VARIATIONAL EDGE DETECTION

by

Niklas Karl Nordstrom

Memorandum No. UCB/ERL M90/43

31 May 1990

VARIATIONAL EDGE DETECTION

by

Niklas Karl Nordstrom

Memorandum No. UCB/ERL M90/43

31 May 1989

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

VARIATIONAL EDGE DETECTION

by

Niklas Karl Nordstrom

Memorandum No. UCB/ERL M90/43

31 May 1989

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Variational Edge Detection

Niklas Karl Nordstrom
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720

26 April 1990

This report is based on the unaltered dissertation by Niklas Karl Nordstrom, submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Electrical Engineering and Computer Sciences in the Graduate Division of the University of California at Berkeley.

The underlying research was in part supported by the National Science Foundation's Presidential Young Investigator Award IRI-8957274, by Lockheed Co. Inc. under grant M900, by the California State MICRO Program, and by DARPA under contract N00039-88-C-0292. This funding was greatly appreciated.

Variational Edge Detection

Niklas Karl Nordstrom

Variational Edge Detection

by

Niklas Karl Nordstrom

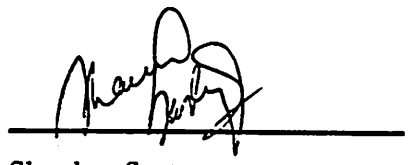
Abstract

In this dissertation we propose and study the properties of two global edge detection methods. Both methods are based on variational regularization, and result in the mathematical and computational problem of minimizing a cost functional depending on the edges as well as a piecewise smooth estimate of the true image function. The total cost consists in both cases of the sum of three separate subcosts—an edge cost penalizing the extent of the edges, a deviation cost promoting close approximation of the true image function by the estimated image function and a stabilizing cost favoring a smooth estimated image function.

In the first method, the edges are represented by parametrized curves in $\mathbf{R}^2$. We consider quite general such curves as well as the special case of spline curves. The central theoretical contribution is a proof of the existence of a solution to the cost minimization problem in a solution space containing a relatively large class of possible detected edges. By applying techniques of variational calculus we furthermore derive a number of optimality conditions for the edges as well as the estimated image function. Based on this analysis we develop a steepest descent type algorithm for detecting and locating edges represented by splines. We also describe a software implementation of the algorithm and present our experimental results.

In the second method, the edges are represented by a continuity control function defined on the entire image domain. While this method was originally intended as a general improvement on variational edge detection, it can also be viewed as a *biased anisotropic diffusion* method. This unification of the seemingly very different regularization and diffu-

sion approaches is remarkable. The new method furthermore shares the better properties of both these approaches. Indeed, it only requires the solution of a *single* boundary value problem on the *entire* image domain, and it converges to a solution of interest. As a consequence the new method is computationally less expensive than most, if not all, of the other regularization-based edge detection methods. It is also better for circuit implementations than previous (anisotropic) diffusion methods.



Shankar Sastry
Thesis Chair

Preface

This thesis is concerned with global edge detection. The first chapter contains a brief introduction to the topic. The four following chapters describe two distinct variational approaches to the edge detection problem. The major difference between the two approaches is in the representation of the edges. In chapter 2–4 we consider *curve-represented edge detection* in which the edges are represented by parametrized curves in the image domain, that is the region in $\mathbf{R}^2$ occupied by the image. The edges are thus formed by the union of the ranges of a collection of $\mathbf{R}^2$-valued functions. In the method in chapter 5, which we will refer to as *biased anisotropic diffusion*, the edges are instead represented by a single real valued (continuity control) function defined on the entire image domain. The edges consist in this case of the inverse image of some interval under this function.

Both the curve representation and the continuity control function representation have their merits, and are at present worth while pursuing. While the curve-based methods yield edge descriptions which are both compacter and likely to be more suitable for further processing, the biased anisotropic diffusion method requires less computation. A curve-based method could also use a continuity control function-based global edge detection method, such as biased anisotropic diffusion, as a preprocessing stage. This could provide it with a relatively accurate initial estimate of the edges which in turn would make it converge faster than if the initial estimate were obtained with a less accurate local edge detection method.

Acknowledgements

First of all I want to express my deepest gratitude to my wife Deborah Ellan for her wonderful love, patience and support during the four years we have known each other. This dissertation is dedicated to our new-born son Devlin Karl Kuumba. May he never have to read beyond this line!

I am most grateful to my academic advisers Professor Shankar Sastry and Professor Jitendra Malik for their inspiration, support, technical advice and careful review of this dissertation. The help from Shankar Sastry dates back to my earliest research in discrete time systems and in adaptive control. His impressive analytical ability, critique and encouragement have been a great source of confidence ever since. He also served on my qualifying committee. Jitendra Malik has been particularly helpful with guidance throughout my work in computer vision. I would also like to take this opportunity to express my appreciation for his sense of humor and his clever reflections on the world and its citizens, which I have always enjoyed during our many meetings.

During my graduate study at U. C. Berkeley I have also enjoyed generous help from other faculty members in the department of electrical engineering and computer sciences. I would like to thank Professor David Messerschmitt for advice and collaboration on some earlier research on echo cancellation as well as for serving on my qualifying committee and Professor Bob Brodersen for general advice and for reviewing some of my recent work on analog integrated circuits. I am moreover happy to credit Professors Charles Desoer and Jean Walrand for serving on my qualifying committee.

One of the greatest benefits of doing engineering research *at U. C. Berkeley* is the access to the expertise of one of the best mathematics institutions in the world. I would in particular like to thank Professor Alberto Grünbaum for reviewing this dissertation and Professor Gabriella Tarantello for taking special interest in and sharing her intuition about

the existence proof in chapter 3 during its early development. I must also credit Professor Alexandre Chorin for giving me the initial directions towards relevant mathematical literature to study and Professor Donald Sarason for serving on my qualifying committee.

A substantial portion of the work on this dissertation has involved the use of computers. These efforts would have been both excessively painful and time consuming, had it not been for the generous help from some of my fellow students. I owe most debt to Marc Singer who gave me a sophisticated and extremely useful skeleton program for the user interface to the computer implementations of the edge detectors described in chapters 4 and 5. I am also indebted to Richard Murray for his endless patience with and fast, polite and accurate answers to all my C-, LATEX-, UNIX- and X11-questions.

While being in graduate school in Berkeley I have met a number of interesting fellow students that in one way or another have contributed to my academic and personal development. I would especially like to mention Lester Ludwig and Bill Baird with whom I have shared interests far (out) beyond the school curricula. I would also like to name Venkat Anantharam, Er-wei Bai, Saman Betash, Marc Bodson, Myra Boenke, Randy Iceman Cieslak, Charles Coleman, Curt Deno, Nazli Gündes, Greg Heinzinger, Paul Jacobs, Güntekin Kabuli, Paul Kube, Stephane Lafortune, Edward Lee, Vijay Madisetti, Jeff Mason, Bob Minnichelli, Vallath Nandakumar, Andy Packard, Pietro Perona, Kris Pister, Tim Salcudean, Shahram Shahruz, Lars Svensson and Mats Torkelsson.

Before I started graduate school I was employed at a division of Telefonaktiebolaget LM Ericsson in Sweden I am very grateful to them for their financial support during my early period in Berkeley and for letting me be on leave until completing my Ph. D. degree requirements.

Berkeley is far from merely a university town. The east-bay is extremely rich in its variety of coexisting cultures and life styles. I feel very fortunate to have met a number of people outside academia who have given me a sense of American roots. I am most excited about all the remarkably talented local string benders and horn blowers who revived my music interest, taught me plenty and let me trash their performances.

Contents

List of Figures

Notation

Matrices

M^H	Hermitian transpose of matrix M	
M^T	transpose of matrix M	
$R_\times$	90° clockwise rotation matrix	$R_\times \doteq \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$

Vectors

e_n	outward normal unit vector at set boundary	
e_τ	unit vector tangential to directed curve in $\mathbf{R}^2$	
e_ν	unit vector normal to directed curve in $\mathbf{R}^2$	$e_\nu \doteq R_\times^T e_\tau$

Miscellaneous

$\lvert\alpha\rvert$	magnitude of multi-index $\alpha = \langle\alpha_k\rangle_{k=1}^n$	$\lvert\alpha\rvert \doteq \sum_{k=1}^n \alpha_k$
C^l	continuously differentiable l times	
C^∞	continuously differentiable any number of times	
class $C^{0,1}$	see page 76	
C^l-continuity	see page 108	

Derivatives

$\dot{f}$	first derivative of f
$\ddot{f}$	second derivative of f
$f^{(l)}$	lth derivative of f
$\partial f/\partial e_n$	directional derivative of f in outward normal direction

$\partial f/\partial x_k$	pointwise partial derivative of f with respect to x_k	
$D_k f$	distributional partial derivative of f with respect to kth argument	
$D^\alpha f$	$D_1^{\alpha_1} \cdots D_n^{\alpha_n} f, \quad \alpha = \langle \alpha_k \rangle_{k=1}^n$	

Relations

$\doteq$	equal to by definition	
$\hookrightarrow\!\!\hookrightarrow$	compactly embedded in	
$\subset\!\subset$	compact subset of open set	
$\cong$	congruent to	
$\bowtie$	joins	$(n,s) \bowtie (p,t)$ iff $\mathcal{N}(n,s) = \mathcal{N}(p,t) \neq \mathcal{N}_{\partial B}$
$\sim$	similar to	$(n,s) \sim (p,t)$ iff $\mathcal{N}(n,s) = \mathcal{N}(p,t)$
$\leftarrowtail$	"corresponds" to	$(\mathcal{B},\mathcal{N}) \leftarrowtail (n,s)$ iff $(\mathcal{B},\mathcal{N}) = (\mathcal{B}(n),\mathcal{N}(n,s))$

Operators

$\blacktriangle$	difference	
$\vee$	maximum	
$\wedge$	minimum	
$\nabla\cdot$	divergence	$\nabla \cdot f \doteq \sum_{k=1}^n \partial f_k/\partial x_k$
∇	gradient	$\nabla f \doteq [\partial f/\partial x_1 \cdots \partial f/\partial x_n]$
Δ	Laplacian	$\Delta f \doteq \sum_{k=1}^n \partial^2 f/\partial x_k^2$
$\tilde{}$	trivial extension	$\tilde{f}(x) \doteq \begin{cases} f(x) & \text{if } x \in \text{domain of } f \\ 0 & \text{otherwise} \end{cases}$
$\circ$	function composition	$(f \circ g)(x) \doteq f(g(x))$

Sets

$B_X(x,r)$	open ball of radius r centered at x in metric space X	
$\complement S$	complement of set S	
∂S	boundary of set S	
S°	interior of set S	
$\overline{S}$	closure of set S	
$\underline{f}$	support of function f	$\underline{f} \doteq \overline{f^{-1}(\complement\{0\})}$
F_f	graph of function f	

Spaces

$\mathbf{C}$	all complex numbers	
$\mathbf{N}$	all natural numbers	$\mathbf{N} \doteq \{1, 2, 3, \ldots\}$
$\mathbf{N}_0$	$\{0\} \cup \mathbf{N}$	
$\mathbf{R}$	all real numbers	$\mathbf{R} \doteq]-\infty, \infty[$
$\mathbf{R}_+$	all strictly positive real numbers	$\mathbf{R}_+ \doteq]0, \infty[$
$\mathbf{R}_-$	all strictly negative real numbers	$\mathbf{R}_- \doteq]-\infty, 0[$
$\mathbf{R}^n_\pm$	$\mathbf{R}^{n-1} \times \mathbf{R}_\pm, \quad n-1 \in \mathbf{N}$	
$\mathbf{R}^n_0$	$\mathbf{R}^{n-1} \times \{0\}, \quad n-1 \in \mathbf{N}$	
$\mathbf{Z}$	all integers	$\mathbf{Z} \doteq \{0, \pm 1, \pm 2, \pm 3, \ldots\}$
$C(\Omega)$	all continuous real-valued functions on Ω	
$C^l(\Omega)$	$\{f \in C(\Omega) : D^\alpha f$ is continuous , $\quad \alpha \in \mathbf{N}_0^n, \quad \lvert\alpha\rvert \leq l\}$	
$C^l_R(\Sigma)^2$	$\{f \in C^l(\Sigma)^2 : 0 \notin \dot{f}(\Sigma)\}$, Σ compact interval	
$C^l_R(\Sigma)^{2N}$	$[C^l_R(\Sigma)^2]^N, \quad N \in \mathbf{N}_0$	
$C^{l,h}(\Omega)$	$\{f \in C^l(\Omega) : D^\alpha f$ is h-Hölder continuous , $\quad \alpha \in \mathbf{N}_0^n, \quad \lvert\alpha\rvert \leq l\}$	
$C^\infty(\Omega)$	$\bigcap_{l \in \mathbf{N}} C^l(\Omega)$	
$C^\infty_0(\Omega)$	$\{f \in C^\infty(\Omega) : \underline{f} \subset\subset \Omega\}$	
$C^\infty_0(\mathbf{R}^n)\lvert\Omega$	$\{f\lvert\Omega : f \in C^\infty_0(\mathbf{R}^n)\}$	
$\mathcal{L}(X, Y)$	all bounded linear maps from X to Y	
X'	dual of space X	$X' \doteq \mathcal{L}(X, \mathbf{R})$
X^*	algebraic dual of space X consisting of all linear maps from X to $\mathbf{R}$	
$L_1^{\mathrm{loc}}(\Omega)$	all locally integrable real-valued functions on Ω	
$L_2(\Omega)$	all square integrable real-valued functions on Ω	
$\mathcal{H}^l(\Omega)$	$\{f \in L_2(\Omega) : D^\alpha f \in L_2(\Omega), \quad \alpha \in \mathbf{N}_0^n, \quad \lvert\alpha\rvert \leq l\}$	

Functions

$\langle \cdot, \cdot \rangle_\Omega$	inner product on $L_2(\Omega)$	$\langle f, g \rangle_\Omega \doteq \int_\Omega f g \, dx$
$\langle \cdot, \cdot \rangle_{\Omega,l}$	inner product on $\mathcal{H}^l(\Omega)$	$\langle f, g \rangle_{\Omega,l} \doteq \sum_{\substack{\alpha \in \mathbf{N}_0^n \\ \lvert\alpha\rvert \leq l}} \int_\Omega D^\alpha f D^\alpha g \, dx$
$m(\cdot)$	Lebesgue measure	
$\rho(\cdot, \cdot)$	Euclidean distance between points and/or sets	

Norms

$\|x\|_p$	$(\sum_{k=1}^{n} x_k^p)^{\frac{1}{p}}$, $x \in \mathbf{R}^n$, $p \in [1, \infty[$
$\|x\|_\infty$	$\bigvee_{k=1}^{n} \|x_k\|$, $x \in \mathbf{R}^n$
$\|x\|$	$\|x\|_2$, $x \in \mathbf{R}^n$
$\|\gamma\|_C$	$\sqrt{\sum_{n=1}^{N} \sum_{m=3o_n}^{M_n-1} \|v_{nm}\|^2 + \sum_{j=1}^{J} \|w_j\|^2}$, γ of configuration C
$\|f\|_{C(\Omega)}$	$\sup_{x\in\Omega} \|f(x)\|$, $f \in C(\Omega)$, Ω compact
$\|f\|_{C(\Omega)^K}$	$\bigvee_{k=1}^{K} \|f_k\|_{C(\Omega)}$, $f = [f_1 \cdots f_K]^T \in C(\Omega)^K$, Ω compact
$\|f\|_{C^l(\Omega)}$	$\bigvee_{L=0}^{l} \sup_{x\in\Omega} \|f^{(L)}(x)\|$, $f \in C^l(\Omega)$, Ω compact
$\|f\|_{C^l(\Omega)^K}$	$\bigvee_{k=1}^{K} \|f_k\|_{C^l(\Omega)}$, $f = [f_1 \cdots f_K]^T \in C^l(\Omega)^K$, Ω compact
$\|f\|_{C^{l,h}(\Omega)}$	$\|f\|_{C^l(\Omega)} + \bigvee_{\substack{\alpha\in\mathbf{N}_0^n \\ \|\alpha\|\le l}} \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{\|D^\alpha f(x) - D^\alpha f(y)\|}{\|x-y\|^h}$, $f \in C^{l,h}(\Omega)$, Ω compact
$\|f\|_{\mathcal{L}(X,Y)}$	$\sup_{\substack{x\in X \\ \|x\|\le 1}} \|f(x)\|_Y$, $f \in \mathcal{L}(X,Y)$
$\|f\|_{L_2(\Omega)}$	$\sqrt{\langle f, f\rangle_\Omega}$, $f \in L_2(\Omega)$
$\|f\|_{\mathcal{H}^l(\Omega)}$	$\sqrt{\langle f, f\rangle_{\Omega,l}}$, $f \in \mathcal{H}^l(\Omega)$

Chapter 1

Introduction

Before diving into the details of our paradigms we will briefly review the reasons for processing of image data, edge detection in general and global edge detection in particular. We will also review some of the previous approaches to global edge detection in some detail. This will put our paradigms in perspective, and thereby motivate our efforts.

1.1 Why Process Image Data?

There are basically two main reasons for processing image data:

1. To generate descriptions of the environment.
2. To enhance certain features of interest in a given image.

The first of these two tasks, which we will discuss further below, falls in the domain of computer vision. The second task belongs to "classical" image processing, and is per se not considered to be computer vision. It can be applied to images to be inspected by humans or as a preprocessing stage in computer vision.

