances with that for the fixed grid algorithm: the edges of the image again show poor IM values, and low IM values are associated with good matches between the original image data and the image generated from the estimated p and q values. If we increase the noise maximum to .02, we find convergence only about 20% of the time.

Smoothing of the image is useful in the deforming grid algorithm as well. With noise maximum of .02, smoothing the image changes the convergence rate from about 20% to just over half. Much more useful is a reduction in step size: all the above results were achieved using $h = 1.0$; if we drop to $h = .25$, we get dramatic improvements in noise immunity: with noise maximum of .04, for example, we get good convergence of the image all the time, as well as quite good integrability results 80% of the time without smoothing the image, and all the time with image smoothing. Even for noise maximum .08 the image appears to converge correctly 80% of the time without image smoothing; now, however, the IM pictures generally look completely white, meaning the p and q arrays converged to after 1200 iterations are much less integrable than in the lower noise conditions. In Figure 5.24 we show an example



a. b.

Figure 5.25. Distorting grid algorithm: incorrect reflectance function: images of p and q arrays with assumed reflectance function. Noise-free, 400 iterations, $h = 1.0$, correct $l_1 = 0.0$. a. $l_1 = .10$. b. $l_1 = .5$.

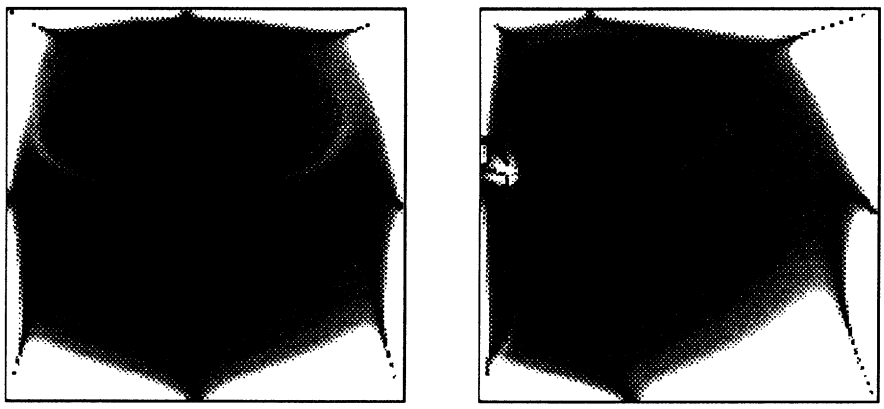


a. b.

Figure 5.26. Deforming grid algorithm: incorrect reflectance function: images of p q arrays with correct reflectance function. Noise-free, 400 iterations, $h = 1.0$, correct $l_1 = 0.0$. a. $l_1 = .10$. b. $l_1 = .5$.

of the convergence of the algorithm for step size $h = .25$ after 1200 iterations on an image with noise maximum of .04.

If we begin with the wrong reflectance function, we get very similar behavior to the fixed grid algorithm: Figures 5.25 to 5.27 show essentially the same performance with the distorting grid algorithm as Figures 5.18 to 5.20 show for the fixed grid algorithm. As before, Figure 5.25 is formed from the converging p and q values with the assumed, incorrect reflectance function: as the error in the reflectance function increases, the convergence becomes slower, but even with l_1 set to .5 instead of the



a. b.

Figure 5.27. Deforming grid algorithm: incorrect reflectance function: Integrability measure pictures. Noise-free, 400 iterations, $h = 1.0$, correct $l_1 = 0.0$. a. $l_1 = .10$. b. $l_1 = .5$.

true 0.0 much of the image is reconstructed. Figure 5.26 is formed using the true reflectance function, and shows how different the solution surface is from the correct one. Figure 5.27 shows the IM pictures; as the error in the reflectance function increases, the integrability measure becomes higher over certain edges of the image; however, much of the center of the image stays integrable. This algorithm as well seems robust to errors in the reflectance function.

The deforming grid algorithm is considerably slower in our implementation than the fixed grid algorithm discussed in the last section. It does appear to converge to somewhat more integrable sets of normal vectors than the fixed grid method except for the occurrence of "cracks" in the integrability picture due to a sudden shift in the neighborhoods used to calculate tangent planes. It also seems quite robust in the face of noise and distortions in the reflectance function.

The difficult part of doing a direct implementation of the Lambda Lemma intuition is filling in the parts of the deforming surface that are vacated by processors flowing away from the critical point. The deforming grid algorithm we implemented flows for a short distance, and then fills in a new rectangular grid consistent with the deformed grid; we have used a local least squares linear approximation to do the filling in using neighborhoods of the integer grid points. One might try to fit a

splined surface of higher order to the data points given by the distorted grid, and then get the rectangular grid values from this.

5.3.4 Conclusions on Implementation

The successful implementation of the Lambda Lemma intuition provides empirical support for the theoretical approach taken in the first two sections of this chapter. Convex/concave solution surfaces around a maximum critical point in the image due to a critical point maximum of the reflectance function correspond to stable/unstable invariant manifolds of the corresponding critical point in the image dynamical system.

The algorithms derived from the Lambda Lemma intuition appear to be relatively stable both with respect to noise and with respect to errors in the reflectance function. In addition, there are several potentially useful properties of these algorithms. One feature is that they do not require the imposition of integrability at each iteration to find a solution. Integrability is a side-effect of a successful convergence: the unstable manifold of a critical point must obey both $E = R$ everywhere on the surface (it is true at the critical point; all the trajectories of the unstable manifold approach this point as t goes to $-\infty$; and $H = E - R$ is constant on each trajectory), and the integrability condition, since the integrability condition is true on all characteristic trajectories and the invariant manifold is made up of these. One can use integrability of the evolving p and q values to monitor the progress of the algorithm, particularly in the presence of noise. Another feature is that we can pick step sizes for the iterations that are different at each point of the grid because the limiting behavior of the convergence is not dependent on the particular time-evolution. This may allow careful handling of tricky sections in an image.

If an image critical point is on the stable manifold, we can reverse time and use the same techniques. If an image critical point is a saddle critical point (for example, at a maximum of the reflectance function and with saddle surface structure at that point), then the invariant manifold is neither a stable or unstable manifold

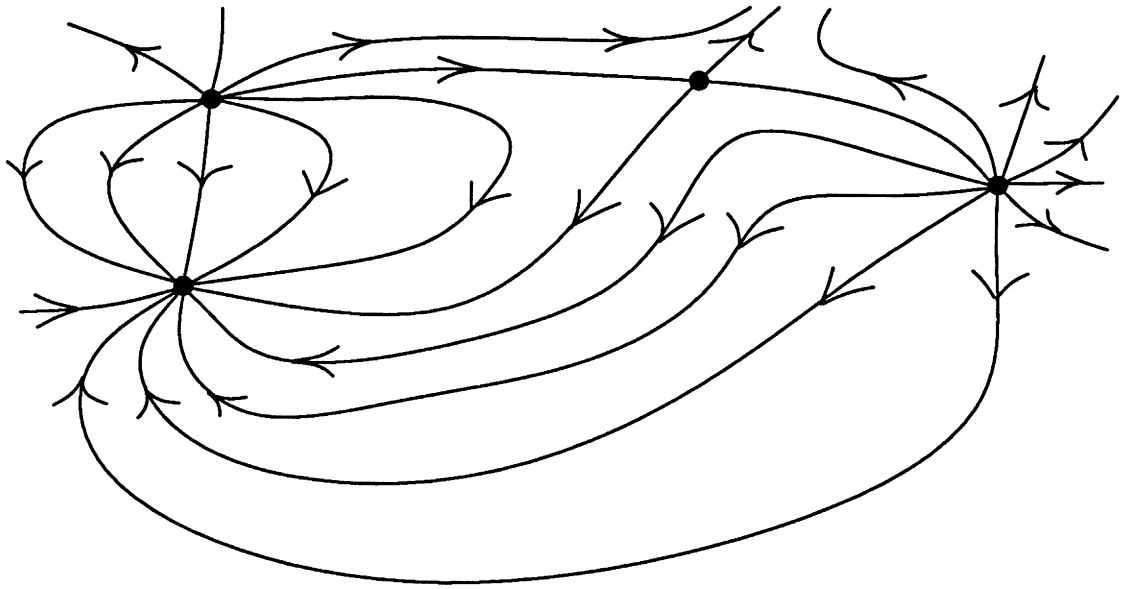


Figure 5.28. 2-d dynamical system restricted to the invariant manifold: in general, most points will be in the stable or unstable manifold of either a source or sink critical point.

for the dynamical system, and the methods described here fail. On the other hand, this saddle invariant manifold has to come from somewhere. A true solution surface will have a two-dimensional restricted image dynamical system defined on its interior. This system will have critical points, and will (probably) be a generic two-dimensional system: this means that almost all the points on the surface will either be in the stable or unstable manifold of a source or sink of the two-dimensional system (Figure 5.28). Almost all the points near the saddle critical point in theory should be reachable by finding the stable and unstable manifolds of other critical points.

The methods described in Sections 5.3.2 and 5.3.3 will only work on some (possibly quite extended) neighborhood of a single critical point. If two critical points are contained in a neighborhood on which these algorithms are applied and the true solution surface is the stable manifold for one of the critical points and the unstable manifold for the other, it is not clear what solution (if any) will be converged to. The Lambda Lemma algorithms will seek simultaneously to find the unstable manifold for both critical points; this will most likely lead to non-convergence.

This suggests dividing the image into regions each centered on a single critical point. We then find the stable and unstable manifolds of each critical point out to the (artificial) boundary of its region, and look to see whether or not these surfaces can be seamlessly merged together.

Treating the shape from shading problem as an image dynamical system leads to new ideas for shape from shading algorithms. These algorithms are very parallel in character, and converge robustly to integrable solution surfaces near the critical points without an externally imposed integrability condition.

Appendix to Chapter 5

A5.1 Similar signatures for B and $A = P^T BP$.

We are interested in the signs of the eigenvalues α_1, α_2 for the 2×2 matrix $A = P^T BP$, where B and P are 2×2 matrices, B is symmetric and P is invertible. Since A is symmetric, we know it has real eigenvalues α_1, α_2. We are interested in the roots of the polynomial

$$\det(I - \lambda A) = \det(I - \lambda P^T BP).$$

Pre-multiplying by P and post-multiplying by P^{-1} does not change the determinant, so we are interested in the roots of the polynomial

$$\det(I - \lambda PP^T B).$$

Since P is invertible, PP^T is positive definite: we have

$$\mathbf{u}^T PP^T \mathbf{u} = (P^T \mathbf{u})^T (P^T \mathbf{u}) \geq 0,$$

and equality only occurs if $P^T \mathbf{u} = 0$, i.e. if $\mathbf{u} = 0$.

We can now use an argument similar to the one in Section 5.2.5: we let Q be a matrix of orthonormal eigenvectors for PP^T, so that $Q\Lambda Q^{-1} = PP^T$ where Λ is a diagonal matrix of the eigenvalues $\gamma_1 > 0$ and $\gamma_2 > 0$ for PP^T. We have

$$\det(PP^T B) = \alpha_1 \alpha_2 = \det(Q\Lambda Q^{-1} B) = \det(\Lambda Q^{-1} BQ).$$

B and $Q^{-1}BQ$, being similar, have the same eigenvalues β_1 and β_2. We have $\alpha_1\alpha_2 = \gamma_1\gamma_2\beta_1\beta_2$; the signs of α_1 and α_2 are the same if and only if the signs of β_1 and β_2 are the same.

We can also look at

$$\begin{aligned}\text{trace}(A) = \alpha_1 + \alpha_2 &= \text{trace}(P^TBP) = \text{trace}(PP^TB) \\ &= \text{trace}(Q\Lambda Q^{-1}B) = \text{trace}(\Lambda Q^{-1}BQ).\end{aligned}$$

If a and b are the diagonal entries of $Q^{-1}BQ$, then we have $\alpha_1 + \alpha_2 = \gamma_1 a + \gamma_2 b$. If B is positive definite (i.e. both β_1 and β_2 posistive), then so is $Q^{-1}BQ$ and hence a and b are both positive (e.g., $a = (1,0)Q^TBQ(1,0)^T > 0$). This means the sum $\alpha_1 + \alpha_2$ is also positive, and since α_1 and α_2 are of the same sign, they must both be positive. Similarly, if β_1 and β_2 are both negative, so are both α_1 and α_2. Finally, if β_1 and β_2 are of opposite sign, so are α_1 and α_2.

A5.2 Eigenvalues of $A = BC$ When B and C Are Symmetric

This is a pictorial look at the following problem (taken in large measure from (Norton, 1988)): say B and C are real 2×2 symmetric matrices. When does their product, $A = BC$, have real eigenvalues?

If A has one complex eigenvalue, then it has two, and $\det A > 0$; thus if A has one real eigenvalue, all its eigenvalues are real. If either B or C is singular, then so is A, and hence A has one real eigenvalue, 0.

Assume both B and C are invertible. We will demonstrate that if either B or C is definite, then A has real eigenvalues. Let us assume that B is definite. If C is not definite, then $\det A = \det B \det C$ is negative, and A must have real eigenvalues.

Assume both B and C are definite. We can look at how B behaves by considering how B acts on the set of straight lines through the origin in the plane: this is also called the projective plane, $\mathbb{P}^1$. We can identify lines through the origin with points on the circle: essentially, we draw a half circle in the plane and look at the

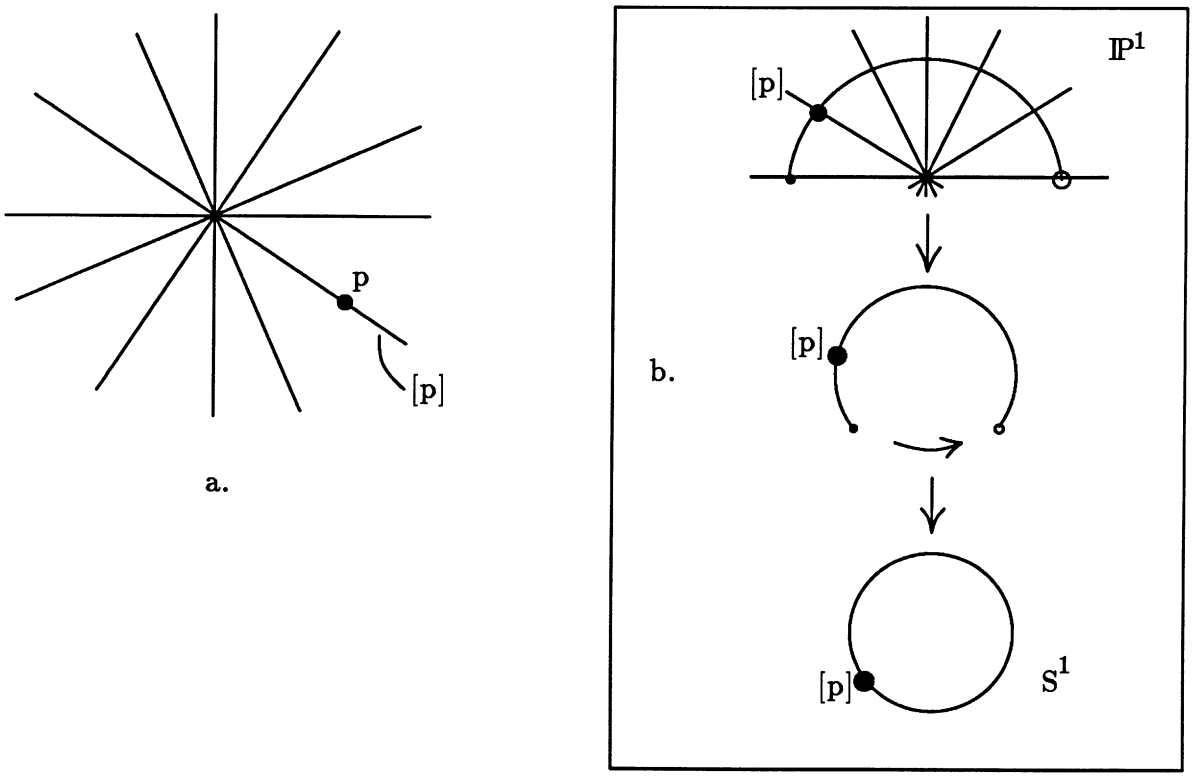


Figure A5.1. a. Lines through the origin. b. S^1 is diffeomorphic to $\mathbb{P}^1$.

intersection of each line through the origin with the circle; we identify the endpoints of the half circle (since the lines through these two points are the same), and this defines topologically a circle (Fig. A5.1). If $p \in \mathbb{R}^2 \backslash 0$, let us define $[p]$ to be the line through the origin containing the point p. B can be considered to act on $\mathbb{P}^1$ since $B\lambda p = \lambda Bp$, so $[Bp] = [B\lambda p]p$ and we can define $B[p] = [Bp]$. If v is an eigenvector for B, then $Bv = \lambda v$, so $B[v] = [v]$, and B leaves the line containing v unchanged. The line through the origin and v is a fixed point of B's action on $\mathbb{P}^1$.