1.1.1 Object Oriented Descriptions of the Environment

We—the human beings—tend to perceive our environment, (at least the part of it, that we are used to manipulate with some degree of success,) as being composed of distinct objects. Whether this is good or bad, we seem to be stuck with this way of thinking, and it is likely that all machines we design will inherit it from us. Detection, location, description and recognition of objects are therefore of interest for a variety of applications for example:

1. automatic manipulation of objects
2. collision avoidance
3. search for and tracking of objects
4. efficient presentation of information to a human operator
5. data reduction for efficient storage of information in a data base

1.1.2 Free-Space Descriptions of the Environment

Collision avoidance plays a vital role in navigation and path planning. In this context the objects in the scene are obstacles, and all that matters is which space they occupy. This information can of course be collected in an object oriented fashion, with the advantage of producing a concise representation. Another approach, which bypasses the object description task, is to generate a so called depth-map of free visible space by means of stereoscopic vision or a laser range finder.

1.2 Why Edge Detection?

Over the last three decades many techniques to detect, locate and link edges have been reported in the computer vision literature. A common concern is to locate the "significant" changes of some property of the image data with respect to position in the image (domain). Why is it of interest to locate such changes? Marr [1] discusses this question at some length, and distinguishes between three fundamental motivations for edge detection. Adding the important task of line drawing generation to his list we obtain the following four reasons:

1. line drawing generation
2. image segmentation
3. computational efficiency
4. imitation of human vision

1.2.1 Line Drawing Generation

It is evident from both the comic book literature and the engineering drawing tradition that plenty, if not most, of the information, which is necessary for understanding an image, can be concentrated in a line drawing. In fact a line drawing reveals a whole lot about the relative location and shape of the objects in a scene, and line labeling algorithms, which attempt to extract this information from line drawings have been developed [2, 3]. In order for these algorithms to work as intended, it is essential that the line drawing in question correctly reflects the smoothness properties of the image data. In other words the line drawing should be composed of piecewise smooth curves with tangent discontinuities in the "right" places. Since the location and the shape description of the objects in the scene represent the most relevant knowledge for automatic manipulation as well as object oriented collision avoidance, and since shape descriptions are useful for object recognition, line drawing generation is possibly the most important reason for edge detection.

1.2.2 Image Segmentation

Image segmentation is closely related to and competing in importance with the generation of line drawings. While line drawing analysis is concerned with the information carried by the edges themselves, the purpose of image segmentation is to partition the image into image segments, that is connected regions, which can thereafter be processed and analyzed one by one. Thus in image segmentation the interest is really in the complement of the edges. As a consequence the smoothness properties of the edges are not of as much concern as in the line drawing generation case.

A good image segmentation should yield image segments that correspond to smooth surface patches in the scene. One can then assume, that each image segment carries information about a single smooth surface of a single object or about the background. This is of importance for all further processing aiming at describing or recognizing the objects in the scene. The fact that each image segment corresponds to a smooth surface, also allows for noise suppression and surface reconstruction by means of linear filters operating locally over the separate image segments, without the errors that such filters are known to cause at discontinuities. (Minimizing a quadratic cost functional over each of the image segments is one example of such an operation.) This can be of value for generation of both object oriented and free-space descriptions of the environment as well as for feature enhancement

by standard image processing techniques.

1.2.3 Computational Efficiency

The edge points are in general sparse in comparison with all the points in the image. Any processing of the image data can therefore benefit in terms of complexity, if it can be cast to operate on edges rather than on whole images. Stereoscopic vision is one such example.

1.2.4 Imitation of Human Vision

Research in biological vision has according to [1] established the existence of anatomical structures that respond to abrupt changes in brightness and color with respect to position in the field of vision. This indicates that the human vision relies on some form of edge detection.

1.3 Edges as Discontinuities

Usually the image can, at least ideally, be represented by an image function, $z : B \rightarrow \mathbf{R}^n$, defined on some bounded connected image domain B in $\mathbf{R}^2$. A simple and by far the most common example of such an image function is a real-valued function, whose value at a point $(x, y) \in B$ represents the image brightness or grey-level at that point, but z can also represent other quantities such as depth (measured with an optical range finder) or disparity (acquired by stereoscopic vision). Other examples include image functions of multidimensional range ($n > 1$), frequently occurring in color vision, and binary image functions ($z : B \rightarrow \{0, 1\}$), which are common in drawing and text processing.

In general the edge detection problem can be thought of as the task of finding the "significant" discontinuities of z and/or its derivatives. Which of these derivatives are to be considered, depends of course on the level of ambition and the computational resources at hand, but also and more importantly on which kind of image data is to be processed. The most interesting discontinuities in the scene are those of the distance to and the tangent plane of the visible surfaces. If z represents depth, these discontinuities obviously correspond to discontinuities in z and ∇z respectively, whereas if z represents brightness, they both cause discontinuities in z itself. As a result the various approaches to edge detection differ

in, which kinds of discontinuities are considered significant. Attempts have for example been made to find idealized discontinuities in z, (step edges,) [4, 5, 6, 7, 8, 9, 10, 11] and ∇z, (roof edges,) [5], points of large magnitudes of the first and second derivatives of z [12] and points of local maxima in $\|\nabla z^T\|$ in the direction of ∇z [12, 13, 14].

Another source of variation in approach is the fact, that in practice the true image function z, that one would obtain by pure projection onto the image plane, is not known. In reality noise is present in the image formation process, which also blurs the image. Finally at some stage the image function is sampled. Thus one is only given the values of a smooth approximation ζ to a noisy version of z on a discrete subset S of B, and exactly what properties of $\zeta|S$ or even ζ that correspond to discontinuities in z and its derivatives is far from obvious. We will refer to ζ as the *original image function* because it represents the original data. (Some authors, for example Geman and Geman, use a different terminology.)

1.4 Why Global Edge Detection?

One can distinguish between four different benefits that global edge detection provides over the older local methods:

1. nondestructive noise suppression
2. smooth edges
3. conceptually appealing models
4. simultaneous image function estimation

Among these benefits the first two are the major reasons for pursuing global edge detection. The other two are hardly strong enough to motivate the relatively costly global techniques on their own.

1.4.1 Non-destructive noise suppression

Until the mid 1980s edge detection was dominated by local methods such as local best fit techniques and linear filtering. (An excellent survey of a large body of the relevant literature is presented in [15].) These methods typically involve convolution with some kernel followed by thresholding. Whatever the initial outlook might have been, the resulting

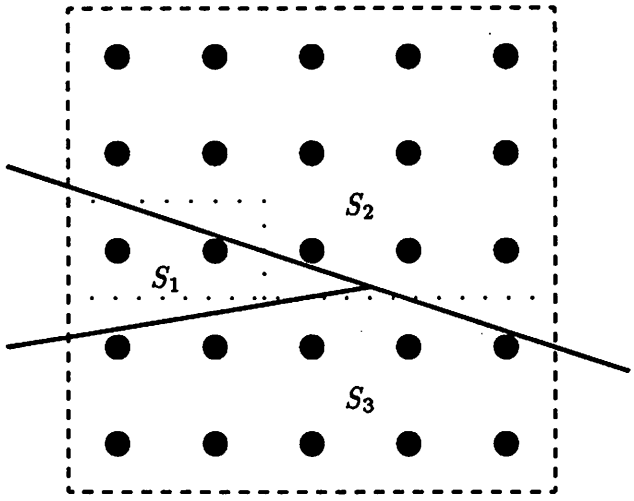


Figure 1.1: T-junction partitioning 5×5 best fit window into three regions (S_1, S_2 and S_3).

convolution basically has one common purpose: to suppress noise prior to the decision about the presence of an edge is made. As a result of this kind of noise reduction L-junctions disappear, and T-junctions are disconnected. The possibility of successful subsequent line drawing analysis is thereby seriously diminished. Another problem with these methods is that they cause dislocation of the edges in the vicinity of high gradients in the image function. One way to attempt to remedy these problems is to introduce nonlinearities in the noise reduction. This can be done locally as well as globally. Nalwa and Binford [9] apply best fit techniques, which are nonlinear in the estimated parameters. They fit functions, which are constant along some direction in $\mathbb{R}^2$, to the data. Such a function is an adequate model only for a short smooth segment of an edge. Hence the technique will have potential problems with both L- and T-junctions.

If one were to do a proper job for all junctions of interest with a local best fit method, it should be possible to obtain a good fit for each one of them. Besides being substantially more expensive than the common simpler best fit techniques this would inevitably require, that (basis) function families be included in the model, for which some of the fitted parameters depend on the original image function ζ over only a tiny portion of its domain, thus in practice on maybe just one or two samples of ζ on the grid S. Such a scenario is depicted in figure 1.1. Here some of the best fit parameters must obviously depend solely on the values of the two pixels centered in the region S_1. They are therefore very sensitive to

noise—a contradiction of the main purpose of the best fit idea. This problem could be cured if some samples in the extension of S_1* outside the window were included. In fact, why not make use of all the samples in the extension of S_1? Well, this is what a number of recent global edge detection methods are all about. In effect these methods seek to suppress the noise in the image function by smoothing it separately within each image segment, without smoothing across the segment boundaries. Since these boundaries, that is the edges, are not known in advance, but to be detected, this problem is nontrivial, and as a consequence a variety of approaches has emerged.

1.4.2 Smooth Edges

In all descriptions (of the environment) there is a trade-off between accuracy and simplicity. This condition is reflected explicitly in most global edge detection schemes by means of a penalty for including edges in the description. Without such a penalty the edges would fill up the image domain, yielding the most accurate, the least simple and a completely useless description of the environment.

Usually the edge penalty is proportional to the length of the edges or something similar. This is sufficient to encourage edges, that are smooth in some sense. In its weakest interpretation, and this is the rule among the global edge detection approaches to date, this means, that the edges are represented by samples on a grid in such a way, that it is easy to fit a smooth or piecewise smooth curve to the sample points. However, stricter smoothness conditions can readily be achieved by explicitly incorporating such requirements in the edge penalty [16]. We will have more to say about this later.

1.4.3 Conceptually Appealing Models

Global edge detection lends some intuitively very appealing models and mathematically well-developed ideas from other fields of science. These include variational regularization techniques, mechanical models of solids, probabilistic models of random Markov fields and physical models of diffusion.

Many problems in early vision such as edge detection, computation of lightness, computation of optical flow, shape from shading, stereopsis and surface reconstruction are

*By the extension of S_1 we mean the (unique) component of the intersection of the image domain and the complement of the edges that contains S_1.

ill-posed in the sense of Hadamard, that is their solutions fail to exist, to be unique or to depend continuously on the given data [17, 18]. In the case of edge detection it is the last of these three conditions that is of concern. To find a satisfactory solution to such an ill-posed problem the solution space has to be restricted by imposing constraints representing some kind of a priori knowledge about the possible solutions—a process known as regularization. Such methods have been studied in mathematics, and a common technique is to replace the original problem by that of minimizing a (cost) functional including stabilizing terms of the kind proposed by Tikhonov [19]. In the edge detection context this technique is often given the interpretation that it resembles problems of continuum mechanics [20, 21, 22, 23, 24]. Another way to regularize the edge detection problem is to introduce a probabilistic model for the solution space, and solve an estimation problem [25, 26].

Aside from regularization, physical diffusion models related to the heat equation have been explored [27, 28, 29, 30].

Finally it is intuitively satisfying, that all the information carried by the image function, not only that associated with a neighborhood of its (and some of its derivatives') discontinuities, is applied in determining the edges. This idea is also certain to find some support from simple psychophysical experiments on human vision.

1.4.4 Simultaneous Image Estimation

Besides finding the discontinuities of the image function and possibly some of its derivatives, typically the global edge detection algorithms also produce a piecewise smoothed version of the image function, which preserves the detected discontinuities. This "side effect", usually thought of as estimation, reconstruction, recovery or restoration of an image function, which is sparsely sampled and/or corrupted by noise, has different applications depending on, what kind of data is being processed. Reconstruction of the visible surfaces from sparse depth data can be used to generate a dense free-space description of the environment [20, 21, 26]. Noisy images can be restored by estimation of the brightness function [25, 26]. The same can of course be done with noisy dense depth data [31].

In some instances the estimation of the image function seems to be the major concern. Some authors argue, however, that the main purpose of estimating the image function is to find its discontinuities. [24].

1.5 Previous Global Edge Detection Approaches

By a *global* edge detection method we understand an edge detection method that utilizes *all* the given information about the original image function to detect and/or locate the edge elements or segments in the image domain.

1.5.1 Early Efforts

Until the early 1980s global edge detection in the sense of the "definition" above was restricted to relatively simplistic methods such as search for special features by means of accumulator arrays [32, 33, 34, 35, 36] and histogram techniques [37, 38]. A brief presentation of these methods can be found in [39, p123–131, p152–153]. In the methods based on accumulator arrays the space of possible edges is limited by the type(s) of features one is looking for—typically lines, circles or other simple geometric curves—and thus far too restricted to account for any but rudimentary descriptions of the environment. The histogram techniques suffer from instability with respect to the given data, unless the image function is known to represent very simple scenes arranged with the particular technique in mind, for example parts on a conveyer belt under controlled light conditions.

1.5.2 Recent Efforts

In recent years a number of interesting and mathematically more sophisticated global edge detection methods with far more general application domains have appeared in the computer vision literature. These include various regularization techniques and anisotropic diffusion.

Variational Regularization

Terzopoulos [20] reconstructs a piecewise C^1-surface from sparse depth data by minimizing a functional representing the potential energy of a thin plate under tension and a collection of springs suspended between the plate and the given original data points. He solves the problem for known but arbitrarily irregular domains, that is the edges, which in his case are depth and orientation discontinuities in the reconstructed surface, are prespecified—to be neither detected nor located.

In a later paper [21] his energy functional is somewhat different. First of all sparse

surface orientation data is incorporated along with the sparse depth data. Secondly and more importantly the internal potential energy of the thin plate is replaced by a so called *"controlled-continuity stabilizer"* defined as a sum of spatially weighted generalized spline functionals [40]. The weighting functions are also referred to as continuity control functions. Since the edges can be represented by the sets at which the weighting functions vanish, this paradigm allows for edge detection by adjusting these functions, so as to lower or minimize the energy functional. Two methods for this adjustment are discussed.

The first method essentially applies local edge detection techniques to the reconstructed surface. Depth discontinuities are detected where "opposing bending moments are imparted to the surface", and the magnitude of the gradient of the estimated image function, (that is the function, whose graph is the reconstructed surface,) is greater than a certain threshold. The procedure turns out to be equivalent to detecting zero-crossings of the Laplacian of the estimated image function, and then weeding out inflections due to small insignificant ripples in the reconstructed surface by thresholding the gradient magnitude at these zero-crossings. Since the reconstruction of a surface without the presence of edges is equivalent to filtering with a second order Butterworth low-pass filter [40]—a smoothing operation not too different from convolution with a Gaussian kernel, this method ends up being similar to that proposed by Marr and Hildreth [14], and thus prone to exhibit similar drawbacks.

The second weighting function adjustment method attempts to minimize the energy functional—now augmented with a term penalizing the presence of edges—with respect to the weighting functions as well as with respect to the estimated image function. This method basically makes sense only after the problem has been discretized. The adjustment is carried out by flipping the values of the discretized weighting functions between their two only possible values—from zero to one or vice versa—at all points of their (common) discrete domain, where such a flip lowers the value of the energy functional. A new controlled-continuity stabilizer is thereby obtained, according to which the surface is subsequently reconstructed. This adjustment-reconstruction procedure is then repeated until convergence is achieved. However, since the flip decisions are made without regard to the alteration in the reconstructed surface implied by such a flip, optimality of the solution cannot be claimed in general.

In both the papers the surface reconstruction (for fixed continuity control functions) amounts to solving an elliptic Euler-Lagrange equation. This is done by means of the

finite element method. Multigrid methods are also being used in order to give descriptions of the environment at different scales of resolution, as well as to speed up the computations.

Mumford and Shah [16] locate piecewise smooth edges by minimization of an energy functional of the form

$$E(f,B) \doteq \mu^2 \int_R (f-g)^2\,dx + \int_{R\setminus B} \|\nabla f^T\|^2\,dx + \mu\nu \int_B d\ell$$

where f and g are the estimated and original image functions respectively, $R \subseteq \mathbf{R}^2$ is the image domain, B is the union of a finite number of smooth curves meeting each other and the boundary of R only at their endpoints, μ and ν are strictly positive parameters and $d\ell$ indicates integration with respect to arc length. The paper also treats the simpler case, when $R \subseteq \mathbf{R}$, and B is just a finite collection of points. This analysis can also be found in [41]. Since the space of possible edges in this setting is vast and complicated, they do not attempt a proof of the existence of an optimal estimated image function $\check{f}$ and an optimal set of edges $\check{B}$, which minimize E. While admitting, that this is the "central mathematical problem", they instead conjecture the existence of $\check{f}$ and $\check{B}$, and proceed to develop an algorithm for minimizing E. They observe, that for a given set B there exists a unique estimated image function f_B, such that

$$\check{E}(B) \doteq E(f_B,B) = \inf_f E(f,B)$$

and that f_B and thus $\check{E}(B)$ can be found by solving an elliptic boundary value problem. Starting with a set B obtained by means of a local edge detector, derived by asymptotic analysis ($\mu \longrightarrow \infty$), they continue to minimize $\check{E}(B)$ by updating B according to a steepest descent rule. The update of B depends on f_B. Hence the elliptic boundary value problem has to be solved at each iteration. This is done numerically with a standard multi-grid elliptic problem solver program.

Blake and Zisserman [24] reconstruct image functions from brightness data, sparse depth data, dense depth data and edge data by fitting a weak membrane, (appropriate for brightness and edge data,) or a weak plate, (most appropriate for depth data) to the given data. This is done mainly in the interest of edge detection. An exception is the fitting of a one-dimensional weak membrane (weak string) to the ('angle', 'arc length')-data associated with a jagged edge in order to obtain a piecewise smooth edge. In this case one is interested

in the reconstructed real valued (image) function representing the smoothed edge as well as in the discontinuities representing the corners of the edge.