If we take a line $[p] \in \mathbb{P}^1$ we can find out where $B^k[p]$ goes as k goes to infinity. Assuming B is definite, we can pick an orthogonal basis of eigenvectors $\{v_1, v_2\}$ such that in this basis

$$B = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},$$

where λ_1 and λ_2 are the eigenvalues of B with $|\lambda_1| > |\lambda_2|$. (If $\lambda_1 = \lambda_2$, then B is a multiple of the identity, and A has real eigenvalues immediately as a scalar multiple

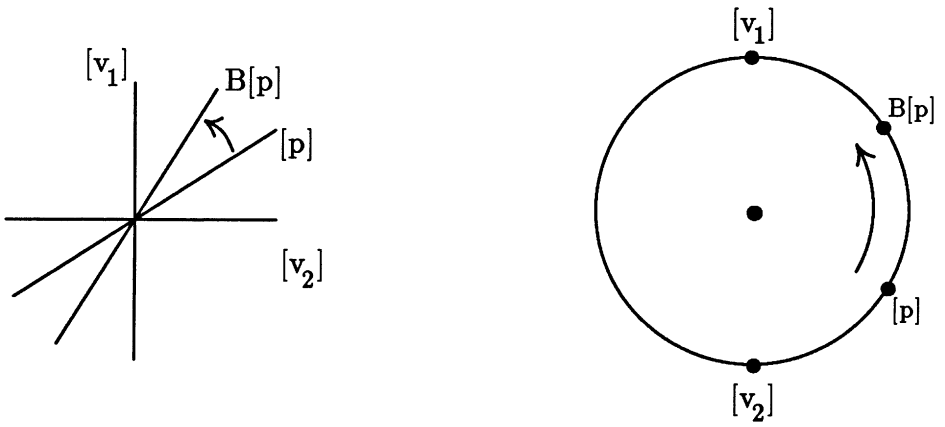


Figure A5.2. Action of symmetric, definite B: a. Lines picture. b. $S^1 \simeq \mathbb{P}^1$ picture. B leaves the eigenspaces $[v_1]$ and $[v_2]$ alone; if $|\lambda_1| > |\lambda_2|$, then for $p \neq v_2$, $\lim_{k\to\infty} B^k[p] = [v_1]$.

of C.) If B is definite, then λ_1 and λ_2 are the same sign; we can multiply B by -1 if necessary to get them positive without affecting the action on lines. In fact, the action of B on lines through the origin is unaffected by scaling B by any constant, so we can look at the action of

$$\bar{B} = \begin{bmatrix} 1 & 0 \\ 0 & \lambda_2/\lambda_1 \end{bmatrix};$$

we have

$$\bar{B}^k = \begin{bmatrix} 1 & 0 \\ 0 & (\lambda_2/\lambda_1)^k \end{bmatrix}.$$

From this we see that if $p \neq v_2$,

$$\lim_{k\to\infty} B^k[p] = \lim_{k\to\infty} \bar{B}^k[p] = [v_1]$$

since $\bar{B}^k$ will squash the v_2 component of p while leaving the v_1 component fixed. Since v_2 is an eigenvector for B, $\lim_{k\to\infty} B^k[v_2] = [v_2]$. Looking at how B^k acts on lines, we see that it moves all of them except $[v_2]$ monotonically towards $[v_1]$ along the half-circle between $[v_1]$ and $[v_2]$ containing p (Figure A5.2a). In Figure A5.2b, we show this on the circle representation of $\mathbb{P}^1$ by putting arrows along the circular

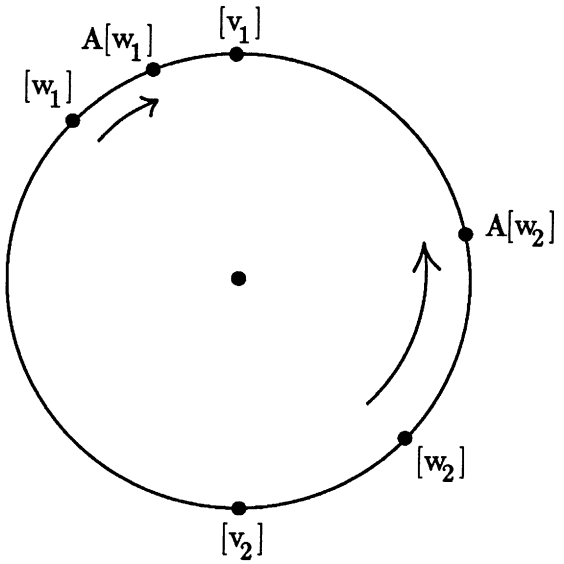


Figure A5.3. Combined action of definite B and C: circle representation. $[v_1]$ and $[v_2]$ are at opposite sides of the circle, and the dynamics of iterations of B is shown again by arrows. The eigenspaces $[w_1]$ and $[w_2]$ for C are placed on the same circle. The combined map $A = BC$ acts as follows: C leaves $[w_i]$ alone, and B moves $[w_i]$ towards $[v_1]$. A must have a fixed point between $A[w_1]$ and $A[w_2]$.

arcs. In a sense, the point $[v_1]$ on the circle $S^1 \simeq \mathbf{IP^1}$ is a sink for the iterations of B on $\mathbf{IP^1}$, while $[v_2]$ acts as a source.

Now we look at the composition BC. C is symmetric, so it also has orthogonal eigenspaces $[w_1]$ and $[w_2]$. Let us assume these are different from $[v_1]$ and $[v_2]$ (otherwise B and C are simultaneously diagonalizeable; hence they commute; hence A is symmetric and has real eigenvalues). In Figure A5.3 we show how the combined action $A = BC$ moves the points $[w_1]$ and $[w_2]$: it moves them both towards the point on the circle $[v_1]$. Since B and C are both definite, they preserve the order of points around the circle, and so does $A = BC$. As a result, A must map the arc between $[w_1]$ and $[w_2]$ containing $[v_1]$ into the smaller arc between $A[w_1]$ and $A[w_2]$. This means A has a fixed point on this interval. To see this, pick a coordinate chart for the circle containing the interval, and say θ_1 is the coordinate on the circle for $[w_1]$ and θ_2 is the coordinate on the circle for $[w_2]$. If we define the function $g(\theta) = A(\theta) - \theta$ (where we think of $A(\theta)$ as the coordinatized version of A), then

$g(\theta_1)$ and $g(\theta_2)$ are of opposite sign. By the intermediate value theorem, there must be a $\theta \in [\theta_1, \theta_2]$ such that $g(\theta) = 0$, so there must be a $[p]$ between $A[w_1]$ and $A[w_2]$ such that $A[p] = [p]$; i.e., A has a fixed point. A fixed point for $A = BC$ acting on lines means that A has a real eigenspace, and hence real eigenvalues.

If B and C are both not definite, then this argument does not work. For example, if

$$B = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$
$$C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

then

$$B^k = \begin{bmatrix} 1 & 0 \\ 0 & -1^k \end{bmatrix}$$

so that $B^k[v]$ will never converge. In this case

$$A = BC = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

is a rotation matrix which does have complex eigenvalues.

A5.3 Eigenvalues of $A = BC$, $C = Q^T BQ$, B symmetric.

In Section 5.2.4 we are interested in the case where $B = D^2R$, and $C = D^2E$; from Section 5.2.2, we know that $C = Q^TBQ$ and hence (Appendix A5.1) has the same signature as B. We want to know when A has positive eigenvalues. We can once again use an argument similar to that in Section 5.2.5.

From Appendix A5.2, we know that if B (and therefore C) is definite, A has real roots. Assume B is definite. We have

$$\det(A) = \alpha_1\alpha_2 = \det(BC) = \beta_1\beta_2\gamma_1\gamma_2,$$

where β_1 and β_2 are the eigenvalues for B and γ_1 and γ_2 are the eigenvalues for C. Since B and C have the same signature, $\alpha_1\alpha_2$ is poistive, so α_1 and α_2 have the same sign. Picking coordinates to diagonalize B, we can write

$$\text{trace}(A) = \alpha_1 + \alpha_2 = \text{trace}\, BC = \beta_1 a + \beta_2 b,$$

where a and b are the diagonal entries of C in this coordinate system. If B (and hence C) is positive definite, then all the values β_1, β_2, a, b are positive, and so α_1 and α_2 must both be positive; if B is negative definite, all the values β_1, β_2, a, b are negative, and α_1 and α_2 are again both positive. Thus, if B is definite, A has positive real roots.

If B is indefinite, this result does not hold. The example at the end of Appendix A5.2 shows that if B and C are both indefinite, BC can have complex eigenvalues. If

$$B = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}$$

$$C = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

then

$$BC = \begin{bmatrix} -2 & 0 \\ 0 & -1 \end{bmatrix}$$

has negative eigenvalues.

Chapter 6

Image Data at the Bounding Contour

In Chapter 5 we discussed the contribution of critical points to the determination of shape from shading. This chapter discusses the contribution of the image data near and at the bounding contour. A motivating example is presented, followed by two different analyses, the first a linearization analysis of the characteristic vector field at the bounding contour, the second a power series approach. These suggest that given a known reflectance function, a patch of bounding contour image data is no more useful than a patch of data from the interior of the image not containing a critical point: different solution surfaces can be determined by the choice of depth values along a curve in the image, in particular the bounding contour. This reinforces the theoretical importance of the critical points on the interior of the image as the major determinants of solution surfaces.

6.1 Introduction

There are two problems with working (either theoretically or practically) with the bounding contour image information: the $\Pi^C : (x, y, z, p, q) \longmapsto (x, y)$ rectilinear coordinate representation for the surface, the image, and the image projection does not have p and q well-defined at the bounding contour; and, although the image

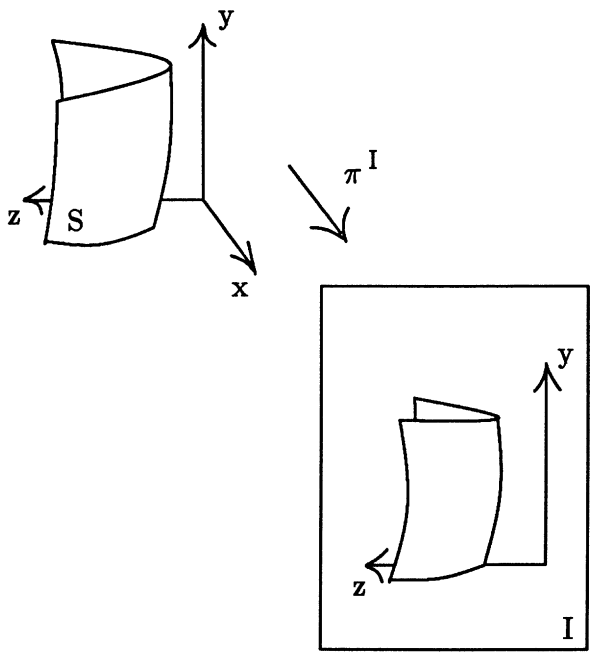


Figure 6.1. The turned standard coordinate system for viewing the bounding contour. The image plane is now the $y - z$ plane.

intensity is theoretically well-defined at the bounding contour, the image brightness derivatives explode.

The first problem is surmountable: our approach has been to try to think in a coordinate-free way and choose coordinates to answer particular questions. In this case, we can turn the coordinates sideways. We use an (x, y, z, p, q) coordinate system as before, with tangent spaces coordinatized by p and q, where

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ p & q \end{bmatrix}$$

spans a tangent space of interest at the point (x, y, z) in space, and the matrix is written using vectors $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}$ as a basis. However, now we take the image projection to be

$$\pi^I : (x, y, z) \longrightarrow (y, z).$$

Effectively, we are projecting down the x direction, and points in $\mathcal{C}(\mathbb{R}^3, 2)$ on the bounding contour all have coordinate $p = 0$ since by definition the $\frac{\partial}{\partial x}$ direction is contained in a surface tangent plane at the bounding contour.

We need to recalculate the coordinate expression for the vector field X defining the characteristic dynamical system. In Chapter 4 we defined the image dynamical system function as $H = E \circ \pi^C - R$, where π^C is the projection from the five dimensional space $\mathcal{C}(\mathbb{R}^3, 2)$ to the image, and R is the reflectance function defined on $\mathcal{C}(\mathbb{R}^3, 2)$. We discussed how finding solution surfaces for the first order partial differential image irradiance equation can be viewed as finding a solution surface for a differential ideal generated by $\mathrm{d}H$ and θ, where θ is the contact 1-form defined in our coordinate system as $\theta = \mathrm{d}z - p\mathrm{d}x - q\mathrm{d}y$. This in turn gave rise to a closed ideal with the same solution surfaces, generated by $\mathrm{d}H, \theta$, and $\mathrm{d}\theta$. This ideal has one-dimensional solution curves, the characteristics, that can be assembled together to make up solution surfaces for the original first order equation. These characteristic curves can be defined by a characteristic vector field X. A characteristic vector field for an ideal of differential forms is one that preserves the ideal under contraction: i.e., if ρ is a differential form in the ideal, then $\mathbf{i}_X\rho$ is in the ideal as well. Thus

$$\mathbf{i}_X\mathrm{d}\theta = \alpha\theta + \beta\mathrm{d}H,$$

since these last two generate all the 1-forms in the image ideal. We can expand both sides with our new coordinate system: we have $H = E(y, z) - R(x, y, z, p, q)$ and collecting terms we get

$$\begin{aligned} X_x\mathrm{d}p - X_p\mathrm{d}x + X_y\mathrm{d}q - X_q\mathrm{d}y = \\ (-\alpha p - \beta R_x)\mathrm{d}x + (-\alpha q + \beta E_y - \beta R_y)\mathrm{d}y + (\alpha + \beta E_z - \beta R_z)\mathrm{d}z \\ - \beta R_p\mathrm{d}p - \beta R_q\mathrm{d}q, \end{aligned}$$

where we have allowed R to be space varying, as we will need the results later on. We can collect coefficients of the basis 1-forms $\mathrm{d}x, \mathrm{d}y, \mathrm{d}z, \mathrm{d}p, \mathrm{d}q$ to get $\alpha = -\beta(E_z - R_z)$, and hence (using $\mathbf{i}_X\theta = X_z - pX_x - qX_y = 0$ to get the components of X_z):

$$X = -\beta \begin{bmatrix} R_p \\ R_q \\ pR_p + qR_q \\ -R_x + p(E_z - R_z) \\ E_y - R_y + q(E_z - R_z) \end{bmatrix}.$$

Even in the space invariant case, $R_x = R_y = R_z = 0$, this vector field "looks" different from the characteristic vector field used to analyze the critical point case in Chapter 5; this is because the coordinates have been turned to allow the bounding contour to be contained in the coordinate chart.

The other problem with working at the bounding contour, the fact that derivatives of the image intensities explode as shown in Section 3.2.1, is more difficult to handle. Infinities keep cropping up when dealing with the bounding contour. Our analysis in Chapter 5 of the critical point case depended on the dynamical system generated by a vector field defined on $\mathcal{C}(\mathbb{R}^3, 2)$. This vector field depends on the image derivatives no matter what the coordinate system, as can be seen from the expression derived in our new coordinate system in the previous paragraph, and these image derivatives blow up in general at the bounding contour. The analogy with the behavior of $\sqrt{x}$ as x goes to zero is very strong, as we shall see.

6.2 A Motivating Example

To begin to get an understanding of what might happen at the bounding contour, we can work through an example. Consider a surface defined by

$$z = \frac{1}{2}(ax^2 + by^2),$$

a paraboloid. (Figure 6.2) At each point (x, y, z) on the surface, we will have

$$p = z_x = ax$$

$$q = z_y = by.$$

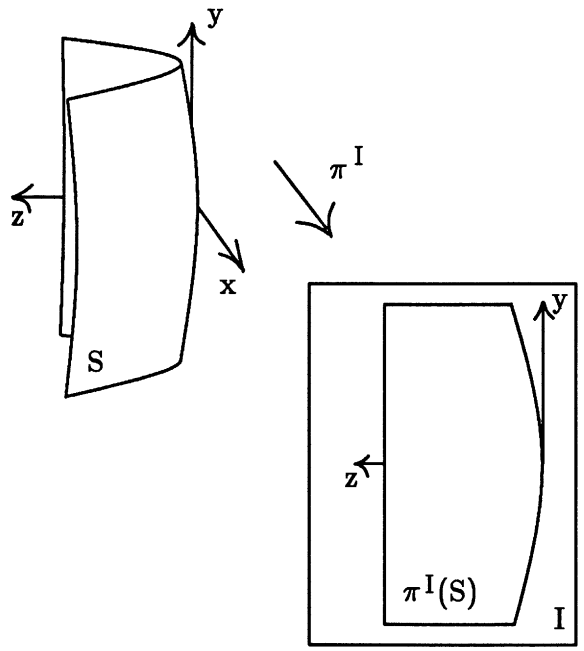


Figure 6.2. Example of bounding contour analysis: a paraboloid. The $y-z$ plane is the projection plane.

Consider also a very simple reflectance function,

$$R(p,q) = pl_1 + ql_2.$$

The image in our turned coordinate system will be

$$E(y,z) = l_1 a x(y,z) + l_2 by;$$

note that we must express x as a function of the image coordinates (y,z) in order to define the image. Since

$$x = +\frac{1}{\sqrt{a}}\sqrt{2z - by^2},$$

assuming we are interested only in the view of the positive x sheet of the surface (remember that we are projecting along the x direction), we have

$$E(y,z) = \frac{l_1 a}{\sqrt{a}}\sqrt{2z - by^2} + l_2 by.$$

The bounding contour in the image will be places in the image (the y–z plane) where the surface has $p = 0$; i.e. places where $x(y,z) = (1/\sqrt{a})\sqrt{2z - by^2} = 0$. These points lie on the parabola $z = \frac{1}{2}by^2$ in the image.

We can find the components of the characteristic vector field using the expressions developed last section and taking $\beta = -1$:

$$X = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \\ \dot{p} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} l_1 \\ l_2 \\ pl_1 + ql_2 \\ p\left(\dfrac{l_1\sqrt{a}}{\sqrt{2z - by^2}}\right) \\ \dfrac{-l_1\sqrt{a}by}{\sqrt{2z - by^2}} + l_2 b + q\left(\dfrac{l_1\sqrt{a}}{\sqrt{2z - by^2}}\right) \end{bmatrix}.$$

As the bounding contour is approached, both p and $\sqrt{2z - by^2}$ approach 0; thus, the vector field is undefined at the bounding contour.

The behavior is similar to the behavior of the vector field $X = (x/y, 1)^T$ on $\mathbb{R}^2$: this two dimensional vector field is also undefined at the origin. In this case, we can get out of difficulty by multiplying the vector field by the scalar y to get a new vector field parallel everywhere (except at the origin) to the old one and given by $X = (x, y)^T$. This is well behaved on the plane, but has a critical point at the origin.

We can multiply the characteristic vector field of our image dynamical system by $v = \sqrt{2z - by^2}$ and consider instead the o.d.e

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \\ \dot{p} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} l_1 v \\ l_2 v \\ v(pl_1 + ql_2) \\ pl_1\sqrt{a} \\ l_2 bv + l_1\sqrt{a}(q - by) \end{bmatrix},$$

Since $p = 0$, $v = 0$, and $q - by = 0$ at the bounding contour, from the expression for our vector field we see that the bounding contour is a curve of critical points. The multiplication by v has made the vector field more tractable. If we replace the $\dot{z}$ equation with one for $\dot{v}$, and substitute $z = (1/2)(v^2 + by^2)$, we will have made a

new related vector field that is actually Lipschitz at the bounding contour. We have $z = (1/2)(v^2 + by^2)$, so

$$v(pl_1 + ql_2) = \dot{z} = v\dot{v} + by\dot{y}.$$

Using the vector field for $\dot{y}$, we can solve for $\dot{v}$ to get the vector field

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{v} \\ \dot{p} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} l_1 v \\ l_2 v \\ pl_1 + ql_2 - bl_2 y \\ pl_1\sqrt{a} \\ l_2 bv + l_1\sqrt{a}(q - by) \end{bmatrix}.$$

We can solve for $p(t) = p_0 \exp(l_1\sqrt{a}t)$. We also note that

$$\dot{q} = l_2 bv + l_1\sqrt{a}(q - by) = b\dot{y} + l_1\sqrt{a}(q - by),$$

so

$$\frac{\mathrm{d}}{\mathrm{d}t}(q - by) = l_1\sqrt{a}(q - by),$$

and we have

$$q - by = c_0 \exp(l_1\sqrt{a}t).$$

We also have $v^2 = 2z - by^2$, so

$$\begin{aligned} v\dot{v} &= \dot{z} - by\dot{y} \\ &= v(pl_1 + ql_2) - l_2 byv \\ \dot{v} &= pl_1 + l_2(q - by) \\ &= l_1 p_0 \exp(l_1\sqrt{a}t) + l_2 c_0 \exp(l_1\sqrt{a}t) \\ &= (l_1 p_0 + l_2 c_0)\exp(l_1\sqrt{a}t); \end{aligned}$$

hence

$$v = v_0 + \frac{l_1 p_0 + l_2 c_0}{l_1\sqrt{a}}(\exp(l_1\sqrt{a}t) - 1).$$

From this and the vector field we can solve for x and y:

$$x(t) = d_0 + v_0 l_1 t + l_1 \frac{l_1 p_0 + l_2 c_0}{l_1\sqrt{a}} \left(\frac{\exp(l_1\sqrt{a}t)}{l_1\sqrt{a}} - t \right)$$
$$y(t) = e_0 + v_0 l_2 t + l_2 \frac{l_1 p_0 + l_2 c_0}{l_1\sqrt{a}} \left(\frac{\exp(l_1\sqrt{a}t)}{l_1\sqrt{a}} - t \right).$$

Finally, we can solve for

$$\begin{aligned} q(t) &= (q - by)(t) + by(t) \\ &= c_0 \exp(l_1\sqrt{a}t) + be_0 + bv_0 t + \frac{b(l_1 p_0 + l_2 c_0)}{l_1\sqrt{a}} \left(\frac{\exp(l_1\sqrt{a}t)}{l_1\sqrt{a}} - t \right) \\ z(t) &= (1/2)(v^2 + by^2). \end{aligned}$$

We can also find the implicit equations of the characteristic curves from the above parameterized equations: we have $p(t) = p_0 \exp(l_1\sqrt{a}t)$, so

$$t = \frac{1}{l_1\sqrt{a}} \ln(p/p_0),$$

and therefore

$$\begin{aligned} q &= by + \frac{p}{p_0} c_0 \\ v &= v_0 + \frac{l_1 p_0 + l_2 c_0}{l_1\sqrt{a}}((p/p_0) - 1) \\ x &= d_0 + l_1 v_0 \frac{1}{l_1\sqrt{a}} \ln(p/p_0) + l_1 \frac{l_1 p_0 + l_2 c_0}{l_1^2 a}((p/p_0) - \ln((p/p_0))) \\ y &= e_0 + l_2 v_0 \frac{1}{l_1\sqrt{a}} \ln(p/p_0) + l_2 \frac{l_1 p_0 + l_2 c_0}{l_1^2 a}((p/p_0) - \ln(p/p_0)). \end{aligned}$$

We can rearrange and write these equations in a different way. We can define functions $\eta_1, ..., \eta_4$ as real functions on $\mathcal{C}(\mathbb{R}^3, 2)$ that have a constant value on a trajectory: they are called integral invariants of the flow defined by the vector field. Each equation $\eta_i(x, y, v, p, q) = \kappa_i$ defines a hypersurface in $\mathcal{C}(\mathbb{R}^3, 2)$; on the interior of the image (i.e. away from $p_0 = 0$) these will be transverse to each other, so that the intersection of all four hypersurfaces in the five dimensional $\mathcal{C}(\mathbb{R}^3, 2)$ will be a one dimensional curve, the path of the characteristic trajectory.

We begin by defining

$$\eta_1 = \frac{q - by}{p},$$

and we see from the above implicit equations that on a characteristic curve, we have

$$\eta_1 = \frac{c_0}{p_0} = \kappa_1.$$

From the implicit equation for v, we have

$$\begin{aligned} v &= v_0 + \frac{l_1 p_0 + l_2 c_0}{l_1\sqrt{a}}\left(\frac{p}{p_0} - 1\right) \\ &= v_0 + \frac{l_1 p + l_2(q - by)}{l_1\sqrt{a}} - \frac{l_1 p_0 + l_2 c_0}{l_1\sqrt{a}}, \end{aligned}$$

where we have used the fact that $q - by = c_0 p/p_0$. We can define

$$\eta_2 = v - \frac{l_1 p + l_2(q - by)}{l_1\sqrt{a}},$$

and we have $\eta_2 = \kappa_2$ along a characteristic curve. If this is compared with the expressions for $E(y,z)$ and $R(p,q)$ early in the example, it can be seen that $\eta_2 = H/(l_1\sqrt{a})$, where $H = E(y,z) - R(p,q)$ is the coordinate representation of the function $H : \mathcal{C}(\mathbb{R}^3, 2) \to \mathbb{R}$ that defines the image dynamical system. We know that H must be a constant on trajectories of the system; in this example, it has explicitly emerged.

We can simplify the coordinate representation of η_2 by explicitly using the definition of η_1:

$$\eta_2 = v - \frac{p(l_1 + l_2\eta_1)}{l_1\sqrt{a}}.$$

We can perform the same kind of rearrangement and substitution on the implicit equation for x to get

$$\begin{aligned} \eta_3 &= x - l_1\left(v - \frac{l_1 p + l_2(q - by)}{l_1\sqrt{a}}\right)\frac{\ln p}{l_1\sqrt{a}} - l_1\frac{l_1 p + l_2(q - by)}{l_2^2 a} \\ &= x - l_1\eta_2\frac{\ln p}{l_1\sqrt{a}} - \frac{l_1}{l_1\sqrt{a}}(v - \eta_2), \end{aligned}$$

with, again, $\eta_3 = \kappa_3$ along a characteristic curve. Finally, rather than using the equation for y to generate a fourth invariant, we can observe (either from the original differential equation or from the implicit solutions themselves) that

$$\eta_4 = l_2 x - l_1 y$$

is a constant along the characteristic trajectories; this and the previous implicit equation defining η_3 give the implicit equation for y.

We write these four invariants here for reference:

$$\begin{aligned}
\eta_1 &= \frac{q - by}{p} \\
\eta_2 &= v - \frac{p(l_1 + l_2\eta_1)}{l_1\sqrt{a}} \\
\eta_3 &= x - \frac{\eta_2}{\sqrt{a}}(\ln p + 1) - \frac{v}{\sqrt{a}} \\
\eta_4 &= l_2 x - l_1 y.
\end{aligned}$$

Given $p \neq 0$, and the constants κ_i, we can solve sequentially for the other values x, y, v, q :

$$\begin{aligned}
v &= \kappa_2 + \frac{p(l_1 + l_2\kappa_1)}{l_1\sqrt{a}} \\
x &= \kappa_3 + \frac{\kappa_2}{\sqrt{a}}(\ln p + 1) + \frac{v}{\sqrt{a}} \qquad (\dagger) \\
y &= -\frac{1}{l_1}\kappa_4 + \frac{l_2}{l_1}x \\
q &= by + p\kappa_1.
\end{aligned}$$

For $p \neq 0$, the four invariants $\eta_i = \kappa_i$ describe the one dimensional characteristic curves.

These equations define the characteristic curves for this image dynamical system. We are interested in behavior near the bounding contour, places where $p = 0$ in our coordinate system. We can, for example, find out which contours will actually approach the bounding contour by examining what happens with the invariants as

p approaches 0. From the definition of η_1, assuming a finite κ_1, we must have $q - by \rightarrow 0$ too. From the definition of η_2 we also see that $v \rightarrow \kappa_2$. From the definition of η_3, in order for κ_3 to be finite, we must have $\kappa_2 = 0$; since η_2 is an invariant of our characteristics, it is not a question of η_2 approaching zero along a characteristic; we must have $\eta_2 = \kappa_2$ identically equal to 0 along the characteristic in order for κ_3 to be finite. This forces $v \rightarrow 0$ and $x \rightarrow \kappa_3$ as $p \rightarrow 0$. Finally, we have $y \rightarrow (\kappa_4/l_1) - (l_2/l_1)\kappa_3$

The assumption that the bounding contour of the image in this example is due to the projection of finite points on characteristic trajectories has forced $\eta_2 = 0$ on trajectories that satisfy this. Since $\eta_2 = H$, the image dynamical system function, this geometrical assumption has restricted us precisely to the characteristics that are consistent with the image: those with $H = 0$, i.e. with $E \circ \Pi^C = R$ along them.

Consider a curve in space parameterized as

$$s \longmapsto (x^0(s), s, \frac{1}{2}s^2, 0, bs),$$

with $x^0(s)$ an arbitrary function. This curve will project to the image of the bounding contour of our simple surface, and also has the same normal along the bounding contour in space as our original surface. Which characteristics will approach this curve? We can parameterize the set of characteristics by s where $(\kappa_1, \ldots, \kappa_4)(s)$ for a fixed s describes a characteristic that approaches the point $(x^0(s), s, (1/2)s^2, 0, bs)$ on the bounding contour. From the above calculation as $p \rightarrow 0$, we see that

$$\kappa_3(s) = x^0(s)$$
$$\kappa_4(s) = l_2 x^0(s) - l_1 s$$
$$\kappa_2(s) = 0$$

and apparently arbitrary $\kappa_1(s)$ will give characteristic contours that approach the given bounding contour in $\mathcal{C}(\mathbb{R}^3, 2)$.

The degeneracy of the image at the bounding contour is shown again by the fact that κ_1 appears to be arbitrary in determining characteristics that approach

a point on the bounding contour. There appears to be a one-parameter family of characteristic contours (with κ_1 as parameter) that approach a given point on the contour.

However, from the equations (†) giving x, y, v, and q in terms of p for a given choice of the κ_i, we see that the space coordinates x, y, and v (and hence z) are determined solely by $\kappa_2 = 0, \kappa_3$, and κ_4; the "ambiguous" κ_1 does not appear to enter in the determination of the space part of any of the characteristic curves approaching the bounding contour. This leads one to suspect that in this example κ_1 must in some way be determined by the integrability condition $\mathrm{d}z = p\mathrm{d}x + q\mathrm{d}y$.

In the interior of the image, the choice of an initial strip determines the values of both H and the contact 1-form θ at each initial point along the strip; these values are constant along each (unique) characteristic emanating from that initial point. In our example, at the bounding contour we have a more degenerate situation: we have a one-parameter family of characteristics emanating (in the limit) from each point on the bounding contour. Using $z = (1/2)(v^2 + y^2)$, the integrability constraint along the bounding contour is

$$z' = vv' + byy' = px' + qy'$$
$$byy' = byy',$$

where x', etc. are derivatives with respect to s along the bounding contour, and we use the fact that $v = p = 0$ and $q = by$ along the bounding contour to get the second equation. Thus, the integrability condition $z' = px' + qy'$ is apparently satisfied for the whole 1-parameter family of different trajectories emanating from a single point on the bounding contour. This gives us the apparent ambiguity in κ_1.

The calculation of the integrabilty equation $z' = px' + qy'$ at the bounding contour is not quite justified, however, since at the bounding contour our invariant characterization $\eta_i = \kappa_i$ of the characteristic trajectories breaks down. We have to be more careful with the limiting behavior of the integrability condition.

We can examine the integrability condition on the interior of the image near the bounding contour along a curve $(x, y, v, p, q)(s)$ that runs transverse to the characteristics. We have functions $\kappa_i(s)$ defined as the characteristics are crossed. We know that in order for these characteristics to approach the bounding contour in the limit we must have $\kappa_2(s) = 0$ for all s. We can write the integrability condition using the flow invariants $\eta_i = \kappa_i$ and substituting for v, v', and y' using different equations from (†):

$$z' = vv' + byy' = px' + qy'$$

$$vv' = px' + (q - by)y' = p(x' + \kappa_1 y')$$

$$\frac{p(l_1 + l_2\kappa_1)}{l_1\sqrt{a}}(\sqrt{a}x' - \sqrt{a}\kappa_3') = p\left(x' + \kappa_1\left(\frac{l_2}{l_1}x' - \frac{\kappa_4'}{l_1}\right)\right).$$

Solving for κ_1,

$$\kappa_1 = \kappa_3' \cdot \frac{1}{(1/l_1)\kappa_4' - (l_2/l_1)\kappa_3'}.$$

Thus, κ_1 is determined by the integrability condition on the interior of the image together with derivatives of the curves $\kappa_3(s)$ and $\kappa_4(s)$. We can take the limit of this as our transverse contour approaches the bounding contour: in this case, we have $x(s) = x^0(s)$, $\kappa_3(s) = x^0(s)$, and $\kappa_4(s) = l_2x^0(s) - l_1s$, and so we see that

$$\kappa_1(s) = -x^{0\prime}(s).$$

This is another sign of the degeneracy at the bounding contour: we need to pass to the derivative of the bounding contour data to actually tie down $\kappa_1(s)$, in contrast to the image interior where $\kappa_1(s)$ is determined just by the values on the initial strip at each s.

Note that a large family of solution surfaces gives the same image as our original paraboloid: we can pick any initial depth curve $x^0(s)$ along the bounding contour and find a solution surface containing it. The equation for the solution surface can be found from the equations (†) relating x and y to v and the invariant coefficients

κ_i. Since $v = \sqrt{2z - by^2}$, by squaring and solving we can find an expression for z as a function of x and y along the characteristic trajectory fixed by s. From the equation relating y and x we can solve for s as a function of y and x (except perhaps where $x^0(s)$ has zero derivative) and substitute this into the preceding equation to get the general equation of the surface. This surface will not fold over itself (except where x^0 does), and so describes a solution surface consistent with the image data and reflectance function assumed.