The term "weak" is used to indicate, that the membrane and the plate satisfy so called "weak continuity constraints". This means, that they are allowed to fracture in the presence of high stresses. The weak plate is also allowed to crease in the presence of high bending moments.

The fitting is in the sense of minimizing functionals. The membrane functional is in its continuous version identical to the functional considered by Mumford and Shah in [16]. The plate functionals on the other hand are basically identical to the edge penalty augmented functional with controlled-continuity stabilizers considered by Terzopoulos in [21]. Thus there is little new in the paradigm. What is new, and very interesting, with this approach, is the minimization technique. While Terzopoulos, Mumford and Shah and others as well use the fact, that the reconstructed image function can be eliminated from the energy functional by solving a well-behaved elliptic problem, Blake and Zisserman go the opposite way, and eliminate the continuity control functions, by them referred to as the line process. Once the reconstructed image function has been found, the line process, which incidentally is of the most interest, can easily be recovered.

The line process elimination leads to a nonconvex minimization problem, to which calculus of variations does not readily apply. In order to solve it a method, referred to as *graduated nonconvexity (GNC)*, is introduced. This involves approximating the true energy E, now expressed as a function of the values of the discretized reconstructed image function, by a family $\{E_p\}_{p\in[0,1]}$ of functions, such that $E_0 = E$, and E_1 is convex. Since E_1 is convex, an image function that minimizes E_1 can be found by means of a steepest descent algorithm. The GNC-method then proceeds to minimize a sequence $\langle E_{p_i}\rangle_{i=1}^{I}$, where $p_i \downarrow$ as $i \uparrow$, using the image function resulting from minimizing $E_{p_{i-1}}$ as the starting point for the minimization of E_{p_i}.

While unable to guarantee convergence to the global minimum, the GNC-method is still quite appealing for two very important reasons.

1. Unlike Terzopoulos' and Mumford and Shah's algorithms it both *detects* and locates the edges in a truly global fashion.

2. If the approximating family $\langle E_{p_i}\rangle_{i=1}^{I}$ is well chosen, the authors claim, the GNC-procedure "pulls" towards a "good" local minimum of E, without the costly search

required by algorithms like simulated annealing, which promise convergence to the global minimum.

It should be noted that the trick of eliminating the line process, results in "thin" edges only if it is applied after the problem has been discretized. For the continuous problem its formal counterpart could yield edges consisting of quite arbitrary, in particular nonmeager, sets. In contrast, the elimination of the estimated image function works in both the continuous and the discrete cases for any chosen class of edges.

Lee and Pavlidis [42] claim to improve on Terzopoulos' surface reconstruction method by proposing *"discrete regularization"* as a preprocessing technique for prior discontinuity detection. According to the authors such a method is to be preferred over postvalidation, for example by detecting "opposing bending moments". However, the discrete regularization technique per se does not detect the discontinuities. Instead it produces a *"smoothing polygon"* approximating the data, or more precisely, a piecewise affine approximation, which avoids large slope changes between neighboring linear segments of its graph. This problem is resolved by resorting to postvalidation by means of local methods quite similar to those proposed by Terzopoulos [21]—a somewhat contradictory turn-around, especially as little analysis is presented to support the idea that the proposed local methods are better than any others. The authors argue, that postvalidation at this stage has less disadvantages, than if applied as part of the original surface reconstruction, because the discrete regularization is much cheaper in computation, and one can thus afford to iterate the polygon-smoothing (followed by discontinuity detection) many more times, than one would have been able to iterate the corresponding surface reconstruction. This argument is, however, less convincing, than one would desire.

For dense data the discrete regularization leads to exactly the same linear system of equations, as does the finite difference scheme, that Terzopoulos employs for solving the Euler-Lagrange equation for the reconstructed surface. Thus in this case the smoothing polygon is just the reconstructed surface, and the computation just as expensive.

For sparse data on the other hand a couple of serious questions arise, which are not being addressed in the paper. First of all, while the generalization to two dimensions is straight forward for data sampled on a regular grid, this is not the case for irregularly sampled data. Since sparse data is commonly gathered by stereopsis, this should be of major

concern. Secondly for sparse data one would suspect, that the piecewise affine approximation represented by the smoothing polygon is in general quite crude in comparison with that represented by a piecewise smooth reconstructed surface, and that this could easily lead to dislocation and spurious detection of edges, and possibly prevent edges from being detected. Thus there is a question of, whether the proposed reiteration of the discrete regularization with intermediate updates of the edges can make up for these potential drawbacks, and if so, whether in the end any computation is being saved. Since the discrete regularization technique is meant to replace later postvalidation in the surface reconstruction process, this question is of utmost importance.

Incidentally all the examples of experimental results in the paper exhibit data sampled on regular grids, whence one tends to suspect, that the generic sparse data case has hardly been studied enough.

Probabilistic Regularization

Geman and Geman [25] restore images from dense brightness data corrupted by noise. More precisely, they solve for (f, l), where f is a piecewise constant image function of finite range defined on the grid of pixel sites, and l is a function that explicitly marks the discontinuities of f. The function l, referred to as a (sample function of a stochastic) "line process", is defined on the grid associated with the mutual boundaries between neighboring pixels, and takes values in a finite set corresponding to the possible edge elements at each grid point. The functions f and l thus play the roles of the reconstructed surface and the continuity control function(s) respectively in the mechanical models discussed above. In this case, however, the problem is discretized from the very outset. The problem is regularized by introducing probabilistic models for the image formation process as well as for the prior knowledge about the solution space. The solution is then defined to be the *maximum a posteriori probability (MAP)* estimate of the true image function and its discontinuities given the data. For the image formation process the authors adopt a quite general model incorporating the effects of blurring, nonlinear sensors and "cancelable", (for example additive or multiplicative,) sensor noise. The solution space and the noise process they model by independent *Markov random fields (MRFs)* specified in terms of their independent *Gibbs distributions.* These models lead to a Gibbsian posterior distribution, and the MAP estimate can thus be found by minimizing the associated energy function.

This is done numerically by stochastic optimization using simulated annealing and the Metropolis algorithm.

Marroquin [26] also estimate images from data corrupted by noise. His paradigm is basically the same as that of Geman and Geman. There are, however, three essential differences:

1. The estimated image function is basically piecewise continuous.
2. The given data may be sparse.
3. The MAP estimator is replaced by the optimal Bayesian estimator with respect to some arbitrary cost function.

The first two of these differences are important generalizations of Geman and Geman's approach. While piecewise continuity of the estimated image function is insufficient for reconstruction of the visible surfaces from depth data, it is adequate for restoration of images and their discontinuities from brightness data.

The third difference is in theory also a generalization, but with the particular cost functional that Marroquin chooses as his favorite, it turns out to be equivalent (modulo quantization effects) to estimating the individual pixel values, (that is the samples of f,) by their posterior mean values, and the individual edge elements, (that is the samples of l,) by their maximum posterior marginal probabilities. Marroquin refers to these estimates as the *"thresholded posterior mean" (TPM)* and the *"maximizer of the posterior marginals" (MPM)* respectively. When he finally discusses an algorithm for computing these estimates, however, he restricts attention to image formation in the presence of additive zero-mean Gaussian white noise. Under these circumstances the TPM estimate reduces to the MAP estimate, and the energy function associated with the Gibbsian posterior distribution is the same as that of a membrane under tension. The estimation problem therefore ends up being solved by minimizing this energy function.

Since the target space of the estimated image, though finite, is of considerable cardinality, the computational burden would be increased enormously, if the optimization method of Geman and Geman were applied without modification. Similarly to Terzopoulos and Mumford and Shah, Marroquin therefore makes use of the fact, that for given edges, that is for a given sample function l of the line process, an optimal estimate $\hat{f}_l$ of the

image function can be found by solving a linear system of equations. For this task he chooses a deterministic iterative technique. The remaining part of the problem, consisting of minimizing the energy of $(\check{f}_l, l)$ with respect to l, is then done by stochastic optimization using the same methods as Geman and Geman.

Anisotropic Diffusion

Perona and Malik [27, 28, 29] have introduced anisotropic diffusion as a method of detecting discontinuities in image functions at multiple scales of resolution. This method is not global in the sense of our "definition" above. In fact, the diffusion per se does not detect any discontinuities; the detection is done in a postprocessing stage, typically with some local method. Neither does it converge to a globally optimal state of any kind; it converges to a constant image function. Nevertheless it shares the most significant advantage of the global edge detection methods, that is, it suppresses noise without smoothing the original image function across its discontinuities—at least not right away.

The method operates by repeatedly filtering the image function with a *data dependent* kernel of small support. By design of the kernel this is analogous to diffusion governed by the partial differential equation

$$\frac{\partial I}{\partial t}(x, y, t) = \nabla \cdot [c(x, y, t) \nabla I(x, y, t)]$$

where $\nabla\cdot$ and ∇ are the divergence and gradient operators respectively with respect to the spatial coordinates (x, y), t is the time corresponding to the iteration index in the numerical scheme, I is the image function—here thought of as representing density and c is the conductivity. In order to discourage diffusion across the edges the conductivity is made to depend on I according to

$$c(x, y, t) \doteq g(\|\nabla I(x, y, t)^T\|)$$

where g is a strictly positive strictly decreasing function.

The filtering stages produce a sequence of diffused images of successively lower resolution, each of which can be subject to edge detection in a postprocessing stage. After only a few iterations the result is the same as that of a "discontinuity preserving" local smoothing operator. Later on, as data propagate throughout the image domain, noise is efficiently suppressed, while the more interesting large scale discontinuities still remain, or

are even sharpened. Finally all discontinuities disappear, and the image function approaches a constant. At some stage in the iteration remarkably impressive results can be obtained by postprocessing with the most rudimentary local edge detector. The task of finding this stage, however, has so far been a matter of manual inspection.

In Nordstrom [30] I have recently discovered a very close relationship between Terzopoulos' variational regularization approach and Perona and Malik's diffusion method; the former can be modified to yield a truly global edge detection method, which can also be interpreted as an anisotropic diffusion method, very similar to that of Perona and Malik. The global nature of Perona and Malik's method is thereby made transparent. Besides this benefit the new method possesses some of the more attractive properties of both variational regularization and anisotropic diffusion. As the other variational methods it converges—at least in practice—to a "good" local minimum of a cost functional, and the (diffused) image function reaches a limit, which is very appropriate for rudimentary local edge detection. Unlike other existing diffusion methods the new method does thus not require premature termination of the diffusion process. Supervision is consequently unnecessary. At the same time the new method shares the relatively low computational cost of Perona and Malik's anisotropic diffusion method.

The new method, referred to as *biased anisotropic diffusion (BAD)*, will be analyzed in detail in chapter 5.

Discussion

The regularization methods described above have a lot in common. In all of them the solution space is of the form $F \times L$, where F and L are the spaces of possible estimated image functions and possible detected edges respectively. Furthermore the sought solution minimizes an energy functional $E : F \times L \to \mathbf{R}$ of some sort. It is also the case, that the energy for given edges, that is the functional $E_l : F \to \mathbf{R} : f \mapsto E(f, l)$, is a positive definite quadratic form (plus some arbitrary constant). With the single exception of the approach by Mumford and Shah it is assumed somewhere along the line, that the edge space L is finite in an essential way—not just because of the quantization necessitated by machine computation. This finite edge space assumption greatly simplifies the problem both theoretically and computationally.

The theory is simplified, because the question of existence of a solution, that is

whether there exists an optimal pair $(\check{f}, \check{l}) \in F \times L$ that minimizes the energy E, is reduced to the question, whether there for each $l \in L$ exists an optimal $\check{f}_l \in F$ that minimizes E_l. For the energy functionals in question this is a well-studied problem; given l an optimal image function $\check{f}_l$ does indeed exist, and in the case of dense (or continuous) data $\check{f}_l$ is also unique. A detailed analysis of this problem is presented in chapter 3. In addition to the finite edge space assumption Geman and Geman further assume, that the image function space F is finite. Thus in their case the existence problem is completely trivial.

The computational problem is also simplified, because the finite edge space assumption grants successful utilization of a number of algorithms and problem formulations, which would otherwise fail. Some examples are:

1. Terzopoulos' method of flipping the values of the continuity control functions
2. Blake and Zisserman's line process elimination
3. Lee and Pavlidis' discrete regularization method
4. the Metropolis algorithm and its various versions used by Geman and Geman as well as Marroquin

Besides having the advantage of greater simplicity, the finite edge space assumption, at least in its forms discussed above, also has some less tractable implications. Indeed, the detected edges merely consist of an unorganized collection of linear edge elements located on a grid and each of length (approximately) equal to the width of a pixel. Consequently they lack several properties of importance for later processing, for example:

1. piecewise smoothness
2. subpixel localization
3. one-dimensional structure
4. a compact parametric analytic description

These inadequacies, at least the first three, can be remedied by edge linking and various other postprocessing techniques. See for example [43, 44, 45, 46, 47]. However with such an approach there is no mechanism that governs the trade-off between optimal edge location with respect to the fit of the estimated image function versus optimal edge location with respect to parametric description, smoothness, sensible linking, etc.

Anisotropic diffusion provides an interesting alternative to the global edge detection methods based on regularization. While preserving junctions and locating edges correctly, problems of thinning, linking, smoothing and ultimately of generating a parametric description of the edges still remain.

Chapter 2

A Variational Approach to Curve-Represented Edge Detection

In this chapter we present our paradigm for detecting and locating curve-represented edges in images. The method is global and based on regularization. It is adequate for dense data, and in its present version most suitable for processing of brightness data. However, heuristic generalizations better suited for other data types, such as depth data, are straight forward, and rigorous generalizations conceivable.

As it has turned out, our approach is very similar to that of Mumford and Shah [16]★. In fact the paradigm to be presented includes theirs as a special and probably the most interesting case. Our contributions to curve-represented edge detection are the following:

1. Our framework is more general.
2. We have solved the central problem of existence of optimal edges for a relatively large class of possible detected edges.
3. We have developed an algorithm for locating spline curve-represented edges.

★At the time the paradigm presented herein was developed, this report by Mumford and Shah was not widely circulated, if at all printed.

We begin by choosing representations for the edges to be detected thereby implicitly selecting appropriate solution spaces for the global edge detection problem. We then regularize the problem by posing it as one of minimizing a cost functional. Various possible such costs are presented, and their variations, that is their differentials, calculated. Although this analysis is in part heuristic, it is well motivated for two good reasons. It provides valuable guidelines for further strict mathematical treatment, where the intuition is often buried in the formalism. This is exemplified by the analysis of the existence of optimal edges in chapter 3. The heuristic optimality analysis also serves as a vehicle for developing global edge detection methods such as the algorithm to be described in chapter 4.

The functions that we choose to represent both the image and the edges throughout most of our analysis are defined on connected, (in particular nondiscrete,) domains, and we postpone the eventually necessary discretization of the problem until after the development of most of the theory. We thus end up with an algorithm for a "continuous" problem. By doing so, we believe, we solve the mathematically most relevant problem, and thereby generate a method with which every method originating from a "discrete" approach to the same problem should be consistent.

2.1 Edge Representations

A principal immediate purpose of detecting edges in any image is to find the curves in the image domain which correspond to significant discontinuities in depth and orientation of the visible surfaces in the scene. Sometimes discontinuities in reflectance properties and illumination of the visible surfaces are also of interest. In the case of brightness data, which is our major concern in this thesis, all these discontinuities give rise to discontinuities in the image function itself, at least in the *true image function* one would have obtained by pure projection onto the image plane without the corruptive influences of blurring and noise. Our goal is therefore to detect and locate the set on which the (true) image function exhibit significant discontinuities. Throughout the rest of this part we will refer to this set informally as "the edges" and formally as the *discontinuity set.* Naturally the final product of our edge detection method will depend on how we choose to represent this set. We have in this approach chosen to work with edge representations in terms of general parametrized curves and spline curves.

2.1.1 Parametrized Curves

It was mentioned in chapter 1 that finite edge space models, that is models including only finitely many discontinuity sets, lead to solutions which are often inappropriate for later processing. One way to introduce an infinite edge space is to allow all possible discontinuity sets in the plane and represent them by their characteristic functions. Although this continuity control function representation, (which is part of the paradigm to be considered in part 5,) provides analytic descriptions and permits subpixel localization of the edges, it has some disadvantages. First of all it overlooks the fact that the edges of interest consist of piecewise smooth curves in the plane. In other words, this edge space is too large to guarantee meaningful solutions. Secondly parametrization of the edges in terms of their spatial coordinates in $\mathbf{R}^2$ leads to a cumbersome data structure in which the most important information about the geometric interrelationships among the points in the discontinuity set, viz. the linking information, is not explicit. Thus even if edges consisting of piecewise smooth curves are detected, this is by no means apparent from the representation of the solution. Finally it is hard to see how the infinity of an edge space based on parametrization by the spatial coordinates would be able to "survive" the for computational purposes necessary discretization of the image domain. A curve-represented edge space will of course in any practical implementation also be finite. However, this restriction, due to quantization rather than sampling, is independent of the discretization of the image domain, and does hardly affect the possible structures and locations of the edges.