For example, we can assume $x^0(s) = \alpha s$. Then

$$\kappa_1(s) = -\alpha$$
$$\kappa_2(s) = 0$$
$$\kappa_3(s) = \alpha s$$
$$\kappa_4(s) = (l_2\alpha - l_1)s,$$

and

$$v = \frac{p(l_1 - l_2\alpha)}{l_1\sqrt{a}}$$
$$x = \alpha s + (v/\sqrt{a})$$
$$y = \frac{l_2}{l_1}x - \frac{(l_2\alpha s - l_1 s)}{l_1}$$
$$q = by - p\alpha.$$

We can solve for s:

$$s = \frac{x - (v/\sqrt{a})}{\alpha},$$

and we can write

$$y = \frac{l_2 x}{l_1} - \frac{(l_2\alpha - l_1)}{l_1}\frac{(x - v/\sqrt{a})}{\alpha}$$
$$l_1 y = \frac{l_2}{\sqrt{a}}v + \frac{l_1}{\alpha}x - \frac{l_1}{\alpha\sqrt{a}}v$$
$$v = \frac{l_1(y - \frac{x}{\alpha})\sqrt{a}}{(l_2 - \frac{l_1}{\alpha})}.$$

The depth for this new surface is given by $z = \frac{1}{2}(v^2 + by^2)$. We can look directly at the image of the new surface to see if it is the same as the image of the original surface: we have $z_x = vv_x$ and $z_y = vv_y + by$, so

$$\begin{aligned} E(y,z) &= l_1 z_x + l_2 z_y = v(l_1 v_x + l_2 v_y) + l_2 by \\ &= \frac{v l_1 \sqrt{a}}{(l_2 - \frac{l_1}{\alpha})} \left(\frac{-l_1}{\alpha} + l_2 \right) + l_2 by \\ &= l_1 v \sqrt{a} + l_2 by \\ &= l_1 \sqrt{a} \sqrt{2z - by^2} + l_2 by, \end{aligned}$$

which is the same as the image of the original paraboloid. .

How does this motivating example connect with the critical point theory worked out in Chapter 5? It turns out that there is no "good" critical point on the interior of the image: our reflectance function does not have a critical point, and indeed the image does not have a critical point anywhere either (e.g. no "bad" critical points caused by parabolic lines as discussed in Section 5.2.1). Thus there are no stable/unstable or saddle manifolds to constrain the possible solution surfaces, and so it is quite possible to have a large class of globally consistent solutions.

Several features of this example will be important in the more general work to follow. The simple geometry of the original surface considerably helped the analysis: we were able easily to calculate the actual image intensities and find the derivatives. The extra function v appears in the denominator of the original vector field and is useful in analyzing it. v contains a square root responsible for the finite limit of the image intensities as the bounding contour is approached while giving infinite image derivatives. We multiplied the whole vector field by v and effectively replaced the z coordinate with v to generate a Lipschitz vector field which has a curve of critical points along the bounding contour; including the depth coordinate x, we actually have a plane of critical points for the new vector field. We will try to follow this spirit of analysis for the more general case.

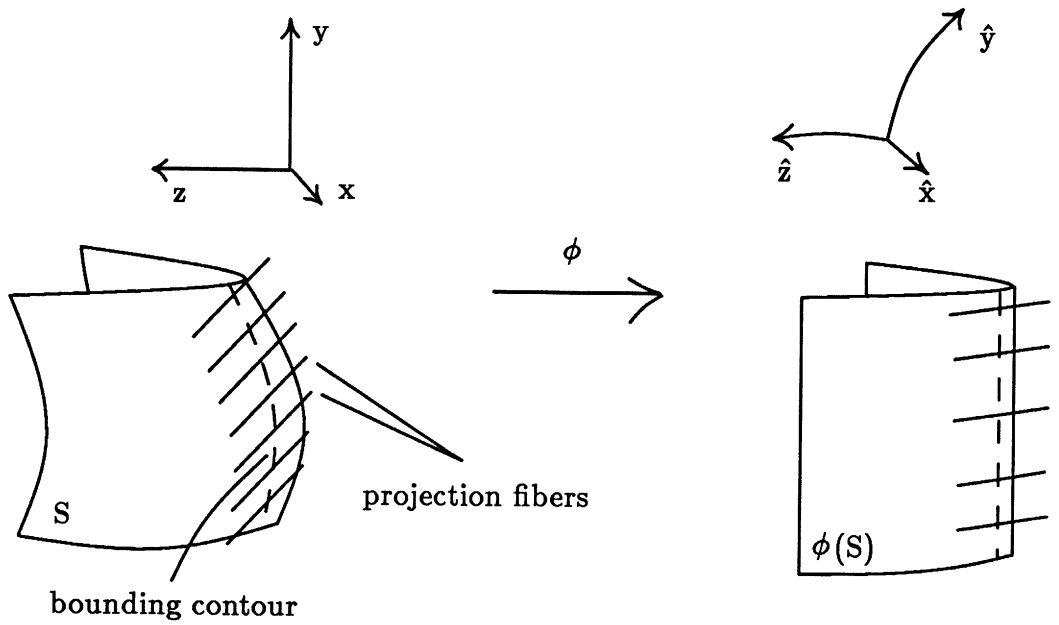


Figure 6.3. Surface, bounding contour, and twisted coordinate system.

6.3 The General Case: Linearization Approach

From Whitney's theorem (Golubitsky and Guillemin, 1973) about the generic nature of singularities of maps between two dimensional smooth surfaces, we know that there are two kinds of singularities generically: a fold and a cusp. The bounding contour points on the image are curves of fold points except where they appear to end in the image interior: these are cusp points (which we will not be examining).

Here we look at the image near generic fold points of the surface $z = f(x, y)$. We will assume that there is a smooth bounding contour visible in the neighborhood of the origin, that the surface does not extend in the negative z direction from the bounding contour (the surface runs to the "left" of the bounding contour), and that we see the image only of the more positive x sheet of the surface.

We will generate a new local set of coordinates for space and the image in which the surface near and at the bounding contour has a particularly simple coordinate representation (this is essentially what Whitney proved to be possible): $z = \frac{1}{2}x^2$. This new (x, y, z) coordinate system generates a new (p, q) system for the surface normals

at each point (in Section 3.1.2 we saw that the (x, y, z, p, q) coordinates for $\mathcal{C}(\mathbb{R}^3, 2)$ works for curvilinear (x, y, z) coordinates as well). In the new (x, y, z, p, q) coordinate representation, the reflectance function will no longer be (x, y, z)-invariant, and so the characteristic vector field will include terms involving space derivatives.

The required coordinate change can more intuitively be thought of as a transformation of the original surface near the origin from a surface $z = f(x, y)$ to a surface $z = \frac{1}{2}x^2$. We assume that our original surface contains the origin and has a fold point bounding contour element at the origin. We want to maintain the image projection as simple orthographic projection along the x coordinate direction. This restricts the transformations we can use to ones which map the null fibres $\{(\lambda, y, z) \text{ for all } \lambda\}$ of the image projection to themselves. Thus, if ϕ is such a transformation with $\phi(x, y, z) = (\hat{x}, \hat{y}, \hat{z})$, we have

$$\pi^I(\phi(x, y, z)) = (\hat{y}, \hat{z}) = \pi^I(\phi(\lambda, y, z)),$$

for all $\lambda \in \mathbb{R}$, and so in the new coordinate system $(\hat{x}, \hat{y}, \hat{z})$ the image projection still is projection down the $\hat{x}$ direction.

We can transform the surface in stages: first, we straighten out the bounding contour. Say the bounding contour of the original surface is $(x^0(s), s, z^0(s))$ parameterized with respect to the y coordinate (we may need a rotation around the x axis to make this possible). Consider the map

$$\psi(x, y, z) = (x - x^0(y), y, z - z^0(y)).$$

For fixed y and z, this takes points (λ, y, z) to points $(\lambda - x^0(y), y, z - z^0(y))$, so as required it maps fibres of the orthogonal image projection to other such fibres. It also maps the curve in space $(x^0(s), s, y^0(s))$ to the curve $(0, s, 0)$, so we have succeeded in straightening out the bounding contour.

We now want to transform the surface near this straightened bounding contour into parabolic form. Let us say that we currently have $z = f(x, y)$, with straight

bounding contour $(0, y, 0)$. Let us fix the level $y = c$, and restrict our attention to the (x, z) plane at this level. Suppressing the y coordinate, we seek a transformation ϕ : $(x, z) \longmapsto (\phi_1(x, z), \phi_2(x, z))$ that takes the curve (i.e. the slice of the surface at $y = c$) $z = f(x)$ to the curve $z = \frac{1}{2}x^2$, and that takes projection fibres $\{(\lambda, z)|$ for all $\lambda \in \mathbf{IR}\}$ to other projection fibres. The latter condition means that $\phi_2(x, z) = \phi_2(z)$, i.e. ϕ_2 is independent of x. We can take $\phi_2(x, z) = z$ for example. We now want

$$\frac{1}{2}(\phi_1(x, f(x)))^2 = \phi_2(f(x)) = f(x)$$

on the surface, so $\phi_1(x, f(x)) = \sqrt{2f(x)}$. We can take $\phi_1(x, z) = \phi_1(x) = \sqrt{2f(x)}$, and our full transformation is:

$\phi(x, y, z) = (\sqrt{2f(x, y)}, y, z)$.

Looking at the positive x sheet of f, is this transformation smooth at the origin? We can look at a power series expansion near the origin:

$$f(x, y) = a_1x + a_2y + a_3x^2 + a_4xy + a_5y^2 + \cdots$$

Since $f(0, y) = 0$ and $p(0, y) = f_x(0, y) = 0$ for all y in the region of interest, we must have $f(x, y) = x^2g(x, y)$ for some analytic $g(x, y)$: the coefficients of terms involving y^i or xy^i must be 0. We assume that $a_3 = g(0, 0) \neq 0$ so that f does have second order behavior and $g = a_3 + h(x, y)$, where $h(x, y)$ is first order in x and y. We can write

$$\begin{aligned}\sqrt{f(x, y)} &= |x|(a_3 + h(x, y))^{1/2} \\ &= |x|(\sqrt{a_3} + (1/2\sqrt{a_3})h(x, y)) + \cdots,\end{aligned}$$

using the power series expansion for the square root. Looking only at the positive (visible) sheet, $x \geq 0$, $\sqrt{f(x, y)}$ will have (one-sided) derivatives of all orders if f is smooth.

Looking at the diffeomorphism ϕ on the visible sheet of the surface, if a_3 is not 0 the x component of the vector $\frac{\partial}{\partial x}\phi$ is not zero either, and

$$D\phi = \begin{pmatrix} & 0 & 0 \\ \frac{\partial}{\partial x}\phi & 1 & 0 \\ & 0 & 1 \end{pmatrix}$$

will have full rank at the origin; thus, by the inverse function theorem, ϕ is a diffeomorphism of the visible sheet at the origin.

This transformation will take the visible part of a surface $z = f(x, y)$ with bounding contour $(0, s, 0)$ and map it to a surface $z = \frac{1}{2}x^2$. By composing the first bounding contour straightening map ψ with ϕ, we have a smooth map that takes our original surface with bounding contour $(x^0(s), s, z^0(s))$ to the surface $z = \frac{1}{2}x^2$, mapping bounding contours to bounding contours, and preserving the image projection fibres. If we think of this combined map as a coordinate change, how are the (p, q) coordinates affected?

Here is one way to think of the effect: say we start with some other surface $z = g(x, y)$ with $p = g_x$ and $q = g_y$ at some point (x, y, z). Say we transform space coordinates with some diffeomorphism $\Phi(x, y, z) = (\hat{x}, \hat{y}, \hat{z})$. In the new coordinates, we will have $\hat{z} = \hat{g}(\hat{x}, \hat{y})$ defining the transformed surface, assuming that in the new coordinates the $\frac{\partial}{\partial \hat{z}}$ direction is not contained in the new surface. We must have $\hat{p} = \frac{\partial}{\partial \hat{x}}\hat{g}$, $\hat{q} = \frac{\partial}{\partial \hat{y}}\hat{g}$. We can write the surface points in the new coordinate system as $\Phi(x, y, g(x, y)) = \Phi \circ (I \times g)(x, y)$, where $I(x, y) = (x, y)$ is the identity map. Thus

$$(\hat{x}, \hat{y}, \hat{g}(\hat{x}, \hat{y})) \stackrel{\triangle}{=} (\Phi_1 \circ (I \times g)(x, y), \Phi_2 \circ (I \times g)(x, y)),$$

where Φ_1 gives the first two coordinates and Φ_2 gives the last coordinate of Φ. We can solve for $\hat{g}$:

$$\hat{g}(\hat{x}, \hat{y}) = \Phi_2 \circ (I \times g) \circ (\Phi_1 \circ (I \times g))^{-1}(\hat{x}, \hat{y}).$$

We can now use the chain rule to get expressions for $\hat{p} = \hat{g}_{\hat{x}}$ and $\hat{q} = \hat{g}_{\hat{y}}$.

Let us do this for each of the two diffeomorphisms defined above, ψ and ϕ. We have $\psi(x, y, z) = (x - x^0(y), y, z - z^0(y))$, so

$$\psi_1 \circ (I \times g)(x, y) = (x - x^0(y), y)$$
$$D(\psi_1 \circ (I \times g)) = \begin{bmatrix} 1 & -x^{0\prime} \\ 0 & 1 \end{bmatrix}$$

$$
\begin{aligned}
(D(\psi_1 \circ (I \times g)))^{-1} &= \begin{bmatrix} 1 & x^{0\prime} \\ 0 & 1 \end{bmatrix} \\
\psi_2 \circ (I \times g)(x,y) &= g(x,y) - z^0(y) \\
D(\psi_2 \circ (I \times g)) &= (g_x, g_y - z^{0\prime}) \\
D\hat{g} &= D(\psi_2 \circ (I \times g))(D(\psi_1 \circ (I \times g)))^{-1} \\
(\hat{g}_{\hat{x}}, \hat{g}_{\hat{y}}) &= (g_x,\ g_x x^{0\prime} + g_y - z^{0\prime})
\end{aligned}
$$

hence

$$(\hat{p}, \hat{q}) = (p,\ px^{0\prime} + q - z^{0\prime}).$$

In the second case, we assume we have an image due to a surface $z = f(x,y)$ with bounding contour $(0,y,0)$. We have $\phi(x,y,z) = (\sqrt{2f(x,y)}, y, z)$, so

$$
\begin{aligned}
\phi_1 \circ (I \times g)(x,y) &= (\sqrt{2f(x,y)}, y) \\
D(\phi_1 \circ (I \times g)) &= \begin{bmatrix} \dfrac{f_x}{\sqrt{2f(x,y)}} & \dfrac{f_y}{\sqrt{2f(x,y)}} \\ 0 & 1 \end{bmatrix} \\
(D(\phi_1 \circ (I \times g)))^{-1} &= \begin{bmatrix} \dfrac{\sqrt{2f(x,y)}}{f_x} & -\dfrac{f_y}{f_x} \\ 0 & 1 \end{bmatrix} \\
\phi_2 \circ (I \times g)(x,y) &= g(x,y) \\
D(\phi_2 \circ (I \times g)) &= (g_x, g_y) \\
D\hat{g} &= D(\phi_2 \circ (I \times g))(D(\phi_1 \circ (I \times g)))^{-1} \\
(\hat{g}_{\hat{x}}, \hat{g}_{\hat{y}}) &= \left(\frac{g_x}{f_x}\sqrt{2f(x,y)},\ -\frac{g_x f_y}{f_x} + g_y \right)
\end{aligned}
$$

hence

$$(\hat{p}, \hat{q}) = \left(\frac{p\sqrt{2f(x,y)}}{f_x},\ \frac{-pf_y}{f_x} + q \right).$$

The composition of these two transformations gives the actual transformation of p and q after the two diffeomorphisms ψ and ϕ. Notice that a bounding contour element (i.e., a point with coordinate $p = 0$) is mapped to a bounding contour element, as expected; $q = z^{0\prime}$ is then taken to $\hat{q} = 0$ after both transformations,

again as expected. Both transformations of p and q are independent of z, so the total transformation is as well. Also, when $p = 0$, the transformation of q to $\hat{q}$ is independent of the x coordinate; this means that along the projection fibre at the bounding contour, if $\hat{p} = 0$ the transformation of $\hat{q}$ back to q depends only on $y = \hat{y}$.

The reflectance function in these new coordinate is $\hat{R} = R \circ \hat{\Phi}^{-1}$, where $\hat{\Phi} = \hat{\phi} \circ \hat{\psi}$ is the total transformation of the coordinates (x, y, z, p, q) to $(\hat{x}, \hat{y}, \hat{z}, \hat{p}, \hat{q})$ induced by ψ and ϕ. The fact that R is space invariant and that $\hat{\Phi}$ is z invariant acting on p and q means that $\hat{R}$ is z invariant, but does depend on x and y in general. However, at a point $(\lambda, \hat{y}, 0, 0, \hat{q})$ in $\mathcal{C}(\mathbb{R}^3, 2)$ projecting to a bounding contour element $(0, \hat{y}, 0, 0, \hat{q})$, the observation at the end of the last paragraph means that we will have $\hat{R}(0, \hat{y}, 0, 0, \hat{q}) = \hat{R}(\lambda, \hat{y}, 0, 0, \hat{q})$, so that

$$\hat{R}_{\hat{y}}(0, \hat{y}, 0, 0, \hat{q}) = \hat{R}_{\hat{y}}(\lambda, \hat{y}, 0, 0, \hat{q})$$

and

$$0 = \hat{R}_{\hat{x}}(0, \hat{y}, 0, 0, \hat{q}) = \hat{R}_{\hat{x}}(\lambda, \hat{y}, 0, 0, \hat{q}).$$

We will need this a little further on.

In this new coordinate system (we will omit the " $\hat{}$ "), the surface is given as $z = \frac{1}{2}x^2$. This means that $p(x, y) = x$, $q(x, y) = 0$ for this surface, and the image of the surface will be $E(y, z) = R(x, y, \frac{1}{2}x^2, x, 0)$ where now $x = \sqrt{2z}$ is a function of the image coordinates. We can write

$$E(y, z) = R(\sqrt{2z}, y, z, \sqrt{2z}, 0)$$

for the image, and so

$$\begin{aligned} E_y &= \tilde{R}_y \\ E_z &= \frac{\tilde{R}_x + \tilde{R}_p}{\sqrt{2z}} + \tilde{R}_z, \end{aligned}$$

where we write $\tilde{R}$ to indicate the evaluation is at $(\sqrt{2z}, y, z, \sqrt{2z}, 0)$. After substituting into the calculation of the characteristic vector field in our turned coordinates at

the end of Section 6.1, the characteristic vector field for this image and reflectance function is

$$X = \begin{bmatrix} R_p \\ R_q \\ pR_p + qR_q \\ -R_x + p\left(\frac{\tilde{R}_x + \tilde{R}_p}{\sqrt{2z}} + \tilde{R}_z - R_z\right) \\ \tilde{R}_y - R_y + q\left(\frac{\tilde{R}_x + \tilde{R}_p}{\sqrt{2z}} + \tilde{R}_z - R_z\right) \end{bmatrix}.$$

We now proceed in the same spirit as in the motivating example: we define a new coordinate function $v = \sqrt{2z}$, notice that $v\dot{v} = \dot{z}$, replace z by $\frac{1}{2}v^2$, and multiply through by v; we also note that $\tilde{R}_z = R_z = 0$ by space-invariance as discussed above:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{v} \\ \dot{p} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} R_p v \\ R_q v \\ pR_p + qR_q \\ -R_x v + p(\tilde{R}_x + \tilde{R}_p) \\ v(\tilde{R}_y - R_y) + q(\tilde{R}_x + \tilde{R}_p) \end{bmatrix},$$

where the old functions R_i, $i = x, y$, etc. are now evaluated at $(x, y, \frac{1}{2}v^2, p, q)$ and similarly for $\tilde{R}_i$. This new vector field is Lipschitz at all potential bounding contour points $(x, y, 0, 0, 0)$, and clearly has a zero there as well (note that the vector field is really only defined on a half space $v \geq 0$.)

We can attempt to analyze this bounding contour critical point using the linearization methods used on "good" critical points in the image interior: the linearized version of the vector field looks almost like the vector field itself at the bounding contour point because many terms are of the form kR_i, where k is one of the coordinates v, p, q, all of which are zero at the bounding contour; hence, the linear part of such a term is just kR_i. We can write the linearization at the bounding contour as a matrix as we did in the interior critical point analysis:

$$X' = \begin{bmatrix} 0 & 0 & R_p & 0 & 0 \\ 0 & 0 & R_q & 0 & 0 \\ 0 & 0 & 0 & R_p & R_q \\ 0 & 0 & -R_x & \tilde{R}_x + \tilde{R}_p & 0 \\ 0 & 0 & \tilde{R}_y - R_y & 0 & \tilde{R}_x + \tilde{R}_p \end{bmatrix},$$

A point projecting to the bounding contour is of the form $(x, y, 0, 0, 0)$, so $\tilde{R}$ indicates evaluation at $(0, y, 0, 0, 0)$. From the analysis two pages back we have $\tilde{R}_y = R_y$ and $\tilde{R}_x = R_x = 0$ along the projection fibre through the origin.[1]

Thus, our linearized vector field is given by

$$X' = \begin{bmatrix} 0 & 0 & R_p & 0 & 0 \\ 0 & 0 & R_q & 0 & 0 \\ 0 & 0 & 0 & R_p & R_q \\ 0 & 0 & 0 & \tilde{R}_p & 0 \\ 0 & 0 & 0 & 0 & \tilde{R}_p \end{bmatrix}.$$

For convenience we will work at the origin, so that $\tilde{R}_p = R_p$. We can see by inspection that the characteristic polynomial of this matrix is $\lambda^3(\lambda - \tilde{R}_p)^2$. We can look for eigenvectors associated with the eigenvalues: for $\lambda = 0$, we see that the matrix has rank 3, and so the 0 eigenspace is defective, having only two eigenvectors in it, $(1, 0, 0, 0, 0)^T$ and $(0, 1, 0, 0, 0)^T$. For $\lambda = R_p$, we get two eigenvectors for X' from looking at the null space of

$$\begin{bmatrix} -R_p & 0 & R_p & 0 & 0 \\ 0 & -R_p & R_q & 0 & 0 \\ 0 & 0 & -R_p & R_p & R_q \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}:$$

we get $(1, R_q/R_p, 1, 1, 0)^T$ and $(1, R_q/R_p, 1, 0, R_p/R_q)^T$.

Because the matrix is defective, i.e. its eigenvectors do not form a basis, we need to expand our earlier analysis of the allowable two dimensional invariant subspaces.

---

[1] We can also argue directly that $\tilde{R}_x = R_x = 0$ along the projection fibre through the origin: we know that in the original space invariant problem from which this one comes, there are multiple solutions: we just move the correct solution surface back and forth in depth. After transforming all of these solution surfaces (or at least patches of them near the origin), these will again be a stack of solution patches along the projection direction. Consider the question of finding a possible bounding contour curve in space: we seek a curve $(x^0(s), s, 0, 0, 0)$ in $\mathcal{C}(\mathbb{R}^3, 2)$ that is consistent with the image. We must have $E(s, 0) = R(x^0(s), s, 0, 0, 0)$ along the bounding contour. If R_x at any point on the origin's projection fibre is not zero, then by the implicit function theorem we can solve for x as a function of s in the equation $E(s, 0) = R(x, s, 0, 0, 0)$ at least in some neighborhood of the point. This would imply a unique curve $x^0(s)$ as the bounding contour solution; in fact, we have a dense stack of consistent solution surface patches at different depths along the origin's projection fibre; thus there cannot be a unique local bounding contour solution, and so we must have $R_x = 0$ along the origin's (or any bounding contour point's) projection fibre.

In the case of a critical point on the interior of the image discussed in Chapter 5, we were able to focus on a diagonalizable matrix, for which all two-dimensional invariant subspaces are spanned by pairs of eigenvectors. Pairs of eigenvectors still span invariant spaces in the defective case, but there may be others. To begin this analysis, we put the matrix into canonical form by using the basis

$$f_1 = (0,0,1,0,0)^T$$
$$f_2 = X'f_1 = (R_p, R_q, 0, 0, 0)^T$$
$$f_3 = (0,1,0,0,0)^T$$
$$f_4 = (1, R_q/R_p, 1, 1, 0)^T$$
$$f_5 = (1, R_q/R_p, 1, 0, R_p/R_q)^T.$$

In this basis, the matrix becomes

$$X' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & R_p & 0 \\ 0 & 0 & 0 & 0 & R_p \end{bmatrix}.$$

X' has the characteristic polynomial $x^3(x - R_p)^2$. As this matrix is not diagonalizable, the minimal polynomial (the polynomial of least degree that annhilates X') cannot be just the product of the linear factors; by calculation it can be seen that the minimal polynomial is $x^2(x - R_p)$. If W is an invariant subspace of a linear transformation T on a finite vector space, then $T|_W$, the restriction of T to W, has a minimal polynomial that must divide the minimal polynomial of T on the whole vector space (concepts from (Hoffman and Kunzie, 1971)). If W has dimension two, then the degree of the minimal polynomial of $T|_W$ must be two or less. We can classify the invariant two-dimensional subspaces of T by their minimal polynomial, and hence by the factors of degree two or less of the minimal polynomial of $T = X'$.

Consider the minimal polynomial x. If a two dimensional invariant subspace is to have this as the minimal polynomial for $T|_W$, we must have $T(W) = 0$. Looking at

the canonical matrix for $T = X'$, there is only a two dimensional null space for T, and so the only invariant two-dimensional subspace with x as the minimal polynomial is $\text{Span}(f_2, f_3)$. This is the zero eigenvalue eigenspace, W^0.

Consider the minimal polynomial $(x - R_p)$. We have

$$T - R_p I = \begin{bmatrix} -R_p & 0 & 0 & 0 & 0 \\ 1 & -R_p & 0 & 0 & 0 \\ 0 & 0 & -R_p & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and again this matrix only has a null space of dimension two, so the only invariant two-dimensional subspace with $x - R_p$ as minimal polynomial is $\text{Span}(f_4, f_5)$. This is the non-zero eigenvalue eigenspace, W^R.

Consider now the minimal polynomial x^2. We seek invariant subspaces such that $T(W) \neq 0$, but $T^2(W) = 0$. We have

$$T^2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & R_p^2 & 0 \\ 0 & 0 & 0 & 0 & R_p^2 \end{bmatrix}.$$

In order for the above conditions to be true, there must be some vector $\mathbf{v} \in W$ such that $T\mathbf{v} \neq 0$, $T^2\mathbf{v} = 0$. If $\mathbf{v} = (v_1, v_2, v_3, v_4, v_5)^T$, then the first condition implies at least one of v_1, v_4, v_5 must not be 0; the second condition implies in general (assuming we are at a generic point for R so that R_p and R_q are non-zero) that $v_4 = v_5 = 0$. Thus, we must have $f_1 \in W$. Since W is invariant, we must also have $Tf_1 = f_2 \in W$ as well, and since these are independent, the only T invariant two dimensional subspace with minimal polynomial x^2 is $\text{Span}(f_1, f_2)$. As we shall see later, this is not an allowable invariant subspace: it violates an integrability constraint.

For the last potential minimal polynomial $x(x - R_p)$, any pair of eigenvectors, one from the zero eigenvalue eigenspace W^0 and one from the non-zero eigenvalue

eigenspace W^R, spans a subspace with this as its minimal polynomial. Say $u_1 \in W^0, u_2 \in W^R$. Then $(T - R_p)u_1 = -R_p u_1 \neq 0$, $Tu_2 = R_p u_2 \neq 0$, but $T(T - R_p)u_i = 0$, so $W = \text{Span}(u_1, u_2)$ has minimal polynomial $x(x - R_p)$. Is there any other vector $\mathbf{v} = v_1 f_1 + u_1 + u_2$, $u_1 \in W^0$, $u_2 \in W^R$, that could possibly be in an invariant subspace with $x(x - R_p)$ as the minimal polynomial for $T|_W$?

$$\begin{aligned}(T - R_p)(\mathbf{v}) &= v_1 f_2 + R_p u_2 - R_p v_1 f_1 - R_p u_1 - R_p u_2 \\ &= v_1(f_2 - R_p f_1) - R_p u_1 \\ T(T - R_p)\mathbf{v} &= -v_1 R_p f_2 = 0,\end{aligned}$$

so we must have $v_1 = 0$. Thus, the only invariant subspaces with $x(x - R_p)$ as minimal polynomial are those spanned by a pair of eigenvectors, one from W^0 and one from W^R.

This exhausts the possibilities for two-dimensional invariant subspaces of this linear transformation. The subspace W^R can be rejected as deriving from a surface in space because its space projection is one-dimensional. The zero eigenvalue eigenspace W^0 is more of a problem: it expresses the potential complication due to higher order solution surfaces, i.e., surfaces that are third and higher order flat along the bounding contour. In this analysis we concentrate on truly second order solution surfaces if for no other reason than that these are generically the ones we expect to have images of; we will not consider W^0 a possible invariant tangent space for a solution surface. The unusual invariant subspace spanned by f_1 and f_2 will be disallowed when we introduce an integrability 2-tensor constraint which needs to hold at the bounding contour. As we shall see, this will leave only subspaces formed from a choice of eigenvector from W^0 and a choice of eigenvector from W^R as candidates for solution surfaces' tangent spaces.

As suggested above, there is one more constraint that needs to be taken into account. We recall that the linearization is in (x, y, v, p, q) coordinates: we have

replaced z by v. Consider the integrability constraint as expressed by the contact 1-form θ written with v instead of z:

$$\begin{aligned} 0 = \theta(\mathbf{h}) &= (\mathrm{d}z - p\mathrm{d}x - q\mathrm{d}y)(\mathbf{h}) \\ &= (v\mathrm{d}v - p\mathrm{d}x - q\mathrm{d}y)(\mathbf{h}), \end{aligned}$$

where $\mathbf{h}$ is any tangent vector to an allowable solution surface: a two-dimensional surface in $\mathcal{C}(\mathbb{R}^3, 2)$ whose orientation part as specified by the p and q coordinates is consistent with the space part. Written with the v coordinate, we see that this condition is identically satisfied at the bounding contour: any surface (now in the x, y, v, p, q topology) containing the bounding contour will satisfy this condition at the bounding contour, since $v = p = q = 0$ there. This is another reflection of the degeneracy of the bounding contour.

As in the motivating example, another way to view this is to consider a path $(x, y, v, p, q)(s)$ in a possible solution surface. The integrability condition applied to the tangent of this curve gives

$$vv' = px' + qy',$$

and again this provides no constraint on the curve at the bounding contour. However, we can consider the limit of this constraint as the bounding contour is approached by looking at the derivative of the constraint with respect to s:

$$(v')^2 + vv'' = p'x' + px'' + q'y' + qy'';$$

at the bounding contour where $v = p = q = 0$, this becomes

$$(v')^2 = p'x' + q'y'.$$

In a sense, the degeneracy of the integrability condition at the bounding contour allows the next highest order of the integrability condition to appear as the major constraint. This integrability condition must be true for all tangent vectors lying in a potential integrable solution surface.

We can write this condition as a bilinear form or 2-tensor on the tangent space at the bounding contour:

$$(\mathrm{d}p \otimes \mathrm{d}x + \mathrm{d}q \otimes \mathrm{d}y - \mathrm{d}v \otimes \mathrm{d}v)(\mathbf{h}, \mathbf{h}) = 0$$

or as

$$\mathbf{h}^t \mathbf{A} \mathbf{h} = 0,$$

where $\mathbf{A}$ can be thought of as the symmetric matrix representation of this 2-tensor in the coordinate system (x, y, v, p, q),

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Which of our invariant subspaces will satisfy this 2-tensor integrability condition? $f_1 = (0, 0, 1, 0, 0)$ fails the condition, and so the invariant subspace $\mathrm{Span}(f_1, f_2)$ does not correspond to a real surface. Applying the 2-tensor to the basis vectors $f_2 = (R_p, R_q, 0, 0, 0)^T$, $f_3 = (0, 1, 0, 0, 0)^T$, we have

$$f_2^T \mathbf{A} f_2 = 0 \cdot R_p + 0 \cdot R_q - 0 \cdot 0 = 0$$
$$f_3^T \mathbf{A} f_3 = 0 \cdot 0 + 0 \cdot 1 - 0 \cdot 0 = 0,$$

so both these basis vectors satisfy the 2-tensor integrability constraint. Applying the 2-tensor to the basis vectors $f_4 = (1, \beta, 1, 1, 0)^T$, $f_5 = (1, \beta, 1, 0, 1/\beta)^T$ where we define $\beta = R_p / R_q$, we have

$$f_4^T \mathbf{A} f_4 = 1 \cdot 1 + 0 \cdot \beta - 1 \cdot 1 = 0$$
$$f_5^T \mathbf{A} f_5 = 0 \cdot 1 + (1/\beta) \cdot \beta - 1 \cdot 1 = 0,$$

so both of these basis vectors satisfy the 2-tensor integrability condition as well.

This leads us to ask what happens to the 2-tensor integrability constraint acting on linear combinations of vectors. Assume $\mathbf{h}_1$ and $\mathbf{h}_2$ are two vectors satisfying

the 2-tensor constraint $0 = \mathbf{h}_i^T \mathbf{A} \mathbf{h}_i$. Any scalar multiple $\lambda \mathbf{h}_1$ will also satisfy the constraint. Also,

$$\begin{aligned}(\mathbf{h}_1 + \mathbf{h}_2)^T \mathbf{A} (\mathbf{h}_1 + \mathbf{h}_2) &= \mathbf{h}_1^T \mathbf{A} \mathbf{h}_1 + \mathbf{h}_2^T \mathbf{A} \mathbf{h}_2 + 2\mathbf{h}_1^T \mathbf{A} \mathbf{h}_2 \\ &= 2\mathbf{h}_1^T \mathbf{A} \mathbf{h}_2,\end{aligned}$$

so that a linear combination of $\mathbf{h}_1$ and $\mathbf{h}_2$ satisfies the 2-tensor integrability constraint if and only if $\mathbf{h}_1$ and $\mathbf{h}_2$ are $\mathbf{A}$–orthogonal, i.e. $\mathbf{h}_1^T \mathbf{A} \mathbf{h}_2 = 0$. If $\mathbf{h}_1$ and $\mathbf{h}_2$ span a two-dimensional space all of whose vectors satisfy the 2-tensor integrability constraint, we must have $\mathbf{h}_1$ and $\mathbf{h}_2$ $\mathbf{A}$–orthogonal to each other. Correspondingly, given $\mathbf{h}_1$, we can only choose a possible $\mathbf{h}_2$ from the set of $\mathbf{A}$–perpendicular vectors to it, $\mathbf{h}_1^{(\mathbf{A}\perp)}$. Since $\mathbf{A}$ has full rank, $\mathbf{h}_1^{(\mathbf{A}\perp)}$ will have dimension four.