For the reasons just mentioned we have chosen to represent the edges as we would like the solution to the edge detection problem to be represented. More precisely, we represent the edges by a finite collection of *edge segments* each of which is a parametrized curve in $\mathbf{R}^2$. The desired smoothness of the edge segments is controlled by assuming that their Cartesian components in $\mathbf{R}^2$ belong to the function space $C^l(\Sigma)$[†] for some $l \in \mathbf{N}_0$ and some compact interval $\Sigma \in \mathbf{R}$. The discontinuity set is then defined to be the set

$$D_\gamma \doteq \bigcup_{n=1}^{N} \gamma_n(\Sigma)$$

where $\gamma_1, \ldots, \gamma_N \in C^l(\Sigma)^2 \doteq C^l(\Sigma) \times C^l(\Sigma)$ are the parametrizations of the $N \in \mathbf{N}_0$ edge segments and the vector $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T$ is referred to as the *image segmentation.* As usual

[†]For any set $\Omega \subseteq \mathbf{R}^K$ the space $C^l(\Omega)$ is defined to consist of all functions $f : \Omega \to \mathbf{R}$ whose partial derivatives of orders $\leq l$ all exist and are continuous. In particular $C^0(\Omega)$ consists of all continuous functions $f : \Omega \to \mathbf{R}$.

the space $C^l(\Sigma)$ is given the norm

$$\|f\|_{C^l(\Sigma)} \doteq \bigvee_{L=0}^{l} \sup_{\sigma \in \Sigma} |f^{(L)}(\sigma)| \qquad f \in C^l(\Sigma)$$

In order to have a measure of distance between image segmentations (with the same number of edge segments), for $K \in \mathbf{N}$ we also define a norm on

$$C^l(\Sigma)^K \doteq \prod_{k=1}^{K} C^l(\Sigma)$$

by

$$\|f\|_{C^l(\Sigma)^K} \doteq \bigvee_{k=1}^{K} \|f_k\|_{C^l(\Sigma)} \qquad f \doteq [f_1 \cdots f_K]^T \in C^l(\Sigma)^K \tag{2.1}$$

which of course generates the product topology on $C^l(\Sigma)^K$.

The edge representations proposed above have the following advantages:

1. The edges have one-dimensional structure. No further thinning or linking processing is necessary.

2. The edges have desired smoothness.

3. The edges have analytic descriptions which permit prompt calculations of many of their geometric properties such as position, tangent vector, curvature and arc length.

4. The representations allow subpixel localization of the edges.

5. The edge space is given a metric which depends on the geometry of the edges so that it reflects the intuitive notion of distance between edges. As we shall see later in this chapter and in chapter 3, the metrics that we have chosen are also in consonance with the weak notions of distance implied by the kind of cost functionals that typically result from standard regularization techniques. They are therefore adequate as a basis for various optimization strategies aimed at solving the regularized edge detection problem by minimizing such functionals.

The first three of these features are essential for accurate line drawing analysis and therefore of value for various tasks such as automatic manipulation, object description and object recognition.

2.1.2 Splines

For the purpose of storage of the detected edges and for computational feasibility of the global edge detection itself as well as of later processing, it is desirable that the edge representation be as compact as possible. In particular it has to be finite dimensional. If the finite dimensionality is achieved simply by sampling the edge segment parametrizations on a finite subset of Σ, the explicit smoothness of these functions is lost. An inherently finite dimensional model for the edge segments is therefore called for. One simple such model, which is frequently used in computer graphics, is the *B-spline* representation. We have settled for so called *uniform cubic* B-splines for the following two reasons:

1. Uniform cubic B-splines are twice continuously differentiable. This means that the most interesting geometric properties of the edge segments, viz. position, tangent vector, curvature and arc length, are well-defined and easy to compute.

2. Uniform cubic B-splines are easy to implement.

This representation is briefly outlined below. A comprehensive exposition of the theory of B-splines and the related beta-splines, of which the former constitute a special subclass, can be found in [48] and [49] respectively.

The B-spline representation of the edges is essentially the same as the general parametrized curve representation discussed above with the additional requirement, that each edge segment is a planar uniform cubic B-spline curve and thus parametrized by a univariate uniform cubic B-spline with range in $\mathbf{R}^2$. Such a spline is by definition a function $q : [0, M] \to \mathbf{R}^2$, $M \in \mathbf{N}$, specified in terms of a sequence $\langle v_m \rangle_{m=0}^{M+2}$ of *control vertices* in $\mathbf{R}^2$ according to

$$q(m+\sigma) \doteq q_m(\sigma) \doteq \sum_{r=0}^{3} v_{m+r} b_r(\sigma) \qquad m = 0, \ldots, M-1 \qquad \sigma \in [0,1] \tag{2.2}$$

where the so called *basis functions* $b_0, \ldots, b_3 : [0,1] \to \mathbf{R}$ are given by

$$\begin{bmatrix} b_0(\sigma) \\ b_1(\sigma) \\ b_2(\sigma) \\ b_3(\sigma) \end{bmatrix} \doteq \begin{bmatrix} -\sigma^3 + 3\sigma^2 - 3\sigma + 1 \\ 3\sigma^3 - 6\sigma^2 + 4 \\ -3\sigma^3 + 3\sigma^2 + 3\sigma + 1 \\ \sigma^3 \end{bmatrix} \frac{1}{6} \qquad \sigma \in [0,1] \tag{2.3}$$

The basis functions are, as the name suggests, and as is easily verified, linearly independent. Each one of the functions $q_m : [0,1] \to \mathbf{R}$, $m = 0, \ldots, M-1$, therefore

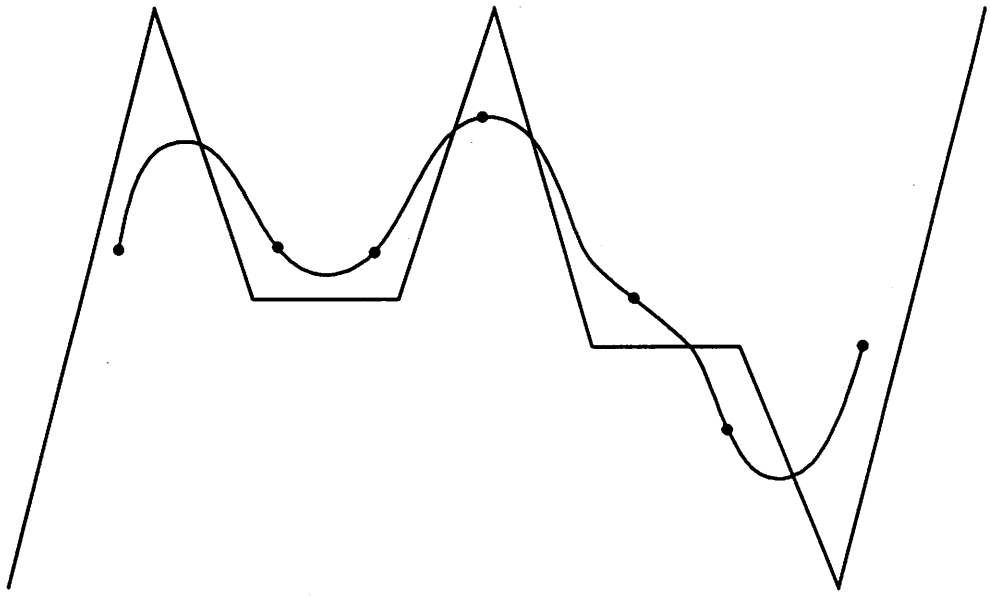


Figure 2.1: Uniform cubic B-spline curve and its defining control polygon.

has a unique representation as a linear combination of the basis functions with its associated control vertices $v_m, \ldots, v_{m+3}$ as combination coefficients. There is thus a one-to-one correspondence between the splines and their defining sequences of control vertices.

The basis functions are chosen to be polynomials of degree ≤ 3, and such that any possible spline q is twice continuously differentiable (regardless of its defining control vertices). These requirements determine them uniquely up to a common factor of which a certain choice yields the expressions in (2.3). This special choice has the intuitively very appealing consequence that every point on the mth *spline curve segment* $q_m([0,1])$ is a convex combination of its associated control vertices—a property referred to as the *convex hull property.*

In general the control vertices do not lie on the spline curve that they specify. Due to the convex hull property, however, the spline curve is a reasonable C^2-smooth approximation of the *control polygon* formed by the line segments joining its control vertices in consecutive order. See figure 2.1.

Before finishing the description of the B-splines themselves a final word has to be said about the end conditions that we use for the spline curves. There are two separate cases to be considered.

For representation of smooth closed curves we use what we naturally shall call *closed splines.* To obtain such a spline from a closed control polygon we simply augment the corresponding control vertex sequence by repeating the three first vertices in their original order at the end of the sequence. Thus for a closed spline $v_{M+m} = v_m, \quad m = 0, 1, 2$. This yields a smooth closed spline curve on which all of its original control vertices $v_0, \ldots, v_{M-1}$ have equally strong influence. Its endpoints are thus pure artifacts.

For representation of smooth curves with distinct endpoints we use what we shall refer to as *open splines.* These are splines with so called *triple vertex end conditions*, which means that the first and last control vertices in the sequence both are repeated twice. In other words, $v_0 = v_1 = v_2$ and $v_M = v_{M+1} = v_{M+2}$. The most important characteristic of this end condition is that the endpoints of the spline curve coincide with those of its control polygon. Indeed, so do the corresponding tangent vectors and (zero) curvatures at these points as well. This makes it easy to constrain the different spline curves in the edge representation to form junctions at any desired points in the plane. It can of course happen that a spline curve of this kind is closed by accident. In this case, however, the curve is in general not smooth at $v_0 = v_{M+2}$.

In many symbolic manipulations it is necessary to distinguish between closed and open splines. With each spline we will therefore associate a binary number o which we shall refer to as the *openness* of the spline (curve). If $o = 0$, the spline is closed, whereas if $o = 1$, it is open.

From the convex hull property it is obvious that too short control vertex sequences result in pathological spline curves. The control polygon of a closed spline must have at least three vertices, or else the spline curve collapses to two line segments lying on top of each other. Likewise the control polygon of an open spline must have at least two vertices in order not to collapse to a point. Since the number of genuine vertices in the control polygon, as can be seen, equals $M - o$, we will therefore always demand that $M \geq 3$.

In order to represent the edges by B-splines, we associate with the nth edge segment, $(n = 1, \ldots, N \in \mathbb{N}_0,)$ a sequence of control vertices $\langle v_{nm} \rangle_{m=0}^{M_n+2}$, $(M_n \geq 3,)$ and parametrize it by the corresponding spline $\gamma_n : \Sigma_n \doteq [0, M_n] \rightarrow \mathbb{R}^2$ defined by

$$\gamma_n(m + \sigma) \doteq \gamma_{nm}(\sigma) \doteq \sum_{r=0}^{3} v_{n,m+r} b_r(\sigma) \qquad \sigma \in [0, 1] \qquad m = 0, \ldots, M_n - 1 \qquad (2.4)$$

The nth edge segment is thereby also given a certain openness o_n. Since the number of control vertices, that is required for accurate spline representation of a given curve, varies

with the length and the shape of the curve, the control vertex sequences associated with the edge segments are, as indicated by the notation, allowed to depend on n. Consequently the domains of the parametrizations differ from segment to segment. Except for this minor modification, the B-spline representation is just a special case of the general parametrized curve representation discussed above. Thus the discontinuity set is given by

$$D_\gamma \doteq \bigcup_{n=1}^{N} \gamma_n(\Sigma_n)$$

where $\gamma_n \in C^2(\Sigma_n)^2, \quad n = 1, \ldots, N$, and the image segmentation is given by

$$\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in \prod_{n=1}^{N} C^2(\Sigma_n)^2$$

Junctions of various kinds are represented within the spline representation framework by constraining subsets of the endpoints of the edge segments (or equivalently of their control polygons) to coincide. We refer to the common position of such a set of endpoints as a *junction*, and the associated constraints as *interconnection constraints*. For notational and computational convenience we also consider the position of an unconstrained endpoint of an edge segment parametrized by an open spline to be a junction.

For the purpose of metrization of the space of all possible image segmentations we partition this space into equivalence classes of image segmentations with similar configuration with regard to the interconnection constraints. Accordingly we say that two image segmentations $\beta \in \prod_{n=1}^{N_\beta} C^2([0, M_{\beta n}])$ and $\gamma \in \prod_{n=1}^{N_\gamma} C^2([0, M_{\gamma n}])$ have the same configuration if:

1. $N_\beta = N_\gamma = N$
2. $M_{\beta n} = M_{\gamma n}, \quad n = 1, \ldots, N$
3. The control vertices associated with β and γ are subject to identical (spline curve) end conditions and interconnection constraints.

It is clear from the earlier discussion that there is a one-to-one correspondence between the splines and their control polygons. This correspondence permits the construction of a simple norm on the space of image segmentations of any given configuration in terms of the Euclidean norms (of parts) of the control vertex sequences associated with the splines of the image segmentations. Since some of the vertices are *dependent*, in the

sense that they are forced by the end conditions and the interconnection constraints to coincide, one has the choice, and it is geometrically most adequate, to exclude from the Euclidean norms the extra copies of the therefore duplicated vertices. Thus for any image segmentation $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T$ of a given configuration $\mathcal{C}$ and with associated openness variables $o_1, \ldots, o_N$, control vertex sequences $\langle v_{nm} \rangle_{m=0}^{M_n+2}$, $n = 1, \ldots, N$, and junctions $w_1, \ldots, w_J \in \mathbf{R}^2$ we define

$$\|\gamma\|_{\mathcal{C}} \doteq \sqrt{\sum_{n=1}^{N} \sum_{m=3o_n}^{M_n-1} \|v_{nm}\|^2 + \sum_{j=1}^{J} \|w_j\|^2} \tag{2.5}$$

If the class of image segmentations of configuration $\mathcal{C}$ is equipped with the obvious appropriate addition and scalar multiplication operations, it becomes a normed vector space with norm $\|\cdot\|_{\mathcal{C}}$. Since the splines approximate their control polygons reasonably well, the metric induced by this norm shares most of the advantages of the $C^l(\Sigma)^{2N}$-norms discussed earlier.

2.2 Cost Functional Problem Formulation

In the "continuous" edge detection problem that we are addressing the objective is, as mentioned earlier, to find the significant discontinuities of the true image function one would obtain by pure projection onto the image plane. Since we are interested in brightness data, we will assume that this image function is real-valued. However, the true image function is not known. Instead the image formation process yields an *original image function* $\zeta : B \to \mathbf{R}$ which is corrupted by blurring as well as measurement noise. We will assume that the original image function is square integrable, that is $\zeta \in L_2(B)$. Its domain B, henceforth referred to as the *image domain*, will always be assumed to be a connected open bounded set in $\mathbf{R}^2$. Besides the effects due to imperfection of the image formation process there are additional disturbances caused by true but insignificant discontinuities which are irrelevant for the generation of a useful description of the environment at a given scale of resolution. Many fine textures in the scene, not to forget dirt and dust, just add unnecessary complexity to the edges which is more likely to hinder than to support any subsequent processing stages. These insignificant discontinuities can in a sense be regarded as a form of noise though quite different in nature from the measurement noise.

Because of the blurring in the image formation process the discontinuity detection

more or less boils down to a search for large magnitudes and/or zero-crossings of various (linear) combinations of partial derivatives of the original image function. This problem is (for the physically relevant topologies on the data space) unstable with respect to the given data, and thus ill-posed in the sense of Hadamard [17, 18]. The presence of noise (including insignificant discontinuities) therefore necessitates some kind of regularization of the global edge detection problem. We have already encountered different approaches to regularization in our review of various global edge detection methods in chapter 1, and there are other techniques as well [17]. In similarity with those methods we attempt to find an "estimate" $z^\ddagger$ of the true image function along with an explicit representation of its associated discontinuity set D_γ by minimizing a *total cost* functional of the form

$$\mathcal{E}_N(\gamma) + \mathcal{D}_{C_\gamma}(z, \zeta) + \mathcal{S}_{C_\gamma}(z)$$

where the *edge cost* $\mathcal{E}_N$ measures the extent and complexity of the discontinuity set, the *deviation cost* $\mathcal{D}_{C_\gamma}$ measures the discrepancy between the estimated and the original image functions, and the *stabilizer* or *stabilizing cost* $\mathcal{S}_{C_\gamma}$ measures the spatial variation of the estimated image function. We will discuss the purposes and definitions of these three separate functionals in more detail in the following subsections. The edge cost depends as indicated on both the number of edge segments N and the image segmentation γ. The other two costs depend on γ only through the complement $C_\gamma \doteq B \setminus D_\gamma$ of the discontinuity set relative to the image domain. This set is referred to as the *continuity set* because it is the domain of the *estimated image function* $z : C_\gamma \to \mathbf{R}$, which is required to be the limit (in some appropriate space to be discussed later) of functions which are continuously differentiable a certain number of times.

Both the deviation cost and the stabilizing cost are essentially associated with the estimated image function z, and depend only implicitly on the edges. They will indeed both be expressed as integrals of z-dependent integrands over the continuity set, which of course is determined by the edges. In spite of the different purposes of the deviation cost and the stabilizer, their sum is therefore most often naturally treated as a single functional. We will refer to this functional as the *image cost.*

[‡]The letter z will henceforth be reserved for the *estimated image function.* Since the true image function—earlier also denoted by z—does not figure in any of the following analysis, this inconsistency will not cause any ambiguities.

2.2.1 Deviation Costs

The purpose of the deviation cost $\mathcal{D}_{C_\gamma}$ is to ensure that the estimated image function z is a faithful approximation of the original image function ζ. In principle it could be chosen so as to discourage discrepancy between the partial derivatives of z and ζ up to any given order. Natural examples of such deviation costs are given by

$$\mathcal{D}_{C_\gamma}(z,\zeta) \doteq \int_{C_\gamma} \sum_{i=1}^{I} \sum_{k_1=1}^{2} \cdots \sum_{k_i=1}^{2} \left[\frac{\partial^i (z-\zeta)}{\partial x_{k_1} \cdots \partial x_{k_i}} \right]^2 dx \qquad I \in \mathbf{N}_0 \tag{2.6}$$

However, the nature of the noise, one is willing to tolerate, puts a limit on which of the partial derivatives of ζ are, (if at all existing,) representative of the corresponding partial derivatives of the true image function that one wants to estimate. In the case of brightness data one can at most assume that the energy and the magnitude of the noise signal, corresponding to its L_1- and L_∞-norms respectively, are small. On the other hand, the derivatives of the noise function can be quite large in spite of blurring. First of all in order for ζ to be a good approximation of the potentially discontinuous true image function the blurring must be kept at a minimum so that even very high frequency components of the input signal are not being attenuated. Secondly, since noise enters the signal throughout the different stages of the image formation process, the blurring of the sensor noise, which is last to enter, might be much less than that of the true image function itself. Under these circumstances only the simplest of the deviation costs in (2.6) is appropriate. For any open subset Ω of the image domain we therefore define

$$\mathcal{D}_\Omega(z,\zeta) \doteq \int_\Omega (z-\zeta)^2 \, dx \qquad z \in L_2(\Omega) \tag{2.7}$$

For any given image segmentation γ the deviation cost is the given by the functional $\mathcal{D}_{C_\gamma}$.