Looking at the zero eigenvalue eigenspace W^0, we have

$$f_2^T \mathbf{A} f_3 = 0 \cdot 0 + \mathbb{R}_p \cdot 0 + 0 \cdot 1 + R_q \cdot 0 - 2(0 \cdot 0) = 0,$$

so the subspace spanned by f_2 and f_3 satisfies the integrability constraint. For the non-zero eigenvalue eigenspace W^R,

$$f_4^T \mathbf{A} f_5 = 1 \cdot 1 + 1 \cdot 0 + 0 \cdot \beta + (1/\beta) \cdot \beta - 2(1 \cdot 1) = 0,$$

so the subspace spanned by e_4 and e_5 also satisfies the integrability constraint. Thus, both eigenspaces W^0 and W^R satisfy the integrability constraint.

Consider now the mixed invariant subspace case, with one basis vector, $\mathbf{h}_1$ chosen from W^0 and another to be chosen from W^R to make the resulting subspace obey the integrability constraint. Given $\mathbf{h}_1 \in W^0$, we seek an eigenvector from the intersection of the non-zero eigenvalue eigenspace W^R with $\mathbf{h}_1^{(\mathbf{A}\perp)}$. Since $\mathcal{C}(\mathbb{R}^3, 2)$ only has dimension five, the minimum dimension of the intersection of a two dimensional subspace and a four dimensional subspace is one. The intersection has dimension two if and only if the entire non-zero eigenvalue eigenspace W^R is $\mathbf{A}$ perpendicular to $\mathbf{h}_1$. This does not happen in general.

To see this, begin with a vector $\mathbf{h}_1$ from the zero eigenvalue eigenspace W^0, $\mathbf{h}_1 = (a, b, 0, 0, 0)^T$. We seek an eigenvector $\mathbf{h}_2$ of the non-zero eigenvalue eigenspace W^R. From the coordinates of the basis vectors $(1, \beta, 1, 1, 0)$ and $(1, \beta, 1, 0, 1/\beta)$ for W^R, $\mathbf{h}_2$ has the form $\mathbf{h}_2 = (c+d, \beta(c+d), c+d, c, d/\beta)^T$ where c and d are constants, and $\beta = R_p/R_q$ as before. We want $\mathbf{h}_2$ and $\mathbf{h}_1$ to violate the integrability constraint; i.e., we want

$$\begin{aligned} 0 &\neq \mathbf{h}_1^T \mathbf{A} \mathbf{h}_2 \\ &\neq 0 \cdot (c+d) + (c+d) \cdot a + 0 \cdot \beta(c+d) + (d/\beta) \cdot b - 2(0 \cdot (c+d)) \\ -db &\neq \beta(c+d)a. \end{aligned}$$

We can clearly choose c and d not both zero so that this latter condition is satisfied when not both a and b are zero. Thus, the non-zero eigenvalue eigenspace W^R is not completely contained in $\mathbf{h}_1^{(\mathbf{A}\perp)}$ for any non-zero $\mathbf{h}_1$, and the intersection $W^R \cap \mathbf{h}_1^{(\mathbf{A}\perp)}$ has dimension one: for each choice of $\mathbf{h}_1 \in W^0$ there will be only a one-dimensional subspace of possible $\mathbf{h}_2 \in W^R$ that makes the subspace spanned by $\mathbf{h}_1$ and $\mathbf{h}_2$ satisfy the integrability constraint.

Note that the space parts of tangent spaces spanned by one vector from W^0 and one from W^R are the same; this reflects the fact that the tangent plane in $\mathbb{R}^3$ is fixed by the surface normal which is known at the bounding contour. The difference between different eigenvector pairs is in the curvature information contained in the orientation part of the eigenvectors.

For surfaces with non-zero second order behavior (i.e. surfaces that have at most one tangent vector from W^0), this suggests there is always a one parameter family of tangent spaces consistent with the image data at the bounding contour point. This is consistent with the results of the motivating example, which indicated solution surfaces were determined only up to a choice of depth values along the bounding contour. This is similar to the case of an image patch in the interior without critical points: by picking different depth curves $x^0(s)$ along a fixed curve in the image $(y(s), z(s))$, we can generate a one-parameter family of tangent planes consistent

with the image at that point: this family is determined by the constant brightness contour of the reflectance map with brightness value equal to that at the image point. This is not similar to the case with image patches containing a critical point: as discussed in Chapter 5, the very good critical points almost always determine a small number of possible solution surfaces consistent with the patch, and therefore there are only a finite number of tangent planes consistent with a given point in the image patch.

How does the original surface, $z = \frac{1}{2}x^2$, fit into all of this? This surface is parameterized in (x, y, v, p, q) space as

$$j : (r, s) \longmapsto \begin{pmatrix} r \\ s \\ r \\ r \\ 0 \end{pmatrix},$$

using $v = \sqrt{2z}$. The invariant tangent space in $\mathcal{C}(\mathbb{R}^3, 2)$ corresponding to this surface is the subspace spanned by the vectors $\mathrm{d}j/\mathrm{d}r|_{r=0,s=0} = (1,0,1,1,0)^T$ and $\mathrm{d}j/\mathrm{d}s|_{r=0,s=0} = (0,1,0,0,0)^T$. This subspace is the same as the invariant subspace spanned by the eigenvectors $(1, R_p/R_q, 1, 1, 0)$ and $(0,1,0,0,0)$ of X', so our original surface has a tangent plane spanned by eigenvectors of X', one from W^0 and one from W^R, consistent with our analysis.

The presence of the zero-eigenvalue eigenspaces for X' is important in allowing our hypothesis to be possible. As we discussed last chapter, the Grobman–Hartman theorem says that a critical point which has no eigenvalues with zero real part has the non-linearized local picture diffeomorphic to the linearized picture, as in the very good image critical point case: an invariant manifold in the linearized picture corresponds uniquely to a local invariant manifold of the non-linearized picture. We are positing more flexibility than that: we are saying that we may have the linearization $x^{0\prime}(0)$ the same for two different depth curves $x^0(s)$ through the same point $(x^0(0), 0, 0)$ and still get different solution surfaces in the general case, matching the

motivating example, but contradicting the result of the Grobman–Hartman theorem. The Grobman–Hartman theorem does not apply here because of the presence of zero-eigenvalue eigenspaces.

Note that we have not proved this assertion about solution surfaces near the bounding contour; we are saying that the assertion is consistent with the degeneracy of the critical point at the bounding contour. Unfortunately, invariant manifolds of a linearized system with zero eigenvalues may disappear once the second-order terms are put back in. One can construct simple examples (e.g. the system $\dot{x} = x^2, \dot{y} = y$) where the linearized picture does not completely capture the actual behavior or the uniqueness of invariant manifolds through a point. There are some results on existence of invariant surfaces in the presence of non-hyperbolic critical points, but they do not seem useful to our problem.

Nonetheless, the motivating example suggests that the above interpretation of the linearized picture may be correct in general. We will pursue the question in a different way using the technique of power series analysis: here too one is hampered by the restriction that the power series analysis only tells about analytic solution surfaces, and unfortunately when zero-eigenvalue eigenspaces of a critical point are present, non-analytic behavior is not uncommon. In addition, a typical difficulty is proving convergence of a derived power series: we have not done this in our case. Nonetheless, the power series analysis in the next section shows the one-parameter ambiguity extending into the third order behavior of analytic solution surfaces, consistent with the hypothesis that generic solution surfaces at the bounding contour are determined only up to a smooth choice of depths along the bounding contour.

6.4 General Case: Power Series Analysis

We begin again with our transformed problem: a surface $z = \frac{1}{2}x^2$ with a space-varying reflectance function. We seek other surfaces $z = f(x, y)$ that could have given rise to the image. We could begin from a general surface and a space-invariant

reflectance function, but the additional complexity of the surface structure makes the calculations even longer than they are here.

We can begin by analyzing the restrictions placed on the first few orders of behavior by the existence of the image bounding contour $(0, s, 0)$; we assume (as before) that this contour in the image is due to a curve in space with $\mathcal{C}(\mathbf{IR}^3, 2)$ coordinates $(x^0(s), s, 0, 0, 0)$ that is contained in the surface. We will assume we are looking at behavior in a neighborhood of the origin.

We begin by expanding both f and x^0:

$$\begin{aligned} f(x,y) &= b_1 x + b_2 y + a_1 x^2 + 2a_2 xy + a_3 y^2 \\ &\quad + a_4 x^3 + 3a_5 x^2 y + 3a_6 xy^2 + a_7 y^3 + \cdots \\ x^0(y) &= \alpha_1 y + \alpha_2 y^2 + \alpha_3 y^3 + \cdots. \end{aligned}$$

Assuming both $f(x,y)$ and $x^0(x,y)$ are analytic, then we have

$$\begin{aligned} f(x^0(y), y) &= 0 \\ f_x(x^0(y), y) &= 0 \\ f_y(x^0(y), y) &= 0 \end{aligned}$$

since $(x^0(y), y, 0)$ lies on the surface and is the bounding contour for the surface. By thinking of f as a function of x only with y as a parameter, these conditions mean that we can write

$$f(x,y) = (x - x^0(y))^2 Q(x,y),$$

where Q is also analytic, since for fixed y the power series for f in x around $x = x^0(y)$ has a zero and a zero in the derivative at $x = x^0(y)$. If we write

$$Q(x,y) = q_0 + q_1 x + q_2 y + q_3 x^2 + 2q_4 xy + q_5 y^2 + \cdots,$$

then we can take the product $(x - x^0(y))^2 Q(x,y)$ term by term as a power series: we get

$$\begin{aligned} f(x,y) &= q_0 x^2 - 2\alpha_1 q_0 xy + \alpha_1^2 q_0 y^2 + q_1 x^3 + (q_2 - 2\alpha_1 q_1) x^2 y \\ &\quad + (\alpha_1^2 q_1 - 2\alpha_2 q_0 - 2\alpha_1 q_2) xy^2 + (2\alpha_1 \alpha_2 q_0 + \alpha_1^2 q_2) y^3 + \cdots \end{aligned}$$

Taking the derivatives with respect to x and y we get

$$\begin{aligned} f_x(x,y) &= 2q_0x - 2\alpha_1 q_0 y + 3q_1x^2 + 2(q_2 - 2\alpha_1 q_1)xy \\ &\quad + (\alpha_1^2 q_1 - 2\alpha_2 q_0 - 2\alpha_1 q_2)y^2 + \cdots \\ f_y(x,y) &= -2\alpha_1 q_0 x + 2\alpha_1^2 q_0 y + (q_2 - 2\alpha_1 q_1)x^2 \\ &\quad + 2(\alpha_1^2 q_1 - 2\alpha_2 q_0 - 2\alpha_1 q_2)xy + 3(2\alpha_1\alpha_2 q_0 + \alpha_1^2 q_2)y^2 + \cdots. \end{aligned}$$

The decomposition $f(x,y) = (x - x^0(y)^2 Q(x,y)$ takes care of the geometric information available to us: that $(x^0(y), y, 0)$ is the bounding contour for the surface defined parametrically as $(x, y, f(x,y))$. Now let us add in the image data. We know the image is due to the surface $z = \frac{1}{2}x^2$, so, as before, we have

$$\begin{aligned} p(y,z) &= x(y,z) = \sqrt{2z} \\ q(y,z) &= 0 \\ E(y,z) &= R(\sqrt{2z}, y, z, \sqrt{2z}, 0) \end{aligned}$$

Replacing $z = f(x,y)$ to find the new surface, the image irradiance equation can be written as

$$\begin{aligned} E(y, f(x,y)) &= R(x, y, f(x,y), f_x(x,y), f_y(x,y)) \\ R(\sqrt{2f(x,y)}, y, f(x,y), \sqrt{2f(x,y)}, 0) &= R(x, y, f(x,y), f_x(x,y), f_y(x,y)). \quad (*) \end{aligned}$$

The next step is to expand this equation in powers of x and y.[2]

We need an expansion for $\sqrt{2f(x,y)}$. After gathering terms, we have

$$\begin{aligned} \sqrt{2f(x,y)} &= |x - x^0(y)|\sqrt{2Q(x,y)} \\ &= \sqrt{2q_0}x - \sqrt{2q_0}\alpha_1 y + (1/\sqrt{2q_0})q_1x^2 + (1/\sqrt{2q_0})(q_2 - q_1\alpha_1)xy \\ &\quad - \sqrt{2q_0}(\alpha_2 + (\alpha_1 q_2/2q_0))y^2 + \cdots, \end{aligned}$$

[2] If we had worked with a general initial surface $z = g(x,y)$ instead of $z = \frac{1}{2}x^2$, we would have had a fair bit more trouble writing this equation—effectively, we would be simultaneously solving for $x^g(y,z)$, where for a given y and z denoting a point in the image, $x^g(y,z)$ is the value of x such that $z = g(x^g, y) = f(x,y)$.

using the Taylor series expansion $\sqrt{a+h} = \sqrt{a}(1+(1/2)(h/a)+\cdots$. We will assume that $q_0 \neq 0$ so that $f(x,y)$ does have second order behavior at the bounding contour. As we are interested in the positive x sheet, we have also taken $(x - x^0(y)) > 0$, as this defines (locally) the region of the surface $f(x,y)$ which is more x positive. As expected, $\sqrt{f(x,y)}$ on this sheet is first order in x and y; this suggests some care in expanding the image irradiance equation $(*)$ in powers of x and y to get the coefficients correctly matched.

For convenience, we will assume that R has an expansion of the form[3]

$$R(x,y,z,p,q) = R_0 + R_x x + R_y y + \cdots + R_q q + R_{xx}x^2 + R_{xy}xy + \cdots + R_{qq}q^2 + \cdots.$$

After substituting into the image irradiance equation the series for $\sqrt{2f(x,y)}$, $f(x,y)$, $f_x(x,y)$, and $f_y(x,y)$, we can examine the terms that are linear in x and y:

$$\begin{aligned}(R_p + R_x)(\sqrt{2q_0}x - \sqrt{2q_0}\alpha_1 y) + R_y y \\ = R_x x + R_y y + R_p(2q_0 x - 2\alpha_1 q_0 y) + R_q(-2\alpha_1 q_0 x + 2\alpha_1^2 q_0 y)\end{aligned}$$

Equating coefficients of x and y we have

$$\begin{aligned}(R_x + R_p)\sqrt{2q_0} &= R_x + 2(R_p q_0 - R_q \alpha_1 q_0) \\ -(R_x + R_p)\sqrt{2q_0}\alpha_1 + R_y &= R_y - 2(R_p \alpha_1 q_0 - R_q \alpha_1^2 q_0).\end{aligned}$$

At first glance, this appears to yield two equations in the two unknowns q_0 and α_1; however, after canceling the R_y terms in the second equation, noting that $R_x = 0$, and multiplying the first by $-\alpha_1$ the second equation is seen to be identical to the first. We can use the first equation to determine α_1 as a function of q_0, but we see the one-parameter second order ambiguity in f arising early in the analysis. We have

$$\alpha_1 = \frac{R_p}{R_q}\left(1 - \frac{1}{\sqrt{2q_0}}\right).$$

[3] For bookkeeping conveniance the coefficients $R_{xx}, R_{xy}, \ldots, R_{qq}$ are here just coefficients for the terms of the expansion. $R_{xx} = (1/2)\partial^2/\partial x^2 R$, etc. relates these coefficients to actual derivatives of R.

[Note that we are assuming in general that R_p and R_q are non-zero.]

The second order terms of the image irradiance equation are more of a book-keeping challenge. Rather than write out all the second order terms at once, we will focus on the coefficients for each term x^2, xy, y^2 individually.