2.2.2 Stabilizing Costs

The purpose of the stabilizer $\mathcal{S}_{C_\gamma}$ is to restrict the space of possible estimated image functions and thereby regularize the (edge detection) problem so that, as the name suggests, stability with respect to the initial data is achieved. Several classes of such stabilizing functionals have been studied in the mathematical theory of ill-posed problems. This theory was pioneered by Tikhonov [50, 19]. He proposed a general class of stabilizers for univariate regularization of the form

$$\mathcal{S}(z) \doteq \int_{\mathbf{R}} \sum_{i=0}^{I} \mu_i \left(\frac{d^i z}{dx^i} \right)^2 dx$$

where $\mu_0, \ldots, \mu_I : \mathbf{R} \rightarrow \overline{\mathbf{R}_+}$ are prespecified continuous weighting functions. These sums of spatially weighted Sobolev norms are since then commonly referred to as *Tikhonov stabilizers*. For multivariate regularization *generalized spline functionals* of the form

$$S(z) \doteq \int_\Omega \sum_{k_1=1}^{K} \cdots \sum_{k_I=1}^{K} \left(\frac{\partial^I z}{\partial x_{k_1} \cdots \partial x_{k_I}} \right)^2 dx$$

have been considered with varying generality of the domain $\Omega \subseteq \mathbf{R}^K$, its dimension K and the "order of regularization" I [51, 52, 53, 54]. Other possible stabilizers for multivariate regularization are the functionals

$$S(z) \doteq \int_\Omega \Delta^I z \, dx$$

where $\Omega \subseteq \mathbf{R}^K$, and Δ denotes the K-dimensional Laplacian operator $\sum_{k=1}^{K} \frac{\partial^2}{\partial x_k^2}$. These *"Laplacian" stabilizers* are clearly somewhat simpler than the multivariate generalized spline functionals just mentioned. If one intends to apply direct minimization methods to solve the regularized problem, this could be of some advantage. In contrast the "Laplacian" stabilizers lead to Euler equations which are more complicated than those resulting from the multivariate generalized spline functionals. Thus for indirect minimization via solution of the Euler equation, which is our intended approach, the situation is reversed.

Examples of the stabilizers discussed above have been used for regularization of a wide range of early vision problems with varying degree of success [18]. A common flaw of these stabilizers in this context is that they do not allow the estimated image function z to be discontinuous. This problem was addressed by Terzopoulos [40] who proposed further generalizations of the multivariate generalized spline functionals. His stabilizing functionals, referred to as *controlled-continuity stabilizers* are given by

$$S(z) \doteq \int_{\mathbf{R}^K} \sum_{i=1}^{I} \mu_i \sum_{k_1=1}^{K} \cdots \sum_{k_i=1}^{K} \left(\frac{\partial^i z}{\partial x_{k_1} \cdots \partial x_{k_i}} \right)^2 dx$$

where the (weighting) functions $\mu_1, \ldots, \mu_I : \mathbf{R}^K \rightarrow [0, 1]$, referred to as *continuity control functions* are in general discontinuous. They are in particular able to make jumps to zero, and edges, where the partial derivatives of z of order $\geq j$ are allowed to be discontinuous, are represented by the sets

$$\bigcap_{i=j+1}^{I} \mu_i^{-1}(\{0\}) \qquad j = 0, \ldots, I-1$$

For reasons that we have already discussed we have (in this approach) chosen to avoid the use of continuity control functions. Instead we let the estimated image function depend on the edges through its domain C_γ. For any open set $\Omega \subseteq \mathbf{R}^2$ we therefore define

$$\mathcal{S}_\Omega(z) \doteq \int_\Omega \sum_{i=0}^{I} \mu_i \sum_{k_1=1}^{2} \cdots \sum_{k_i=1}^{2} \left(\frac{\partial^i z}{\partial x_{k_1} \cdots \partial x_{k_i}} \right)^2 dx \qquad I \in \mathbf{N}_0 \tag{2.8}$$

where $\mu_0, \ldots, \mu_I \in \overline{\mathbf{R}_+}$ (are constants), and the zero times iterated sum occurring for $i = 0$ is consistently defined to be equal to its only possible term, that is

$$\sum_{k_1=1}^{2} \cdots \sum_{k_0=1}^{2} \left(\frac{\partial^i z}{\partial x_{k_1} \cdots \partial x_{k_0}} \right)^2 \doteq z^2$$

For any given image segmentation γ the stabilizing cost for our bivariate problem is then given by the functional $\mathcal{S}_{C_\gamma}$.

Since we are only interested in discontinuities of z itself and not its partial derivatives, the most important term in the sum in (2.8) is the term corresponding to $i = 1$. (In the proof of existence of optimal edges in chapter 3 we will with this condition in mind as a matter of fact only consider the case when $I = 1$, $\mu_0 = 0$ and $\mu_1 = \mu > 0$.) The other terms are not useless, however, because sharp changes in the original image function ζ, presumably due to step discontinuities in the true image function, in the absence of an edge segment also cause the cost associated with the higher order terms ($i > 1$) to rise, thereby contributing to the call for an edge segment at the point in question. Such an occurrence due to, what is sometimes referred to as the *Gibbs phenomenon*, is depicted in figure 2.2. In regions of constant gradient $\nabla\zeta$ on the other hand no such phenomenon will occur no matter how large the value of $\|\nabla\zeta^T\|$. The higher order terms can thus serve as a mechanism for distinguishing regions of smooth shading from true edges. The term corresponding to $i = 0$ is more questionable. It basically discourages the most extreme values of ζ from being taken too seriously. If for example those values for some reason were highly suspected of being caused by noise, this could have some merit.

2.2.3 Edge Costs

The purpose of the edge cost $\mathcal{E}_N$ is to limit the extent and complexity of the edges. This is not only important for the goal of generating simple useful descriptions of the environment; it is absolutely necessary for the cost functional minimization approach

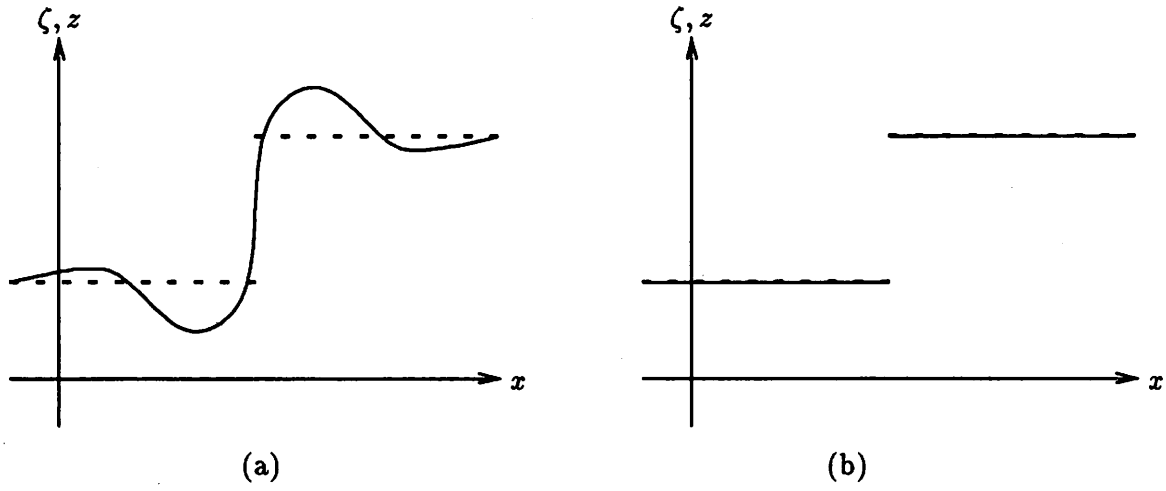


Figure 2.2: Effects of a step discontinuity in the original image function ζ (dashed) on a piecewise C^2-smooth estimated image function z (solid) in the absence (a) vs. presence (b) of an (estimated) edge at the location of the step.

to have any meaning at all. Indeed, the other part of the total cost, that is the image cost $\mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)$ has, as we shall see in chapter 3, a global minimum with respect to z which decreases monotonically as the continuity set C_γ decreases. Thus under any attempt to minimize this cost with respect to both z and γ, the edges would tend to fill up the entire image domain. This would undoubtedly lead to completely useless results.

It is of course desirable that the edge cost only reflects the properties of the discontinuity set, and that it is in other ways independent of the particular parametrizations of the edge segments. This concern leads us to edge costs composed of quite natural quantities associated with the edge segments themselves such as their total number, their total arc length and various parametrization independent measures of their *"curvedness"*. It is also desirable that the form of the edge cost is suitable for the variational methods that we intend to use for solving the regularized edge detection problem. Any appropriate curvedness measure should therefore, just as the arc length, be expressible in terms of integrals of smooth algebraic expressions of the edge segment parametrizations and their derivatives. This suggests the use of integrands involving some power of the curvature which in turn requires that the edge segment parametrizations $\gamma_1, \ldots, \gamma_N$ are regular, that is that

$$\dot{\gamma}_n(\sigma) \neq 0 \qquad \forall \sigma \in \Sigma \qquad n = 1, \ldots, N$$

For $l \in \mathbb{N}$ we therefore define the function spaces

$$C_R^l(\Sigma)^2 \doteq \{f \in C^l(\Sigma)^2 : 0 \notin \dot{f}(\Sigma)\}$$

and

$$C_R^l(\Sigma)^{2N} \doteq \prod_{n=1}^{N} C_R^l(\Sigma)^2 \qquad N \in \mathbb{N}_0$$

where the empty product denotes a space consisting of one single trivial "function".

The edge cost is most easily and quite reasonably defined as a weighted sum of cost functionals each of which is associated with an individual edge segment. We will next discuss some possible choices of such edge segment costs. We begin with costs for general parametrized curves. Thereafter a couple of additional possibilities for spline curves will be considered. Finally we give some examples of edge costs for the entire discontinuity set.

General Parametrized Curve Costs

The simplest possible edge segment cost is a constant *presence cost.* If the constant is the same for all the edge segments, this just results in an edge cost term which is proportional to the number of edge segments.

The next simplest measure of the extent of an edge segment is given by its length. We therefore define the *arc length cost* functional

$$\Lambda : C^1(\Sigma)^2 \to \overline{\mathbb{R}_+} : f \mapsto \int_{f(\Sigma)} d\ell = \int_\Sigma \|\dot{f}(\sigma)\|\, d\sigma \tag{2.9}$$

The variational methods to be used in the next section will require that the integrand $\|\dot{f}(\sigma)\|$ is continuously differentiable with respect to $\dot{f}(\sigma)$ for all $\sigma \in \Sigma$. This is clearly not the case if $0 \in \dot{f}(\Sigma)$. We will, however, only consider regular edge segment parametrizations whence this problem disappears.

For a simple measure of the curvedness of an edge segment we likewise define the *curvedness cost*

$$\mathrm{K} : C_R^2(\Sigma)^2 \to \overline{\mathbb{R}_+} : f \mapsto \int_{f(\Sigma)} \left(\frac{d\theta_f}{d\ell}\right)^2 d\ell = \int_\Sigma \frac{[\dot{f}(\sigma)^T R_\times \ddot{f}(\sigma)]^2}{\|\dot{f}(\sigma)\|^5}\, d\sigma \tag{2.10}$$

where θ_f denotes the tangent orientation angle of the edge segment $f(\Sigma)$, ℓ as before denotes the arc length variable, and the rotation matrix

$$R_\times \doteq \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{2.11}$$

plays the role of the "cross product". (It will later also be used for general rotation purposes.) The choice of $(d\theta_f/d\ell)^2$ as opposed to any other power of $|d\theta_f/d\ell|$ in the integrand is motivated by simplicity as well as variational considerations. It might appear simpler to use the first power, but the resulting integrand

$$\frac{|\dot{f}(\sigma)^T R_\times \ddot{f}(\sigma)|}{\|\dot{f}(\sigma)\|^2}$$

(in the corresponding integral over Σ) fails to be continuously differentiable with respect to $\dot{f}(\sigma)$ and $\ddot{f}(\sigma)$. The problem occurs wherever $\dot{f}(\sigma)^T R_\times \ddot{f}(\sigma) = 0$, which in particular includes all points of zero curvature. Since edge segments possessing such points are common, one cannot, unlike in the case of the arc length cost above, get around this problem by simple restriction of the parametrization space.

Besides desiring discontinuity sets with few edge segments, short total arc length and moderate curvedness, one might have a preference for certain curve shapes over others regardless of curve size. None of the edge segment costs above serves to promote such a preference. The constant cost is obviously independent of any property of a particular edge segment except its existence, and both the arc length and curvedness costs depend on the size of the edge segment in question. Indeed, if $x \in \mathbf{R}^2$, $r > 0$ and f is an edge segment parametrization, then

$$\Lambda(x + rf) = r\Lambda(f) \tag{2.12}$$

and

$$\mathrm{K}(x + rf) = \frac{1}{r}\mathrm{K}(f) \tag{2.13}$$

One way of obtaining a size invariant edge segment cost, which discourages unnecessary complexity of the edges, is to modify the curvedness cost K by replacing the arc length variable ℓ by the parameter σ as independent variable in its defining integral. This modification yields the functional

$$f \mapsto \int_\Sigma \left(\frac{d\theta_f}{d\sigma}\right)^2 d\sigma = \int_\Sigma \frac{[\dot{f}(\sigma)^T R_\times \ddot{f}(\sigma)]^2}{\|\dot{f}(\sigma)\|^4}\, d\sigma$$

A problem with this cost, however, is that it turns out to depend on the particular parametrization f of the edge segment $f(\Sigma)$. This problem can be remedied by selecting a unique canonical parametrization for each edge segment. The only choice that makes sense (from a symmetry point of view) for a given fixed compact parameter interval Σ is the *constant*

velocity parametrization characterized by

$$\|\dot{f}(\sigma)\| = \frac{\Lambda(f)}{m(\Sigma)} \qquad \forall \sigma \in \Sigma \tag{2.14}$$

where m denotes the Lebesgue measure. One then ends up with the *shape cost*

$$C_R^2(\Sigma)^2 \to \overline{\mathbf{R}_+} : f \mapsto \mathrm{K}(f)\Lambda(f) \tag{2.15}$$

Spline Curve Costs

For edge segments represented by splines, edge segment costs defined directly in terms of their associated control vertex sequences offer simple alternatives to the general parametrized curve costs discussed above. (Since each edge segment $\gamma_n([0, M_n])$, $n = 1, \ldots, N$, corresponds to a unique control vertex sequence, the previous definition of the edge cost $\mathcal{E}_N$ as a functional of the image segmentation is strictly speaking still valid even if it is expressed in terms of the associated control vertices.)

The length of a spline curve can for example be approximated by the length of its defining control polygon. We therefore define the *polygon length cost*

$$\Pi : \bigcup_{M=3}^{\infty} \mathbf{R}^{2(M+3)} \to \overline{\mathbf{R}_+} : \langle v_m \rangle_{m=0}^{M+2} \mapsto \sum_{m=2o}^{M-1} \|v_{m+1} - v_m\| \tag{2.16}$$

where o as before denotes the openness of the spline. (As we see, the repeated control vertices do not contribute to the cost.)

Similarly extensively curved shapes of the edge segments can be prevented by penalizing the jaggedness of the control polygons. If consecutive control vertices are constrained to be separated, except when constrained to coincide by the spline end conditions, a simple *jaggedness cost* $\Theta : \bigcup_{M=3}^{\infty} \mathbf{R}^{2(M+3)} \to \overline{\mathbf{R}_+}$ can for example be defined by

$$\Theta(v_0, \ldots, v_{M+2}) \doteq \sum_{m=1+2o}^{M-o} \frac{(v_{m+1} - v_m)^T (v_{m-1} - v_m)}{\|v_{m+1} - v_m\| \|v_{m-1} - v_m\|} \tag{2.17}$$

Edge Cost Examples

Of the many possible edge costs that can be composed as weighted sums of the various edge segment costs presented above we will henceforth consider only a few practical examples. In particular, with the exception of the constant presence cost we will not mix the general parametrized curve costs with the spline curve costs. In the absence of more

specific information about the individual edge segments we will moreover choose weighting coefficients which are uniform with respect to the edge segments.

For edges represented by general parametrized curves we will consider the edge cost

$$\mathcal{E}_N(\gamma) \doteq \sum_{n=1}^{N}[\nu + \lambda\Lambda(\gamma_n) + \varkappa\mathrm{K}(\gamma_n) + \iota\dot{\mathrm{K}}(\gamma_n)\Lambda(\gamma_n)] \qquad N \in \mathbb{N}_0 \tag{2.18}$$

where $\nu, \lambda, \kappa, \iota \geq 0$ and $\nu + \lambda > 0$. The interesting special case

$$\mathcal{E}_N(\gamma) \doteq \sum_{n=1}^{N}[\nu + \lambda\Lambda(\gamma_n)] \qquad N \in \mathbb{N}_0 \tag{2.19}$$

corresponding to $\kappa = \iota = 0$ will also be given some extra attention.

For edges represented by splines we will only consider one basic edge cost which we shall define as a function directly in terms of the control vertex sequences $\langle v_{nm}\rangle_{m=0}^{M_n+2}$, $n = 1, \ldots, N$. Thus we choose

$$\mathcal{E}_N(\gamma) \doteq \sum_{n=1}^{N}[\nu + \varpi\Pi(v_{n0}, \ldots, v_{n,M_n+2})] \qquad N \in \mathbb{N}_0 \tag{2.20}$$

where $\nu, \varpi \geq 0$ and $\nu + \varpi > 0$.

2.3 Variations

Having regularized the the global edge detection problem by posing it as a cost functional minimization problem, the next natural step is to seek conditions for optimality. In other words we want to see if the solution, that is the detected image segmentation and the estimated image function that minimize the total cost functional $\mathcal{E}_N(\gamma)+\mathcal{D}_{C_\gamma}(z,\zeta)+\mathcal{S}_{C_\gamma}(z)$, (for a given original image function,) satisfy some manageable set of equations. A solution or at least part of it can then be found by solving these equations. A standard method for deriving such conditions for local minima of the kind of cost functional that we are concerned with is to use calculus of variations.

Before getting into the details of calculations of the sort, we should note that calculus of variations typically yields conditions only for local minima in the *interior* of the domain of the functional in question. In order to get the most out of a variational method the problem should therefore as far as possible be formulated so that at least the vast majority of the local minima of interest are located in the interior of this domain. For continuity control function-represented edge detection this is an important issue. Although

it can be dealt with successfully, it has so far been overlooked by the variational edge detection approaches in the vision literature. (This is the main motivation behind the biased anisotropic diffusion method to be presented in part 5.) For curve-represented edge detection, which is our present concern, the consequences are much milder, but also harder to get around. Since $\mathbb{N}_0$ has empty interior, there is no hope of finding the optimal number of edge segments directly by calculus of variations. For similar reasons the variational methods do not provide direct means for selecting an optimal *edge segment interconnection*, that is the set of interconnection constraints each of which forces a set of edge segment endpoints to coincide at a junction. However, for a given interconnection of $N \in \mathbb{N}_0$ edge segments, calculus of variations yields, (at least for sufficiently nice edges,) necessary conditions for both the optimal (N-segment) image segmentation and the optimal estimated image function. As for the majority of the global edge detection methods discussed in chapter 1 these conditions determine for each given image segmentation the associated estimated image function as the solution of a partial differential equation. While they do not support a similar technique for finding the optimal image segmentation, they provide valuable directions for a structured search. Calculus of variations is thus useful for finding the optimal image segmentation and the optimal estimated image function for each given number of edge segments and each edge segmentation interconnection of interest. One can then select the best among as many such solution candidates as one can afford to compute by simply comparing their associated costs.

In the next three subsections we will derive expressions for the variations of most of the cost functionals presented in section 2.2 with respect to the image segmentation γ, the control vertices v_{nm}, $m = 0, \ldots, M_n$, $n = 1, \ldots, N$, and the estimated image function z. In the following section we will then use these expressions to establish the necessary conditions for optimality. At this point we are mainly interested in what these conditions are, and not in under exactly which additional conditions they are valid. We will therefore make the simplifying assumption that the edges to be detected and the image function to be estimated are sufficiently well behaved, (smooth and integrable,) that we can differentiate and integrate by parts as many times as needed, and that the variation of the image cost $\delta_\gamma[\mathcal{D}_{C_\gamma}(z,\zeta)+\mathcal{S}_{C_\gamma}(z)]$ due to a displacement $\delta\gamma$ of the edges can be evaluated by integrating the difference between the local *image cost density*, that is the integrand of the image cost, on each side of the edge times the normal component of the local edge displacement along the edges.

2.3.1 Variation with Respect to the Image Segmentation

Since all the cost functionals that depend on the edges only depend on the edge segments themselves and not on their particular parametrizations, the same must be the case with the variation due to a certain displacement of the edges. For the purpose of deriving the optimality conditions we can therefore assume any convenient parametrization of the edge segments. The simplest choice that complies with our earlier definitions is the constant velocity parametrization characterized by (2.14). This parametrization will be used throughout this subsection as well as in the sequel wherever reference is made to the results derived herein. Since we have already assumed that $\Sigma = [0,1]$, the constant "velocity" of the nth edge segment parametrization is given by

$$\|\dot{\gamma}_n(\sigma)\| = \Lambda_n \qquad \sigma \in \Sigma, \qquad n = 1, \ldots, N$$

where $\Lambda_n \doteq \Lambda(\gamma_n)$ denotes the (total) arc length of the nth edge segment.

The variation (of any of the cost functionals) with respect to the image segmentation will not surprisingly be expressed as (a sum of) line integrals along the edge segments. For more convenient and intuitive notation we therefore define the tangential and normal unit vectors $e_{\tau n}$ and $e_{\nu n}$ respectively of the nth edge segment $\gamma_n(\Sigma)$, $(n = 1, \ldots, N,)$ according to

$$\left.\begin{aligned} e_{\tau n}(\sigma) &\doteq \frac{\dot{\gamma}_n(\sigma)}{\Lambda_n} \\ e_{\nu n}(\sigma) &\doteq \frac{R_x^T \dot{\gamma}_n(\sigma)}{\Lambda_n} \end{aligned}\right\} \qquad \sigma \in \Sigma \tag{2.21}$$

where R_x is the 90° clockwise rotation matrix defined in (2.11). The orientation of these vectors relative to the nth edge segment and its tangent vector is depicted in figure 2.3, and their first and second derivatives (with respect to σ) are easily found to be given by

$$\dot{e}_{\tau n} = e_{\nu n}\Lambda_n\kappa_n \tag{2.22a}$$

$$\dot{e}_{\nu n} = -e_{\tau n}\Lambda_n\kappa_n \tag{2.22b}$$

$$\ddot{e}_{\tau n} = -e_{\tau n}\Lambda_n^2\kappa_n^2 + e_{\nu n}\Lambda_n\dot{\kappa}_n \tag{2.22c}$$

$$\ddot{e}_{\nu n} = -e_{\tau n}\Lambda_n\dot{\kappa}_n - e_{\nu n}\Lambda_n^2\kappa_n^2 \tag{2.22d}$$

where

$$\kappa_n \doteq \frac{\dot{\gamma}_n^T R_x \ddot{\gamma}_n}{\Lambda_n^3}$$

is the curvature of $\gamma_n(\Sigma)$.

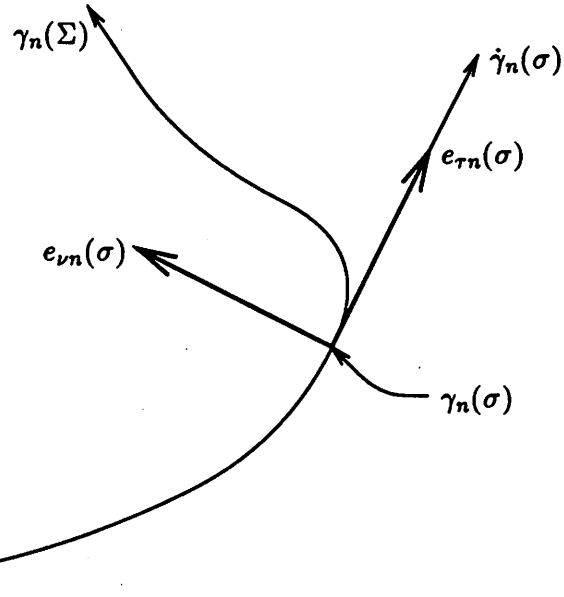


Figure 2.3: Orientation of the tangential and normal unit vectors $e_{\tau n}(\sigma)$ and $e_{\nu n}(\sigma)$ relative to the edge segment $\gamma_n(\Sigma)$ and its tangent vector $\dot{\gamma}_n(\sigma)$.

The subcosts, of which the variation with respect to the image segmentation is of primary interest, are the arc length cost $\Lambda(\gamma_n)$, the curvedness cost $\mathrm{K}(\gamma_n)$, the shape cost $\mathrm{K}(\gamma_n)\Lambda(\gamma_n)$ and the image cost $\mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)$.

Length Cost Variation

The variation of the arc length cost (2.9) with respect to the image segmentation is given by

$$\delta_\gamma \Lambda(\gamma_n) = \delta_\gamma \int_0^1 \sqrt{\dot{\gamma}_n^T \dot{\gamma}_n}\, d\sigma = \int_0^1 \frac{\dot{\gamma}_n^T \delta\dot{\gamma}_n}{\Lambda_n}\, d\sigma$$

Integrating this expression by parts and using (2.21) and (2.22) we obtain

$$\delta_\gamma \Lambda(\gamma_n) = \delta\gamma_{n\tau}|_{\sigma=0}^{1} - \Lambda_n \int_0^1 \kappa_n \delta\gamma_{n\nu}\, d\sigma = \delta\gamma_{n\tau}|_{\ell=0}^{\Lambda_n} - \int_{\gamma_n(\Sigma)} \kappa_n \delta\gamma_{n\nu}\, d\ell \qquad (2.23)$$

where $\delta\gamma_{n\tau} \doteq e_{\tau n}^T \delta\gamma_n$ and $\delta\gamma_{n\nu} \doteq e_{\nu n}^T \delta\gamma_n$ are the tangential and normal components respectively of the variation in the edge parametrization γ_n. This result is very intuitive; in order to lower the arc length cost the edge segment $\gamma_n(\Sigma)$ should be adjusted as indicated in figure 2.4.

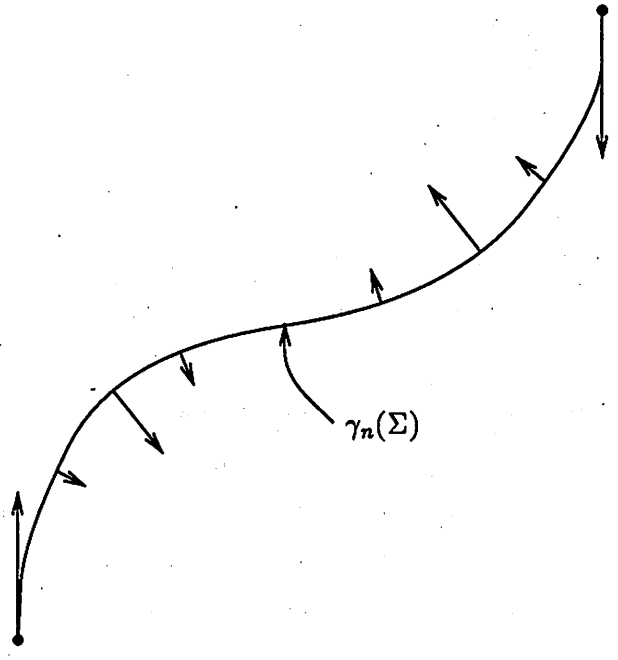


Figure 2.4: Appropriate adjustment of edge segment $\gamma_n(\Sigma)$ for lowering the arc length cost $\Lambda(\gamma_n)$.

Curvedness Cost Variation

For the curvedness cost (2.10) a calculation similar to that above yields

$$\begin{aligned}
&\delta_\gamma \mathrm{K}(\gamma_n) = \\
&\quad = \delta_\gamma \int_0^1 \left(\kappa_n^2 \sqrt{\dot\gamma_n^T \dot\gamma_n}\right) d\sigma \\
&\quad = \int_0^1 \left[\frac{2\kappa_n(\delta\dot\gamma_n^T R_\times \ddot\gamma_n + \dot\gamma_n^T R_\times \delta\ddot\gamma_n)}{\Lambda_n^2} - \frac{5\kappa_n^2 \dot\gamma_n^T \delta\dot\gamma_n}{\Lambda_n}\right] d\sigma \\
&\quad = \left(-\kappa_n^2 \delta\gamma_{n\tau} - \frac{2\dot\kappa_n}{\Lambda_n}\delta\gamma_{n\nu} + \frac{2\kappa_n}{\Lambda_n}\delta\dot\gamma_{n\nu}\right)\Big|_{\sigma=0}^{1} + \int_0^1 \left(\Lambda_n\kappa_n^3 + \frac{2\ddot\kappa_n}{\Lambda_n}\right)\delta\gamma_{n\nu}\, d\sigma \\
&\quad = \left(-\kappa_n^2 \delta\gamma_{n\tau} - 2\frac{d\kappa_n}{d\ell}\delta\gamma_{n\nu} + 2\kappa_n \delta\frac{d\gamma_{n\nu}}{d\ell}\right)\Big|_{\ell=0}^{\Lambda_n} \\
&\qquad + \int_{\gamma_n(\Sigma)} \left(\kappa_n^3 + 2\frac{d^2\kappa_n}{d\ell^2}\right)\delta\gamma_{n\nu}\, d\ell
\end{aligned} \tag{2.24}$$

In this case the intuition is not as immediate as for the arc length cost, and we will in fact only attempt an interpretation of four of the five terms.

As we see from (2.13) the curvedness cost decreases if the edge segment is magnified. The first of the three "endpoint terms" and the first term in the integrand above clearly correspond to this effect. Indeed, in order to lower the cost associated with these terms the edge segment should be expanded as indicated in figure 2.5 (a). Naturally the response to such an adjustment is strongest where the magnitude of the curvature is the largest.

The last of the "endpoint terms" and the second term in the integrand are more related to the shape than the size of the edge segment. The former decreases if the edge segment is "straightened out" near its endpoints. The cost due to the latter decreases if the curvature is redistributed more evenly along the edge. An appropriate adjustment for lowering the cost associated with these terms is depicted in figure 2.5 (b).

Shape Cost Variation

The variation of the shape cost (2.15) with respect to the image segmentation γ, which is easily expressed in terms of those of the arc length and curvedness costs calculated above, is given by

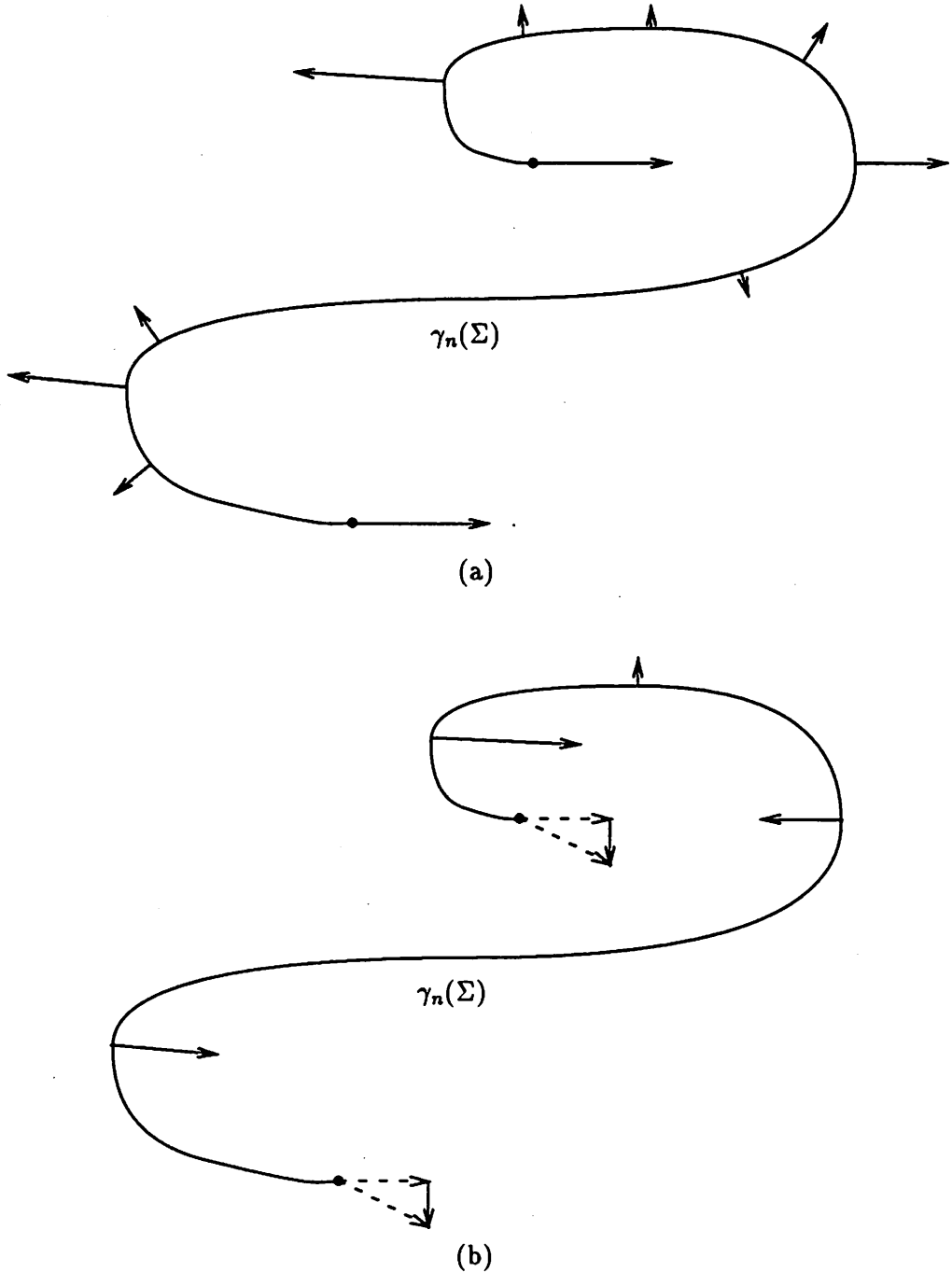


Figure 2.5: Appropriate adjustment of the edge segment $\gamma_n(\Sigma)$ for lowering the curvedness cost $K(\gamma_n)$ associated with the "expansion terms" (a) and the "shape terms" (b).