Coefficients for x^2:

$$\begin{aligned}(R_{xx} + R_{xp} + R_{pp})2q_0 + R_z q_0 + \frac{R_p}{\sqrt{2q_0}}q_1 \\ = R_{xx} + 2R_{xp}q_0 - 2R_{xq}\alpha_1 q_0 + R_z q_0 + 3R_p q_1 + 4R_{pp}q_0^2 - 4R_{pq}\alpha_1 q_0^2 \\ + R_q(q_2 - 2\alpha_1 q_1) + 4R_{qq}\alpha_1^2 q_0^2\end{aligned}$$

Coefficients for xy:

$$\begin{aligned}-4(R_{xx} + R_{xp} + R_{pp})q_0\alpha_1 + (R_{xy} + R_{yp})\sqrt{2q_0} + \frac{R_p}{\sqrt{2q_0}}(q_2 - q_1\alpha_1) - 2R_z\alpha_1 q_0 \\ = R_{xy} - 2R_{xp}\alpha_1 q_0 + 2R_{xq}\alpha_1^2 q_0 + 2R_{yp}q_0 - 2R_{yq}\alpha_1 q_0 \\ - 2R_z\alpha_1 q_0 + 2R_p(q_2 - 2\alpha_1 q_1) - 8R_{pp}\alpha_1 q_0^2 + 8R_{pq}\alpha_1^2 q_0^2 \\ + 2R_q(\alpha_1^2 q_1 - 2\alpha_2 q_0 - 2\alpha_1 q_2) - 8R_{qq}\alpha_1^3 q_0^2\end{aligned}$$

Coefficients for y^2:

$$\begin{aligned}2(R_{xx} + R_{xp} + R_{pp})q_0\alpha_1^2 - (R_{xy} + R_{yp})\sqrt{2q_0}\alpha_1 + R_{yy} \\ + R_z\alpha_1^2 q_0 - R_p\sqrt{2q_0}\left(\alpha_2 + \frac{\alpha_1 q_2}{2q0}\right) \\ = R_{yy} - 2R_{yp}\alpha_1 q_0 + 2R_{yq}\alpha_1^2 q_0 + R_z\alpha_1^2 q_0 \\ + R_p(\alpha_1^2 q_1 - 2\alpha_2 q_0 - 2\alpha_1 q_2) + 4R_{pp}\alpha_1^2 q_0^2 \\ - 4R_{pq}\alpha_1^3 q_0^2 + 3R_q(2\alpha_1\alpha_2 q_0 + \alpha_1^2 q_2) + 4R_{qq}\alpha_1^4 q_0^2\end{aligned}$$

These are also a system of linear equations in q_1, q_2, and α_2. These last three equations together define a linear system of three equations in three unknowns. The equations for this system are

$$\left(\frac{R_p}{\sqrt{2q_0}} - 3R_p + 2\alpha_1 R_q\right) q_1 - R_q q_2 = c_1$$
$$\left(-\frac{R_p\alpha_1}{\sqrt{2q_0}} + 4R_p\alpha_1 - 2R_q\alpha_1^2\right) q_1 + \left(\frac{R_p}{\sqrt{2q_0}} - 2R_p + 4R_q\alpha_1\right) q_2 + 4R_q q_0 \alpha_2 = c_2$$
$$- R_p\alpha_1^2 q_1 + \left(-\frac{R_p\alpha_1}{\sqrt{2q_0}} + 2R_p\alpha_1 - 3R_q\alpha_1^2\right) q_2$$
$$+ \left(-R_p\sqrt{2q_0} + 2R_p q_0 - 6R_q\alpha_1 q_0\right)\alpha_2 = c_3$$

We can show that these three equations form a linear system of rank two, so that there is another one-parameter ambiguity arising at the second order. First, we recall that along the bounding contour, $R_x = 0$. Second, we make use of the first order solution for α_1 :

$$\alpha_1 = \frac{R_p}{R_q}\left(1 - \frac{1}{\sqrt{2q_0}}\right).$$

We can substitute this in to the equations and simplify. The system of three equations becomes

$$\left(-\frac{R_p}{\sqrt{2q_0}} - R_p\right) q_1 - R_q q_2 = c_1 \qquad (*)$$
$$\frac{R_p^2}{R_q}\left(1 - \frac{1}{\sqrt{2q_0}}\right)\left(2 + \frac{1}{\sqrt{2q_0}}\right) q_1 + R_p\left(2 - \frac{3}{\sqrt{2q_0}}\right) q_2 + 4R_q q_0 \alpha_2 = c_2 \qquad (**)$$
$$- \frac{R_p^3}{R_q^2}\left(1 - \frac{1}{\sqrt{2q_0}}\right)^2 q_1 + \frac{R_p^2}{R_q}\left(1 - \frac{1}{\sqrt{2q_0}}\right)\left(-1 + \frac{2}{\sqrt{2q_0}}\right) q_2$$
$$- 4R_p q_0\left(1 - \frac{1}{\sqrt{2q_0}}\right)\alpha_2 = c_3$$

If $q_0 = \frac{1}{2}$, then $1 - (1/\sqrt{2q_0}) = 0$, and the last equation is identically zero. In this case the three linear homogenous equations (i.e., setting the $c_i = 0$) form a dependent, rank 2 system. If $q_0 \neq \frac{1}{2}$, we can factor out $1 - (1/\sqrt{2q_0})$ from the last equation to get

$$- \frac{R_p^3}{R_q^2}\left(1 - \frac{1}{\sqrt{2q_0}}\right) q_1 + \frac{R_p^2}{R_q}\left(-1 + \frac{2}{\sqrt{2q_0}}\right) q_2 - 4R_p q_0 \alpha_2 = d_3 \qquad (***)$$

We can now show that a linear combination of the homogenous versions of equations (**) and (***) will give us a linear multiple of the homogenous version of equation (*), showing that the system is dependent in general. We take (R_q/R_p) times equation (***) and add it to equation (**) to get

$$\frac{R_p^2}{R_q}\left(1-\frac{1}{\sqrt{2q_0}}\right)\left(1+\frac{1}{\sqrt{2q_0}}\right)q_1 + R_p\left(1-\frac{1}{\sqrt{2q_0}}\right)q_2,$$

which is just $-(R_p/R_q)(1-(1/\sqrt{2q_0}))$ times the homogenous version of equation (*). In this case as well the system has rank 2.

In general the fact that the homogenous system has rank two means either that the system of linear equations including the original c_i has no solution, or it has a one-dimensional family of solutions. In the Appendix to this chapter we show that for this singular system written as $Ax = b$, we do have b in the row space of A; thus there is a one-dimensional family of solutions for x, i.e. for q_1, q_2, and α_2. As a result, the third derivatives of an analytic solution surface with non-zero second order behavior also possesses a one-parameter ambiguity, again consistent with the suggestion that every depth curve $x^0(s)$ generates a different solution curve.

6.5 Bounding Contour Conclusions

This analysis suggests (up to third order) that an analytic potential solution surface f with non-zero second order behavior is determined by the image data and the bounding contour in the image up to a choice of depth function $x^0(s)$ along the bounding contour. This corresponds with our linearized analysis and carries that analysis one order further; however, the problems discussed earlier about convergence of power series derived this way, and the potential existence of non-analytic solutions (or solutions with only third order behavior and above) prevents us from concluding with certainty that the general mathematical problem admits solutions unique up to a choice of $x^0(s)$. It is not clear how to extend the power series analysis beyond the third order without a prohibitive notational burden.

This analysis of the bounding contour again emphasizes the theoretical importance of the critical points on the interior of the image as the determinants of actual surface shape. Even if solution surfaces are determined up to a choice of $x^0(s)$, it still suggests that the bounding contour shading information in isolation is of limited value in determining shape. We would expect from this analysis that if subjects are shown bounding contours (not closing on themselves) with image data near them, subjects should be relatively poor at determining the true solution surface even if they know the position of the light source.

If true, our hypothesis says the ambiguity at the bounding contour is about as bad as the ambiguity in the interior of the image on a patch not including a critical point: there as well we can determine surface shape up to the choice of a depth curve by picking a curve on the image, choosing a depth curve along it, and generating a solution surface (locally) using the characteristics.

However, this does not mean the bounding contour image data is useless: the bounding contour may be very important for determining whether a particular reflectance function could have given rise to the image. The image values at the bounding contour (in so far as they can be reliably determined) are the result of known surface orientations; they provide a curve of data from the reflectance function which may quite tightly constrain choices within the class of possible reflectance functions. There may also be global effects of having an entire bounding contour available. For example, in the degenerate case of an image of a sphere lit from the viewing direction, the saddle shaped solutions valid near the critical point at the center of the image cannot be extended out to the bounding contour. More work needs to be done to understand such global effects in general.

Appendix to Chapter 6

A6.1 Linear Dependence of the Power Series Constants

If we have a linear equation $Ax = b$ and A is singular, the equation has solutions if and only if b is in the range of A. In discussing the bounding contour case, we wound up with such a linear equation: the matrix A consisted of the constants attached to the unknown power series coefficients q_1, q_2, α_2, while b is a complicated expression containing second derivatives of R and the coefficients q_0, α_1. It is our aim in this appendix to organize the pieces of b and show that it is indeed in the column range of A.

We will collect the various terms in b and factor out the different R_{ij} second derivatives for each row of b. This will make b the sum of a set of column vectors, each of which is multiplied by some R_{ij}. Note that $R_{xx} = R_{xy} = R_{xq} = 0$ at the bounding contour: we have from the chapter $R(\lambda, 0, y, 0, q) = R(0, 0, y, 0, q)$, so $R_x(\lambda, 0, y, 0, q) = 0$ after taking the derivative with respect to λ. Taking further derivatives with respect to λ, y, and q gives the result.

We will arrange the columns in two tables: the top row gives the R_{ij} to which the column is associated; the bottom three rows give the values of the column of entries of b that have R_{ij} as a factor.

R_{xx}	R_{xy}	R_{xp}	R_{xq}	R_{yy}	R_{yp}
$2q_0 - 1$	0	0	$2\alpha_1 q_0$	0	0
$-4q_0\alpha_1$	$\sqrt{2q_0} - 1$	$-4q_0\alpha_1 + 2\alpha_1 q_0$	$-2\alpha_1^2 q_0$	0	$\sqrt{2q_0} - 2q_0$
$2q_0\alpha_1^2$	$-\sqrt{2q_0}\alpha_1$	$2q_0\alpha_1^2$	0	0	$-\sqrt{2q_0}\alpha_1 + 2\alpha_1 q_0$

R_{yq}	R_{pp}	R_{pq}	R_{qq}
0	$2q_0 - 4q_0^2$	$4\alpha_1 q_0^2$	$-4\alpha_1^2 q_0^2$
$2\alpha_1 q_0$	$-4q_0\alpha_1 + 8\alpha_1 q_0^2$	$-8\alpha_1^2 q_0^2$	$8\alpha_1^3 q_0^2$
$-2\alpha_1^2 q_0$	$2q_0\alpha_1^2 - 4\alpha_1^2 q_0^2$	$4\alpha_1^3 q_0^2$	$-4\alpha_1^4 q_0^2$

We can see that the R_{pp}, R_{pq} and R_{qq} columns are all multiples of the vector $(1, -2\alpha_1, \alpha_1^2)^T$. The R_{yq}, R_{yp}, and R_{xp} are all multiples of the vector $(0, 1, -\alpha_1)^T$. Since R_{xx}, R_{xy}, and R_{xp} are all zero here, these are the only non-zero columns in whose span b lies.

Are the columns $(0, 1, -\alpha_1)^T$ and $(1, -2\alpha_1, \alpha_1^2)^T$ contained in the column range of the matrix A? A has the form

$$\begin{bmatrix} (-R_p\sqrt{2q_0} - R_p) & -R_q q_2 & 0 \\ \frac{R_p^2}{R_q}\left(1 - \frac{1}{\sqrt{2q_0}}\right)\left(2 + \frac{1}{\sqrt{2q_0}}\right) & R_p\left(2 - \frac{3}{\sqrt{2q_0}}\right) & 4R_q \\ -\frac{R_p^3}{R_q^2}\left(1 - \frac{1}{\sqrt{2q_0}}\right)^2 & \frac{R_p^2}{R_q}\left(1 - \frac{1}{\sqrt{2q_0}}\right)\left(-1 + \frac{2}{\sqrt{2q_0}}\right) & -4R_p q_0\left(1 - \frac{1}{\sqrt{2q_0}}\right) \end{bmatrix}$$

From the chapter we know that

$$\alpha_1 = \frac{R_p}{R_q}\left(1 - \frac{1}{\sqrt{2q_0}}\right),$$

so the last column of A is a multiple of $(0, 1, -\alpha_1)$. We also want to see that $(1, -2\alpha_1, \alpha_1^2)^T$ is in the column space of A: we add $-1/R_q$ times the second column of A to $-R_p/(4R_q^2 q_0\sqrt{2q_0})$ times the last column of A:

$$\begin{bmatrix} 1 + 0 \\ -\frac{R_p}{R_q}\left(2 - \frac{3}{\sqrt{2q_0}}\right) - \frac{R_p}{R_q\sqrt{2q_0}} \\ -\frac{R_p^2}{R_q^2}\left(1 - \frac{1}{\sqrt{2q_0}}\right)\left(-1 + \frac{2}{\sqrt{2q_0}}\right) + -\frac{R_p^2}{R_q^2\sqrt{2q_0}}\left(1 - \frac{1}{\sqrt{2q_0}}\right) \end{bmatrix} = \begin{bmatrix} 1 \\ -2\frac{R_p}{R_q}\left(1 - \frac{1}{\sqrt{2q_0}}\right) \\ -\frac{R_p^2}{R_q^2}\left(1 - \frac{1}{\sqrt{2q_0}}\right)^2 \end{bmatrix};$$

this, using the expression for α_1, is just $(1, -2\alpha_1, \alpha_1^2)^T$. Thus, the range of A does include the vector b as required for there to be any solutions of $Ax = b$.

Appendix to Chapter 6

Chapter 7

Conclusions and Pointers

7.1 Conclusions

An image of a surface is the result of the interaction of a number of processes: how points in space are mapped to the image, how light is reflected from a surface and concentrated to form the image, the location and nature of light sources in the environment, and the geometry of the surface sitting in space. In trying to find out the constraints that an image and an assumed reflectance function put on potential solution surfaces, we have examined several local and global sources of information in the image.

We have examined three different sorts of small patches in the image. The first kind of patch contains neither good critical points nor any part of a bounding contour. In this case, the image dynamical system described in Chapter 4 will take any initial strip (positions in space with surface orientations) consistent with the image and not parallel to characteristic trajectories, and extend it to a solution surface patch consistent with the image patch. If we specify a path on the image patch, then the ambiguity of the solution surfaces can be summarized as the choice of a depth function along this curve; the reflectance function and contact 1-form will provide the orientation data needed to make an initial strip, and the characteristic strips will draw out a solution from this patch.

A second kind of patch we have analyzed contains a piece of the bounding contour. In this case we suggest (but have not completely proved) that the ambiguity of solution surfaces is of the same type as for an image interior patch, except that the initial path in the image is the bounding contour. A choice of depths along the bounding contour in the patch should lead to characteristics that draw out a solution surface consistent with the data on this patch. The bounding contour image data does provide extra data about the space-invariant reflectance function that might have generated the image: since we know the surface orientations along the bounding contour, and (theoretically) we know the brightness values along the bounding contour, our assumed reflectance map must match these values.

The third kind of patch is one on the image interior containing a critical point of both the image and the reflectance function. Given a general space invariant reflectance function, we showed in Chapter 5 that critical points of both the image and the reflectance function ("good" critical points) theoretically provide a great deal of information about surfaces that may have given rise to the image. We considered a good critical point as a critical point in a dynamical system on $\mathcal{C}(\mathbb{R}^3, 2)$, the space of all possible two-dimensional tangent spaces at each point in $\mathbb{R}^3$. Possible solution surfaces are surfaces that are invariant manifolds of the image dynamical system and contain the critical point. For very good critical points (those due to local maxima or minima of the reflectance function), this limits the possible solution surface patches consistent with the image of a generic surface to at most four. Two of these will be the unstable and stable manifolds of the dynamical system, and the local surface patch around the critical point can be expanded by following (either in positive or negative time) the trajectories away from the critical point. In the usual case of a maximum in brightness due to a reflectance function maximum, the stable and unstable manifold correspond to the convex and concave concave solution surface shapes. The two remaining possible invariant solution surfaces give rise to a saddle dynamical system on the solution surface, and so the surface cannot be

fully extended just by following the trajectories. In the case of a reflectance function maximum, these solution surfaces are saddle shaped in space at the critical point. In the case when symmetry is present (for example, a sphere lit from the viewing direction) the stable and unstable manifolds (concave and convex solution surfaces) will still be unique, but there may be an infinite number of solution surfaces with saddle dynamics consistent with the image data near the critical point.

In summary, for an interior image patch of a generic surface containing a very good critical point, the surface in general is theoretically determined up to a finite number of possible solutions. For a small image patch from the image interior not containing a critical point, we only determine the surface up to an arbitrary space curve (locally). For a small image patch containing a piece of the bounding contour, the solution surface also appears to be determined only up to an arbitrary space curve.

As described in Chapter 5, we have implemented two related techniques for finding the unstable manifold of an image dynamical system in a region near a critical point, and have seen that they provide robust solutions for the local shape from shading problem. We use the image dynamical flow to deform an initial surface over time; the Lambda Lemma says that as t goes to infinity, the deformed surface should C^1 approach the unstable manifold near the critical point. In one case we used an iterative scheme based on the equations

$$v_p = \lambda(E_x - p_x R_p - p_y R_q)$$
$$v_q = \lambda(E_y - q_x R_p - q_y R_q)$$

where $p(x,y)$ and $q(x,y)$ are the orientation coordinates of the current estimated surface, and $v_p(x,y)$ and $v_q(x,y)$ represent the change in $p(x,y)$ and $q(x,y)$ to get new surface orientation coordinates. λ is a step size parameter, and can be a function of (x,y). In the other case, we used the dynamical system flow to move a mesh of points $(x_i, y_j, p(x_i,y_j), q(x_i,y_j))$ on the solution surface a small amount; we then

reset the mesh using a least squares estimate of the new mesh values $p(x_i, y_j)$ and $q(x_i, y_j)$ at the original (x_i, y_j) sites.

Several constraints on a space invariant reflectance function are also available from the image dynamical system. First, there are the constraints from the brightness at "good" critical points and brightnesses along the bounding contour, where the surface orientations are known. Second, if the image is assumed to come from a single, smooth surface, then this surface is a smooth invariant manifold of the image dynamical system. In order for the image dynamical system defined by the image and an assumed reflectance function to have a solution surface, some subset of the invariant manifolds (including stable and unstable manifolds of convex and concave critical points) must merge rather than intersect in a one-dimensional curve. This is unstable behavior for generic dynamical systems, and so may provide a constraint on the choice of reflectance functions: although numerical errors will prevent exact matching of the invariant manifolds even with exact knowledge of the reflectance function, a wrong choice is likely to prevent the invariant manifolds from being even close together.