$$\begin{aligned}
&\delta_\gamma[\mathrm{K}(\gamma_n)\Lambda(\gamma_n)] = \\
&\quad = \delta_\gamma \mathrm{K}(\gamma_n)\Lambda(\gamma_n) + \mathrm{K}(\gamma_n)\delta_\gamma\Lambda(\gamma_n) \\
&\quad = \left[\left(\kappa_n^2\delta\gamma_{n\tau} - 2\frac{d\kappa_n}{d\ell}\delta\gamma_{n\nu} + 2\kappa_n\delta\frac{d\gamma_{n\nu}}{d\ell}\right)\Big|_{\ell=0}^{\Lambda_n} + \int_{\gamma_n(\Sigma)}\left(\kappa_n^3 + 2\frac{d^2\kappa_n}{d\ell^2}\right)\delta\gamma_{n\nu}\,d\ell\right]\Lambda_n \\
&\qquad + \mathrm{K}(\gamma_n)\left(\delta\gamma_{n\tau}|_{\ell=0}^{\Lambda_n} - \int_{\gamma_n(\Sigma)}\kappa_n\delta\gamma_{n\nu}\,d\ell\right)
\end{aligned}$$

The appropriate edge segment adjustment for reducing the shape cost is thus just a weighted combination of the corresponding adjustments for the arc length and curvedness costs. Because of the special weights, Λ_n and $\mathrm{K}(\gamma_n)$, however, there is a more geometric way to interpret this adjustment as a modification of the appropriate adjustment for the curvedness cost. Indeed, since the average square curvature of the nth edge segment is given by

$$\overline{\kappa_n^2} \doteq \frac{\int_{\gamma_n(\Sigma)}\kappa_n^2\,d\ell}{\int_{\gamma_n(\Sigma)}d\ell} = \frac{\mathrm{K}(\gamma_n)}{\Lambda_n}$$

the variation of the shape cost with respect to γ can be written as

$$\begin{aligned}
&\delta_\gamma[\mathrm{K}(\gamma_n)\Lambda(\gamma_n)] = \\
&\quad = \left(\left[-(\kappa_n^2 - \overline{\kappa_n^2})\delta\gamma_{n\tau} - 2\frac{d\kappa_n}{d\ell}\delta\gamma_{n\nu} + 2\kappa_n\delta\frac{d\gamma_{n\nu}}{d\ell}\right]\Big|_{\ell=0}^{\Lambda_n}\right. \\
&\qquad \left. + \int_{\gamma_n(\Sigma)}\left[(\kappa_n^2 - \overline{\kappa_n^2})\kappa_n + 2\frac{d^2\kappa_n}{d\ell^2}\right]\delta\gamma_{n\nu}\,d\ell\right)\Lambda_n
\end{aligned} \tag{2.25}$$

Comparing this expression with (2.24) we notice two and only two differences. First of all the entire expression is multiplied by the arc length of the edge segment. Although this has no effect on the "direction" of an appropriate edge segment adjustment for lowering the cost, for a fixed step size steepest descent scheme it would affect the size of the update so as to make it proportional to the size of the edge segment. This is perfectly reasonable for a size independent edge cost. Secondly the two "expansion terms" suggesting the adjustment in figure 2.5 (a) are moderated by the square curvature being measured relative to its average value. More precisely, κ_n^2 is replaced by $\kappa_n^2 - \overline{\kappa_n^2}$. The effect that this has on the appropriate adjustment for reducing the cost associated with these two terms is depicted in figure 2.6.

Image Cost Variation

When the image segmentation is displaced from γ to $\gamma + \delta\gamma$, the current continuity set C_γ is replaced by a new continuity set $C_{\gamma+\delta\gamma}$. The "difference set" (in $\mathbf{R}^2$) that the

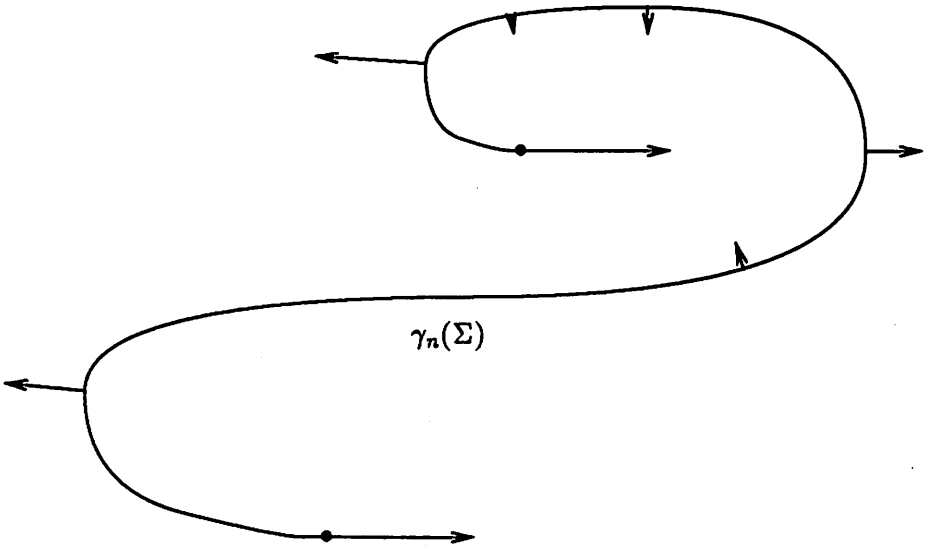


Figure 2.6: Appropriate adjustment of edge segment $\gamma_n(\Sigma)$ for lowering the shape cost $\mathrm{K}(\gamma_n)\Lambda(\gamma_n)$ associated with the "expansion terms".

discontinuity set, that is the edge segments, would have swept over, should this displacement have taken place in a continuous fashion, will by a minor abuse of notation be denoted by δC_γ. An example of such a set is displayed in figure 2.7. Since the continuity set is the domain of the estimated image function z, the perturbation $\delta\gamma$ necessarily affects the image cost density

$$\varrho \doteq (z-\zeta)^2 + \sum_{i=0}^{I} \mu_i \sum_{k_1=1}^{2} \cdots \sum_{k_i=1}^{2} \left(\frac{\partial^i z}{\partial x_{k_1} \cdots \partial x_{k_i}} \right)^2$$

and thus the image cost

$$\mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z) = \int_{C_\gamma} \varrho \, dx$$

If the restriction $z_{C_\gamma}|C_\gamma \setminus \delta C_\gamma$ of the estimated image function $z_{C_\gamma} :\rightarrow \mathbf{R}$ could be extended to a new (admissible) estimated image function $z_{C_{\gamma+\delta\gamma}} :\rightarrow \mathbf{R}$ on the new continuity set, the resulting image cost difference would be given by

$$\mathcal{D}_{C_{\gamma+\delta\gamma}}(z_{C_{\gamma+\delta\gamma}},\zeta) + \mathcal{S}_{C_{\gamma+\delta\gamma}}(z_{C_{\gamma+\delta\gamma}}) - [\mathcal{D}_{C_\gamma}(z_{C_\gamma},\zeta) + \mathcal{S}_{C_\gamma}(z_{C_\gamma})] =$$
$$= \int_{\delta C_\gamma} \left((z_{C_{\gamma+\delta\gamma}} - \zeta)^2 + \sum_{i=0}^{} \mu_i \sum_{k_1=1}^{2} \cdots \sum_{k_i=1}^{2} \left(\frac{\partial^i z_{C_{\gamma+\delta\gamma}}}{\partial x_{k_1} \cdots \partial x_{k_i}} \right)^2 -$$

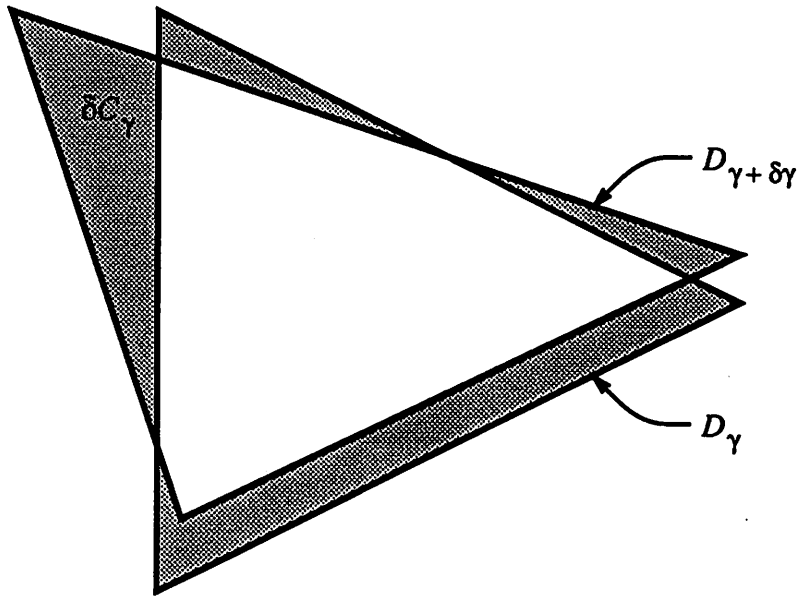


Figure 2.7: A discontinuity set D_γ, a slightly altered discontinuity set $D_{\gamma+\delta\gamma}$ and the resulting "difference set" δC_γ (shaded).

$$- \left[(z_{C_\gamma} - \zeta)^2 + \sum_{i=0} \mu_i \sum_{k_1=1}^{2} \cdots \sum_{k_i=1}^{2} \left(\frac{\partial^i z_{C_\gamma}}{\partial x_{k_1} \cdots \partial x_{k_i}} \right)^2 \right] \Bigg) \, dx$$

We will study such extensions in detail in chapter 3. For now we will just assume that such an extension exists, and that it is sufficiently well behaved that the integral above for an infinitesimal displacement $\delta\gamma$ reduces to a (sum of) line integrals along the edge segments. The variation of the image cost with respect to the image segmentation is then given by

$$\delta_\gamma[\mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)] = -\sum_{n=1}^{N} \int_{\gamma_n(\Sigma)} \blacktriangle\varrho_n \delta\gamma_{n\nu} \, d\ell \tag{2.26}$$

where

$$\blacktriangle\varrho_n \doteq \lim_{h \downarrow 0} [\varrho(\gamma_n(\sigma) + e_{\nu n}(\sigma)h) - \varrho(\gamma_n(\sigma) - e_{\nu n}(\sigma)h)]$$

is the *image cost density difference* across the nth edge segment. In order to lower the image cost the edge segments should thus be shifted towards the side on which the image cost density is the highest. Intuitively enough, the area of relatively low image cost density is thereby enlarged at the expense of the area of relatively high image cost density.

2.3.2 Variation with Respect to the Control Vertices

If the edges are represented by splines, as suggested in section 2.1, the solution to the edge detection problem essentially consists of finding the control vertices that specify the optimal image segmentation. In this context it is desirable that the optimality conditions are expressed directly in terms of the control vertices rather than the spline curves that they define. The variations of both the edge and the image costs with respect to the control vertices are therefore of interest. In the case of spline represented edges an edge cost composed by the special spline costs, discussed in section 2.2, is to be preferred. We will illustrate the calculations by considering the polygon length cost (2.16) in detail. The jaggedness cost (2.17) can be treated in a similar fashion.

Polygon Length Cost Variation

The variation of the polygon length cost with respect to the control vertices exists only at those points in control vertex space where the polygon length cost is differentiable with respect to all its independent[§] control vertices. This means that all independent consecutive control vertices must be distinct. Assuming that this "regularity" condition is satisfied the variation of the polygon length cost with respect to the (independent) control vertices is given by

$$\begin{aligned}\delta_v \Pi(v_0, \ldots, v_{M+2}) &= \\ &= \delta_v \sum_{m=2o}^{M-1} \|v_{m+1} - v_m\| \\ &= \sum_{m=2o}^{M-1} \left(\frac{v_{m+1} - v_m}{\|v_{m+1} - v_m\|} \right)^T (\delta v_{m+1} - \delta v_m)\end{aligned} \tag{2.27}$$

where o is the openness of the spline. Collecting the terms depending on the perturbation of each individual control vertex, for closed splines we obtain

$$\delta_v \Pi(v_0, \ldots, v_{M+2}) = \sum_{m=1}^{M} \left(\frac{v_m - v_{m-1}}{\|v_m - v_{m-1}\|} + \frac{v_m - v_{m+1}}{\|v_m - v_{m+1}\|} \right)^T \delta v_m \tag{2.28}$$

where we have used the fact that $v_M = v_0$. For open splines we similarly get

[§]Two control vertices are said to be independent if they are neither constrained by the spline end conditions nor by the interconnection constraints.

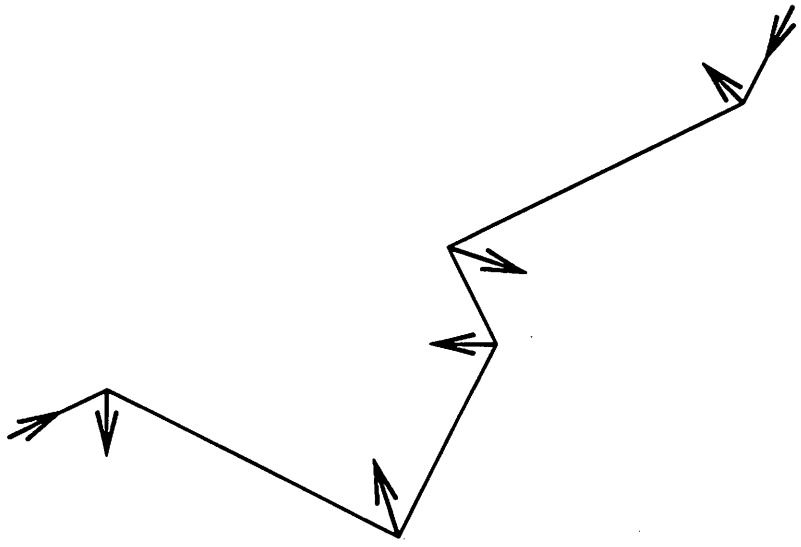


Figure 2.8: Appropriate adjustment of the control vertices for lowering the polygon length cost $\Pi(v_0, \ldots, v_{10})$.

$$\begin{aligned}\delta_v \Pi(v_0, \ldots, v_{M+2}) &= \\ &= \left(\frac{v_2 - v_3}{\|v_2 - v_3\|}\right)^T \delta v_2 + \sum_{m=3}^{M-1} \left(\frac{v_m - v_{m-1}}{\|v_m - v_{m-1}\|} + \frac{v_m - v_{m+1}}{\|v_m - v_{m+1}\|}\right)^T \delta v_m \\ &\quad + \left(\frac{v_M - v_{M-1}}{\|v_M - v_{M-1}\|}\right)^T \delta v_M \end{aligned} \tag{2.29}$$

The expressions above have a very simple interpretation. In order to decrease the polygon length cost the control polygon should be adjusted as indicated in figure 2.8. In other words, intermediate vertices should move along the bisectors of the angles formed by the polygon, and end vertices should move along the polygon itself.

Image Cost Variation

For the image cost we need to reexpress the previously derived variation with respect to the image segmentation in terms of the control vertices. Recalling that the domain of the edge segment parametrization γ_n now is $\Sigma_n \doteq [0, M_n]$ from (2.4) we have that

$$\begin{aligned}
&\int_{\gamma_n(\Sigma_n)} \blacktriangle\varrho_n \delta\gamma_{n\nu}\, d\ell = \\
&\quad = \int_0^{M_n} \blacktriangle\varrho_n \delta\gamma_n^T R_\times^T \dot\gamma_n\, d\sigma \\
&\quad = \sum_{m=0}^{M_n-1} \int_0^1 \blacktriangle\varrho_n(m+\sigma)\delta\gamma_{nm}(\sigma)^T R_\times^T \dot\gamma_{nm}(\sigma)\, d\sigma \\
&\quad = \sum_{m=0}^{M_n-1} \int_0^1 \blacktriangle\varrho_n(m+\sigma) \sum_{r=0}^{3} b_r(\sigma)\delta v_{n,m+r}^T R_\times^T \sum_{s=0}^{3} v_{n,m+s}\dot b_s(\sigma)\, d\sigma \\
&\quad = \int_0^1 \sum_{s=0}^{3} d_n(m,r,s,\sigma)\dot b_s(\sigma)\, d\sigma \qquad n = 1,\ldots,N
\end{aligned} \tag{2.30}$$

where $b_0,\ldots,b_3$ are the basis functions given by (2.3), and

$$d_n(m,r,s,\sigma) \doteq \sum_{r=0}^{3}\sum_{m=0}^{M_n-1} \blacktriangle\varrho_n(m+\sigma) b_r(\sigma)\delta v_{n,m+r}^T R_\times^T v_{n,m+s} \tag{2.31}$$

In order to collect the terms associated with each individual control vertex we substitute m for $m+r$ and change the order of summation in (2.31). Dropping the subscript n and using the fact that $M \geq 3$ we then get

$$\begin{aligned}
&d(m,r,s,\sigma) = \\
&\quad = \sum_{m=0}^{2}\sum_{r=0}^{m} \blacktriangle\varrho(m-r+\sigma) b_r(\sigma)\delta v_m^T R_\times^T v_{m-r+s} \\
&\qquad + \sum_{m=3}^{M-1}\sum_{r=0}^{3} \blacktriangle\varrho(m-r+\sigma) b_r(\sigma)\delta v_m^T R_\times^T v_{m-r+s} \\
&\qquad + \sum_{m=M}^{M+2}\sum_{r=m-M+1}^{3} \blacktriangle\varrho(m-r+\sigma) b_r(\sigma)\delta v_m^T R_\times^T v_{m-r+s}
\end{aligned}$$

For closed splines $v_m = v_{M+m}$, $m = 0,1,2$. Since $m-r \in \{0,\ldots,M-1\}$ in each of the double sums above, we therefore have that

$$d(m,r,s,\sigma) = \sum_{m=1}^{M}\sum_{r=0}^{3} \blacktriangle\varrho((m-r) \bmod M + \sigma) b_r(\sigma)\delta v_m^T R_\times^T v_{(m-r) \bmod M+s} \tag{2.32}$$