The modern methods of differential geometry provide a set of tools for reasoning about geometry and shape without always being tied to coordinate system expressions. They allow one to look at all the information available from the image and reflectance function in the image irradiance problem and see how properties of the dynamical system influence choices of possible solution surfaces; they can provide constraints on reflectance function choices as well. The theoretical view of the image irradiance problem as a dynamical system also suggests computational approaches to finding solution surfaces that are theoretically tractable.

7.2 Future Work

As reported in Chapter 5, we have not considered in detail the case of a good critical point caused by a saddle in the reflectance function. The critical point in this case may no longer be hyperbolic, and the study of the invariant manifolds

may be considerably more complicated. With pure imaginary eigenvalues, it may be possible for the characteristic trajectories to be closed orbits around the critical point. Neither closed orbits nor chaotic critical elements are considered here—it may of interest to find and study generic examples where these actually occur in an image dynamical system.

As discussed in previous Chapters, the image dynamical system provides constraints to restrict the choice of reflectance function based on the image data alone. Some kind of constraint is necessary: a slide projector on a flat white screen can give the impression of any possible three-dimensional scene. A vexing problem in understanding human image understanding is how we tell the difference between painted and unpainted surfaces (consider again Figure 1.1). We can clearly be "fooled"; indeed, we choose to be fooled when we interpret a photograph as a window onto a scene. We also do often have an opinion about whether a surface in an image is painted, or is lit in an unusual way.

What are reasonable reflectance functions to use in the absence of prior information? One might try to work through the physics of image formation as described by Horn and put the different constructions used (bidirectional reflectance map, radiance, and irradiance) into an invariant form connected to the geometry of the surfaces and lighting source distributions: for example, image irradiance becomes a two-form on the image. It would be interesting if reasonable properties of reflectance functions could be phrased in geometrical terms: e.g., symmetries, or other conservation principles, which might be applied to restrict the class of possible reflectance functions.

Another approach to studying the reflectance function question is to set the entire image irradiance problem in an infinite dimensional context instead of the finite dimensional one we have used. In this case, we might consider the image to be a map from the space of possible surfaces and possible reflectance functions to brightness functions on the image. There has been considerable work in looking at

fluid mechanics and the calculus of variations from this perspective (Marsden and Hughes, 1983); it may be useful to try this here as well.

The results of perturbing the image dynamical system with changes in reflectance functions should be examined: what else can go wrong globally if the wrong reflectance function is used, even if the bounding contour and the critical point brightnesses are matched? If there are no additional global inconsistencies, then how different are perturbed solution surfaces from the correct ones? Other global constraints that could be provided by the dynamical system should be explored, e.g., the possibility of trajectories folding over or the possibility of self-intersections of the solution surface in $\mathbb{R}^3$; these are not likely to be correct or stable solutions (unless perhaps for semi-transparent surfaces) given a smooth image.

We have focused on critical points and small patches containing the bounding contour. A critical point in an image can often be inferred from the brightness contours near it even if the critical point itself is obscured. A fuller exploration of the role of brightness contour configurations in determining solution surfaces is needed. One suspicion is that an annular region which "guarantees" a critical point on its missing interior might alone be sufficient to give invariant manifold results like those for critical points themselves, or perhaps provide limits on the behavior of possible invariant manifolds.

In footnote 7 of Chapter 5 (p. 109) we mention that the image dynamical system can be considered as a Hamiltonian dynamical system; this could be explored either on the six dimensional dual space $T^*\mathbb{R}^3$ where the null-space of a 1-form in $T^*\mathbb{R}^3$ at a point p can be used to identify a tangent space at p, or in the space invariant case this could be explored on the four dimensional reduced space $\tilde{\mathcal{C}}(\mathbb{R}^3, 2)$. As such, the behavior of "nearby" Hamiltonian systems as opposed to nearby generic dynamical systems is of interest: changing reflectance functions really corresponds to changing the Hamiltonian. Other standard features of Hamiltonian systems which we have not studied are closed orbits and chaotic orbits (Abraham and Marsden, 1985) as

mentioned earlier: under what conditions do these occur in an image dynamical system? Probably chaotic critical elements are evidence that we are far from the right track in choosing a reflectance function.

The work reported here examines in detail the image interpretation problem defined as exact recovery of the orientation at each point of the image. It is not clear that this is what the human perceptual system is concerned with. What are properties of a solution surface under reflectance function changes that people think remain constant and connected to the shape? How do alterations in the theoretical solution surface under changes in the reflectance function consistent with bounding contour and critical point data compare with human perceptual accuracy from the same image?

In Chapter 5 we suggested a couple of computational methods for finding invariant manifolds. From the theoretical description we have given, invariant manifolds of the dynamical system containing the critical points are important features to look for: better, more efficient ways of finding them would be of interest. We explored first order methods; as suggested in Chapter 5, perhaps higher-order methods similar to the Runge-Kutta vector field integration methods would give more effective results. We used a Connection Machine to explore the methods; other kinds of parallel architectures or neural network techniques could be explored.

The development of a system to actually solve shape from shading using these methods would be very interesting, even if the reflectance function must be provided beforehand. First, some reliable method of finding the stable and unstable manifolds of critical points drawn out as far as possible would be needed, perhaps including adaptive determination of the convergence region. Second, some comparison procedure to decide whether two such manifolds are the same or different would be needed: as discussed in Chapter 5, because of computational errors, exact matching cannot be expected even if the reflectance function is chosen exactly right. One way to accomplish the matching could be to tesselate the image into regions centered over

what one hopes are good critical points; for a typical smooth gently curved surface, there should be few such points (although in theory there could be many). At each critical point, the stable and unstable manifold could be computed; if a region contains bounding contour elements, this may be used to discard certain solutions if the invariant manifold does not come "close enough" to the bounding contour within a feasible distance from the observer. Choices from the remaining sets of invariant manifolds would have to be stitched together along the boundaries of regions: choices that were not "close enough" to each other would be discarded. There is one additional complexity: certain regions will not be either the stable or unstable manifold because the critical point is actually a saddle point on the surface; the surface solution here will have to be drawn out by extending other consistent stable or unstable manifolds to include this region. Some measure of goodness of fit of the solution would be needed based on how difficult it is to stitch the solution surfaces together.

With a system like this in hand, one could try to handle unknown reflectance functions as well: postulate a reflectance function consistent with the bounding contour and the critical points and see how bad the solution surface is. Note that there is a fair bit of specification already involved: a typical parameterized family of reflectance functions might very well be completely determined by the bounding contour data and the reflectance function maxima data. As suggested above, it would be interesting to explore the changes in both the qualitative behavior and the quantitative badness (in the sense of trouble stitching together the segments) as a result of changing the reflectance functions without changing the critical points or the bounding contour data.

This geometric method of analysis can be extended to include other cues. Perhaps the most natural extension is to include time: how do we make use of all the visual information available from a black and white television? As discussed in Chapter 2, one could consider theoretically a time-dependent embedding of a surface S, $i : S \times R \rightarrow \mathbf{IR}^3$; depending on the choice of class for i, we could study rigid motions of

the object or observer or various kinds of distortions of the object. This generates a time-dependent image dynamical system, and it would be interesting to study this to see how the time dimension affected features of the dynamical system.

One could also extend the analysis and methods to the case of flat surfaces or zero Gaussian curvature surfaces. An extended region in the image with constant brightness is not necessarily the image of a planar surface; it could be a carefully shaped and positioned curved surface (with zero Gaussian curvature), one whose normals all lie on a single constant brightness contour of a reflectance function. ((Brooks, 1982) discusses this in some detail.) The image is very unstable: a slight shift in position and the image will probably no longer have an open region of constant brightness, in contrast to the image of a planar surface which will still give a region of constant brightness if it is rigidly shifted. Perhaps by itself this sort of instability justifies a visual system labeling as planar an open region of constant brightness?

Extending this last idea, if issues of "stability" can be connected to the "likelihood" of a particular interpretation of an image being correct, then it is of interest to look at the class of transformations allowed in the definition of stability. If non-rigid transformations of surfaces are considered in deciding on stability of an interpretation, then the plane does not provide a stable constant brightness image either. Our experience that rigid motions of objects are important (ubiquitous because they correspond to the effects of observer motion) may lead us first to interpret images consistent with stability in the class of rigid motions; perhaps there is a hierarchy of object transformations, from rigid motion through area preserving distortions to general diffeomorphisms, which are involved in analyzing an image or moving image using stability. A more careful use of catastrophe theory to study the effects of transformations on the structure of the image dynamical system may be useful.

The ideas and methods of modern differential geometry suggest avenues both for theoretical and computational research in understanding human and machine vision. A visual problem formulated with most of the available information can be examined

both globally and locally with these techniques, as we have done with the image dynamical system for the shape from shading problem. The coordinate independent approach to geometric ideas allows coordinate choices to be made specifically to explore a particular feature of the problem.

There is an entire literature on dynamical systems developed over the last twenty years to study features of complicated dynamical systems. The mathematical tools were developed to help analyze geometric visualizations of complicated physical problems. It is also worthwhile to use these tools to study the principles behind vision itself.

References

Abraham, R., Marsden, J., *Foundations of Mechanics*, 2nd edition, Benjamin Cummings, 1985

Abraham, R., Marsden, J., Ratiu, T., *Manifolds, Tensor Analysis, and Applications*, Addison–Wesley, 1983

Blicher, P., "Edge Detection and Geometric Methods in Computer Vision," Ph.D. thesis, Department of Computer Science, Stanford University, 1985

Blicher, P., "The Stereo Matching Problem from the Topological Viewpoint," International Joint Conference on Artificial Intelligence, pp. 1046–1049, 1983

Brooks, M. J., "Shape from Shading Discretely," Ph.D. thesis, Department of Computer Science, Essex University, Colchester, England, 1982

Bruss, A., "The Image Irradiance Equation: Its Solution and Application," Ph.D. Thesis, Dept. of Electrical Engineering, MIT, 1980

Bülthoff, H. H., Mallot, H. A., "Interaction of Different Modules in Depth Perception," Internation Conference on Computer Vision, London, England, June 8–11, pp. 295–305, 1987

Chen, S., Penna, M., "Shape and Motion of Nonrigid Bodies," Computer Vision, Graphics, and Image Processing, 36:175–207, 1986

Deift, P., Sylvester, J., "Some Remarks on the Shape–from–Shading Problem in Computer Vision," Journal of Mathematical Analysis and Applications, 84:235–248, 1981

Edelen, D. G. B., *Applied Exterior Calculus*, Wiley and Sons, 1985

Frankot, R. T., Chellappa, R., "A Method for Enforcing Integrability in Shape from Shading Algorithms," International Conference on Computer Vision, London, England, June 8-11, pp. 118–127, 1987

Golubitsky, M., Guillemin, V., *Stable Mappings and Their Singularities*, Springer–Verlag, 1973

Guillemin, V., Pollack A., *Differential Topology*, Prentice–Hall, 1974

Haralick, R. M., Watson, L. T., Laffey, T.J., "The Topographic Primal Sketch," International Journal of Robotics Research, 2:50–72, 1983

Hawking, S. W., Ellis, G. F. R., *The Large Scale Structure of Space–Time*, Cambridge University Press, 1973

Hoffman, K., Kunze, R., *Linear Algebra*, Prentice-Hall, 1971

Horn, B.K.P., "Obtaining Shape from Shading Information," Chapter 4 in *The Psychology of Computer Vision*, P.H. Winston (ed.), McGraw–Hill, pp. 115-155, 1975

Horn, B.K.P., Brooks, M. J., "The Variational Approach to Shape from Shading," Computer Vision, Graphics, and Image Processing, 33: 174–208, 1986

Ikeuchi, K., Horn, B.K.P., "Numerical Shape from Shading and Occluding Boundaries," Artificial Intelligence, 17: 141–184, 1981

Koenderink, J.J., Van Doorn, A. J., "Photometric Invariants Related to Solid Shape," Optica Acta, 27: 981–996, 1980

Koenderink, J. J., "Operational Significance of Receptive Field Assemblies," Biological Cybernetics, 58: 163–171, 1988

Lang, S., *Analysis I*, Addison–Wesley, 1968

Marsden, J. E., Hughes, T. J. R., *Mathematical Foundations of Elasticity*, Prentice–Hall, 1983

Mingolla, E., Todd, J. T., "Perception of Solid Shape from Shading," Biological Cybernetics, 53: 137–151, 1986

Norton, A., personal communication, 1988

Palis, J., De Melo, W., *Geometric Theory of Dynamical Systems*, Springer–Verlag, 1982

Pentland, A. P., "The Visual Inference of Shape: Computation from Local Features," Ph.D. Dissertation, Psychology Department, MIT, 1982

Pentland, A. P., "Local Shading Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6: 170–187, 1984

Pentland, A. P., "Shading into Texture," Artificial Intelligence, 29: 147–170, 1986

Pong, T., Shapiro, L. G., Haralick, R. M., "Shape Estimation from Topographic Primal Sketch," Pattern Recognition 18: 333–347, 1985

Spivak, M., *A Comprehensive Introduction to Differential Geometry, Volume I*, Publish or Perish, 1979

Thorpe, J. A., *Elementary Topics in Differential Geometry*, Springer–Verlag, 1979

Warner, F. W., *Foundations of Differential Manifolds and Lie Groups*, Scott, Foresman and Company, 1971

CS-TR Scanning Project
Document Control Form

Date : 7 / 27 / 95

Report # AI-TR-1117

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☒ Technical Report (TR) ☐ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 214 (222-IMAGES)
Not to include DOD forms, printer instructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☒ Laser Print
☐ InkJet Printer ☐ Unknown ☐ Other:_____

Check each if included with document:

☒ DOD Form (2) ☐ Funding Agent Form ☒ Cover Page
☒ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages (by page number): _____

Photographs/Tonal Material (by page number): 10, 129-132, 134-137, 142-146

Other (note description/page number):

Description :	Page Number:
(A) IMAGE MAP	(1-6)
TITLE PAGE	1
UNMARKED BLANK	2
UNMARKED BLANK	3
UNMARKED BLANK	(7-214)
PAGES #'S	4-211
SCAN CONTROL, COVER, SPINE, DOD(2)	(215-219)
TRGT'S (3)	(220-222)
(B) CUT+PASTE F.C.S	ON PAGES 10

Scanning Agent Signoff:
Date Received: 7/27/95 Date Scanned: 7/31/95 Date Returned: 8/3/95

Scanning Agent Signature: Michael W. Cook

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AI-TR 1117	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Modern Differential Geometric Approach to Shape from Shading		5. TYPE OF REPORT & PERIOD COVERED technical report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Bror V.H. Saxberg		8. CONTRACT OR GRANT NUMBER(s) DACA76-85-C-0010 N00014-85-K-0124
9. PERFORMING ORGANIZATION NAME AND ADDRESS Artificial Intelligence Laboratory 545 Technology Square Cambridge, MA 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Blvd. Arlington, VA 22209		12. REPORT DATE June 1989 13. NUMBER OF PAGES 211
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Information Systems Arlington, VA 22217		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

shape from shading
computer vision
differential geometry
dynamical systems

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(see reverse)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 (cont'd)

Our visual system is a remarkably flexible and reliable source of information about the world. One of the major challenges for appreciating how vision contributes to our understanding of the world is understanding how it copes with the wide variety of lighting conditions, surfaces, and surface markings to provide accurate representations of the surfaces around us. The goal of the research reported here is to gain a better theoretical understanding of what lies behind the visual system's ability to generate robust surface interpretations from single grey scale images of smooth surfaces. In the course of doing this, a new robust shape from shading method is developed.

The image irradiance equation is written using coordinate independent notation and concepts from modern differential geometry and global analysis. This is done to help make explicit the assumptions about the image formation process, and to delay making these assumptions as long as possible. The method of characteristic strips used by Horn (Horn, 1975) can be interpreted as a dynamical system on the five-dimensional space of tangent planes, $\mathcal{C}(\mathbf{IR}^3, 2)$. Modern methods for analyzing the behavior of dynamical systems are used to show that solution surfaces for the shape from shading problem are invariant manifolds of the flow generated by the image dynamical system. The rest of the analysis assumes orthographic projection of the image and a space-invariant reflectance function, but does not assume any particular form or symmetry for the reflectance function.

Near critical points in the image dynamical system due to certain critical points in a smooth image, in general (i.e. in the absence of special symmetries) the dynamical system approach implies there will only be four possible smooth solution surfaces for the shape from shading problem. Two of these are the stable and unstable manifolds associated with the image dynamical system critical point. Two implementations for finding the unstable (or stable) manifold in this dynamical system are developed using the image dynamical system directly.

The shading information in a patch containing a piece of the bounding contour is also examined, and it appears to contribute more to an assessment of a reflectance function choice than to the determination of patches of solution surfaces consistent with the image.

Finally directions for future work are suggested, and some guidelines and caveats are provided for the development of image analysis systems based on these ideas.

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**

Scanned

Date: 7/31/1995

M.I.T. Libraries
Document Services