For open splines on the other hand $v_0 = v_1 = v_2$ and $v_M = v_{M+1} = v_{M+2}$. Hence

$$\begin{aligned}
&d(m,r,s,\sigma) = \\
&\quad = \sum_{m=0}^{2}\sum_{r=0}^{m} \blacktriangle\varrho(m-r+\sigma) b_r(\sigma)\delta v_2^T R_\times^T v_{m-r+s}
\end{aligned}$$

$$+\sum_{m=3}^{M-1}\sum_{r=0}^{3}\blacktriangle\varrho(m-r+\sigma)b_r(\sigma)\delta v_m^T R_\times^T v_{m-r+s}$$
$$+\sum_{m=M}^{M+2}\sum_{r=m-M+1}^{3}\blacktriangle\varrho(m-r+\sigma)b_r(\sigma)\delta v_M^T R_\times^T v_{m-r+s} \qquad (2.33)$$

Using (2.26), (2.30), (2.32) and (2.33) the image cost variation with respect to the control vertices can now be expressed as

$$\begin{aligned}
&\delta_v[\mathcal{D}_{C_\gamma}(z,\zeta)+\mathcal{S}_{C_\gamma}(z)] = \\
&\quad = -\sum_{n=1}^{N}\sum_{m=0}^{M_n-1}\sum_{r=0}^{3}\sum_{s=0}^{3}\int_0^1\blacktriangle\varrho_n(m+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma v_{n,m+s}^T R_\times\delta v_{n,m+r} \\
&\quad = -\sum_{n\in N_0}\sum_{m=1}^{M_n}\sum_{r=0}^{3}\sum_{s=0}^{3}\int_0^1\blacktriangle\varrho_n((m-r)\bmod M_n+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma \\
&\qquad \cdot v_{n,(m-r)\bmod M_n+s}^T R_\times\delta v_{nm} \\
&\qquad -\sum_{n\in N_1}\Bigg[\sum_{m=0}^{2}\sum_{r=0}^{m}\sum_{s=0}^{3}\int_0^1\blacktriangle\varrho_n(m-r+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma v_{n,m-r+s}^T R_\times\delta v_{n2} \\
&\qquad\quad +\sum_{m=3}^{M_n-1}\sum_{r=0}^{3}\sum_{s=0}^{3}\int_0^1\blacktriangle\varrho_n(m-r+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma v_{n,m-r+s}^T R_\times\delta v_{nm} \\
&\qquad\quad +\sum_{m=M_n}^{M_n+2}\sum_{r=m-M_n+1}^{3}\sum_{s=0}^{3}\int_0^1\blacktriangle\varrho_n(m-r+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma \\
&\qquad\quad \cdot v_{n,m-r+s}^T R_\times\delta v_{nM_n}\Bigg] \qquad (2.34)
\end{aligned}$$

where $N_t \doteq \{n\in\{1,\ldots,N\} : o_n = t\}$, $t = 0,1$, that is N_0 is the set of edge segments indices corresponding to closed splines, and N_1 is the set of those corresponding to open splines. Unfortunately this expression is too complicated to provide much insight. A few general observations can, however, be made:

1. The image cost variation with respect to the control vertex v_{nm} depends in addition to on v_{nm} itself on the three closest vertices in both the directions along the control polygon. (If v_{nm} is close to the end of the polygon, some of these vertices may of course be identical.)

2. The image cost variation with respect to v_{nm} also depends on the image cost density along a local *segment* of the nth edge segment. This local segment is exactly the portion of $\gamma_n(\Sigma_n)$ that is influenced by the control vertices just referred to.

3. The functions

$$a_{rs}(\sigma) \doteq \int_0^\sigma b_r(\varsigma)\dot{b}(\varsigma)\,d\varsigma \qquad r,s \in \{0,\ldots,3\} \tag{2.35}$$

can be expected to play a significant role in the evaluation of the image cost variation.

2.3.3 Variation with Respect to the Estimated Image Function

The edge cost $\mathcal{E}_N(\gamma)$ is as indicated independent of the estimated image function z. Its variation with respect to z is hence equal to zero. We thus only have to consider the image cost. In this case the deviation and stabilizing costs are most conveniently treated separately.

Deviation Cost Variation

The variation of the deviation cost (2.7) with respect to the estimated image function is extremely easy to calculate. Indeed, for any open domain Ω

$$\delta_z \mathcal{D}_\Omega(z,\zeta) = \delta_z \int_\Omega (z-\zeta)^2\,dx = 2\int_\Omega (z-\zeta)\delta z\,dx \tag{2.36}$$

Stabilizing Cost Variation

The calculation of the variation of the stabilizer (2.8) with respect to the estimated image function is more involved. We begin by defining the differential operators

$$\begin{aligned} Q_0(f,g) &\doteq fg \\ Q_i(f,g) &\doteq \sum_{k_1=1}^{2}\cdots\sum_{k_i=1}^{2} \frac{\partial^i f}{\partial x_{k_1}\cdots\partial x_{k_i}} \frac{\partial^i g}{\partial x_{k_1}\cdots\partial x_{k_i}} \qquad i \in \mathbb{N} \end{aligned}$$

where the functions f and g are allowed to be matrix valued, as long as the dimensions match so that the product fg makes sense. We note that the definition of Q_0 is consistent with our earlier notion of sums of the form $\sum_{k_1}\cdots\sum_{k_i}$ (for $i = 0$). It is also consistent with the simple rule

$$Q_i(f,g) = Q_{i-1}(\nabla f, \nabla g^T) \qquad i \in \mathbb{N} \tag{2.37}$$

for scalar valued functions f and g. Letting $(-\Delta)^j$ denote the j times repeated negative Laplace operator $[-(\partial/\partial x_1)^2 - (\partial/\partial x_2)^2]^j$ we then have the following.

Proposition 2.3.1

$$Q_i(f,g) = \nabla \cdot \sum_{j=1}^{i} Q_{j-1}(f, \nabla(-\Delta)^{i-j}g) + f(-\Delta)^i g$$

Proof: For $i = 0$ the sum on the right hand side vanishes, whence the assertion is trivially true. Suppose now the assertion is true for $i = p \in \mathbb{N}_0$. Then by (2.37)

$$\begin{aligned}
&Q_{p+1}(f,g) = \\
&\quad = \sum_{k_1=1}^{2} \cdots \sum_{k_p=1}^{2} \nabla \frac{\partial^p f}{\partial x_{k_1} \cdots \partial x_{k_p}} \nabla \frac{\partial^p f}{\partial x_{k_1} \cdots \partial x_{k_p}}^T \\
&\quad = \sum_{k_1=1}^{2} \cdots \sum_{k_p=1}^{2} \left[\nabla \cdot \left(\frac{\partial^p f}{\partial x_{k_1} \cdots \partial x_{k_p}} \nabla \frac{\partial^p f}{\partial x_{k_1} \cdots \partial x_{k_p}} \right) \right. \\
&\qquad \left. - \frac{\partial^p f}{\partial x_{k_1} \cdots \partial x_{k_p}} \Delta \frac{\partial^p f}{\partial x_{k_1} \cdots \partial x_{k_p}} \right] \\
&\quad = \nabla \cdot Q_p(f, \nabla g) + Q_p(f, -\Delta g) \\
&\quad = \nabla \cdot Q_p(f, \nabla g) + \nabla \cdot \sum_{j=1}^{p} Q_{j-1}(f, \nabla(-\Delta)^{p-j+1} g) + f(-\Delta)^{p+1} g \\
&\quad = \nabla \cdot \sum_{j=1}^{p+1} Q_{j-1}(f, \nabla(-\Delta)^{p+1-j} g) + f(-\Delta)^{p+1} g
\end{aligned}$$

The proposition thence follows by induction. ■

The calculation of $\delta_z \mathcal{S}_\Omega(z)$ is now straight forward. Indeed, noting that $\delta_z Q_i(z,z)$ $= 2Q_i(\delta z, z)$, $\forall i \in \mathbb{N}_0$, for any open domain Ω we have

$$\delta_z \mathcal{S}_\Omega(z) = \delta_z \int_\Omega \sum_{i=0}^{I} \mu_i Q_i(z,z)\, dx = 2 \int_\Omega \sum_{i=0}^{I} \mu_i Q_i(\delta z, z)\, dx$$

By proposition 2.3.1 and Gauss' divergence theorem it thus follows that

$$\begin{aligned}
&\delta_z \mathcal{S}_\Omega(z) = \\
&\quad = 2 \int_\Omega \sum_{i=0}^{I} \mu_i \left[\nabla \cdot \sum_{j=1}^{i} Q_{j-1}(\delta z, \nabla(-\Delta)^{i-j} z) + \delta z (-\Delta)^i z \right] dx \\
&\quad = 2 \int_{\partial\Omega} \sum_{i=1}^{I} \mu_i \sum_{j=1}^{i} Q_{j-1}(\delta z, \nabla(-\Delta)^{i-j} z) e_n \, d\ell + 2 \int_\Omega \sum_{i=0}^{I} \mu_i \delta z (-\Delta)^i z \, dx
\end{aligned}$$

where e_n is the outward normal unit vector and $d\ell$ indicates integration against the arc length measure along the boundary $\partial\Omega$ of Ω. Changing order of summation and applying the chain rule we then obtain

$$\delta_z \mathcal{S}_\Omega(z) = 2 \sum_{j=1}^{I} \int_{\partial\Omega} Q_{j-1} \left(\delta z, \frac{\partial}{\partial e_n} \sum_{i=0}^{I} \mu_i (-\Delta)^{i-j} z \right) d\ell + 2 \int_\Omega \delta z \sum_{i=0}^{I} \mu_i (-\Delta)^i z \, dx$$

where $\partial/\partial e_n$ denotes the directional derivative in the direction of the outward normal unit vector e_n. In order to see what this really means one can expand the Q_{j-1}-expressions in the integral over $\partial\Omega$. The final result then becomes

$$\begin{aligned}\delta_z \mathcal{S}_\Omega(z) = \\ = 2\sum_{j=1}^{I}\sum_{k_1=1}^{2}\cdots\sum_{k_{j-1}=1}^{2}\int_{\partial\Omega}\frac{\partial^{j-1}\delta z}{\partial x_{k_1}\cdots\partial x_{k_{j-1}}}\frac{\partial^j}{\partial e_n \partial x_{k_1}\cdots\partial x_{k_{j-1}}}\sum_{i=j}^{I}\mu_i(-\Delta)^{i-j}z\,d\ell \\ 2\int_\Omega \delta z\sum_{i=0}^{I}\mu_i(-\Delta)^i z\,dx \end{aligned} \tag{2.38}$$

2.4 Optimality Conditions

With the expressions for all the cost variations in the previous section at hand it is relatively straight forward to compile a collection of optimality conditions. One simply has to add up the total cost variation with respect to each independent variable and set the resulting sum identically equal to zero. The only part of the matter that calls for some special attention is the interdependence among some of the edge variables imposed by the interconnection constraints.

The optimality conditions are naturally divided into those concerning the estimated image function and those concerning the edges. The latter category furthermore takes different forms depending on the edge cost as well as on how the edges are being represented. In this section we will present optimality conditions for the estimated image function, for the image segmentation when the edge cost is given by (2.18), and for the control vertices when the edge cost is given by (2.20).

2.4.1 Estimated Image Function Conditions

Adding (2.36) and (2.38) (to the zero edge cost variation $\delta_z \mathcal{E}_N(\gamma)$) we find that the total cost variation with respect to the estimated image function z is given by

$$\begin{aligned}\delta_z[\mathcal{E}_N(\gamma) + \mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)] = \\ = 2\int_{C_\gamma}\delta z\left[z-\zeta+\sum_{i=0}^{I}\mu_i(-\Delta)^i z\right]dx \\ + 2\sum_{j=1}^{I}\sum_{k_1=1}^{2}\cdots\sum_{k_{j-1}=1}^{2}\int_{\partial\Omega}\frac{\partial^{j-1}\delta z}{\partial x_{k_1}\cdots\partial x_{k_{j-1}}}\frac{\partial^j}{\partial e_n\partial x_{k_1}\cdots\partial x_{k_{j-1}}}\sum_{i=j}^{I}\mu_i(-\Delta)^{i-j}z\,d\ell\end{aligned}$$

For optimality this variation has to vanish for all possible perturbations δz. This in turn requires that the estimated image function satisfies the partial differential (Euler) equation

$$z - \zeta + \sum_{i=0}^{I} \mu_i (-\Delta)^i z = 0 \qquad \text{on } C_\gamma \tag{2.39a}$$

with the boundary conditions

$$\frac{\partial^j}{\partial e_n \partial x_{k_1} \cdots \partial x_{k_{j-1}}} \sum_{i=j}^{I} \mu_i (-\Delta)^{i-j} z = 0 \qquad \text{on } \partial C_\gamma \tag{2.39b}$$

Even for quite small values of I the equations above are fairly complicated, and one is therefore led to consider the simplest examples. One interesting such example, to which we will pay most of our attention from now on, is specified by choosing $I = 1$, $\mu_0 = 0$ and $\mu_1 = \mu > 0$. In this case the optimality conditions above reduce to

$$z - \zeta - \mu \Delta z = 0 \qquad \text{on } C_\gamma \tag{2.40a}$$

$$\frac{\partial z}{\partial e_n} = 0 \qquad \text{on } \partial C_\gamma \tag{2.40b}$$

In chapter 3 we will show that this partial differential equation has a unique solution for all open domains. Since the continuity set C_γ, being the intersection of two open sets, is always open, the optimality condition above does thus by itself not put any restriction on the image segmentation γ.

2.4.2 Edge Conditions

Unlike the optimality conditions for the estimated image function, the conditions for the edges depend on the interconnection constraints. To be specific we will assume that these constraints are of the form discussed in section 2.1, that is each edge segment endpoint, except the artificial ones of closed spline curves in case of spline represented edges, has to coincide with one of $J \in \mathbf{N}_0$ junctions $w_1, \ldots, w_J \in \mathbf{R}^2$. However, more elaborate interconnection constraints, for example such involving the tangents of the edge segments at their endpoints, are of course both possible and useful.

For later reference, for $j = 1, \ldots, J$ we define N_{j0} to be the set of those of the edge segment indices $n = 1, \ldots, N$ for which the "beginning" endpoint $\gamma_n(0)$ is forced to coincide with the junction w_j. Similarly we let N_{j1} be the set of those indices for which the "terminating" endpoint, $\gamma_n(1)$ in case of general parametrized curves or $\gamma_n(M_n)$ in case of splines, is forced to coincide with w_j.

Image Segmentation Conditions

Suppose that the edges are represented by general parametrized curves as described in section 2.1, and that the edge cost is of the form

$$\mathcal{E}_N(\gamma) \doteq \sum_{n=1}^{N} [\nu + \lambda \Lambda(\gamma_n) + \varkappa \mathrm{K}(\gamma_n) + \iota \mathrm{K}(\gamma_n)\Lambda(\gamma_n)]$$

where $\nu, \varkappa, \iota \geq 0$ and $\lambda > 0$. From (2.23), (2.24), (2.25) and (2.26) we then have that

$$\begin{aligned}
&\delta_\gamma[\mathcal{E}_N(\gamma) + \mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)] = \\
&\quad = \sum_{n=1}^{N} (\lambda \delta_\gamma \Lambda(\gamma_n) + \varkappa \delta_\gamma \mathrm{K}(\gamma_n) + \iota \delta_\gamma [\mathrm{K}(\gamma_n)\Lambda(\gamma_n)]) + \delta_\gamma[\mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)] \\
&\quad = \sum_{n=1}^{N} \left[\left(\varphi_n \delta\gamma_{n\tau} - \alpha_n \frac{d\kappa_n}{d\ell} \delta\gamma_{n\nu} + \alpha_n \kappa_n \delta \frac{d\gamma_{n\nu}}{d\ell} \right) \Big|_{\ell=0}^{\Lambda_n} \right. \\
&\qquad \left. - \int_{\gamma_n(\Sigma)} \left(\varphi_n \kappa_n - \alpha_n \frac{d^2\kappa_n}{d\ell^2} + \blacktriangle \varrho_n \right) \delta\gamma_{n\nu} \, d\ell \right]
\end{aligned}$$

where

$$\varphi_n(\sigma) \doteq \lambda - \varkappa \kappa_n(\sigma)^2 - \iota \Lambda_n [\kappa_n(\sigma)^2 - \overline{\kappa_n^2}] \qquad \sigma \in \Sigma, \qquad n = 1, \ldots, N \tag{2.41}$$

and

$$\alpha_n \doteq 2(\varkappa + \iota \Lambda_n) \qquad n = 1, \ldots, N$$

After eliminating the dependence between the endpoints $\gamma_n(t)$, $t = 0, 1$, $n = 1, \ldots, N$, by replacing them by the appropriate junctions, the total cost variation with respect to the image segmentation can thus be written as

$$\begin{aligned}
&\delta_\gamma[\mathcal{E}_N(\gamma) + \mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)] = \\
&\quad = \sum_{j=1}^{J} \sum_{t=0}^{1} \sum_{n \in N_{jt}} (-1)^{1-t} \left[\varphi_n(t) e_{\tau n}^T(t) - \alpha_n \frac{d\kappa_n}{d\ell}(t) e_{\nu n}^T(t) \right] \delta w_j + \sum_{n=1}^{N} \alpha_n \kappa_n \delta \frac{d\gamma_{n\nu}}{d\ell} \Big|_{\ell=0}^{\Lambda_n} \\
&\qquad - \sum_{n=1}^{N} \int_{\gamma_n(\Sigma)} \left(\varphi_n \kappa_n - \alpha_n \frac{d^2\kappa_n}{d\ell^2} + \blacktriangle \varrho_n \right) \delta\gamma_{n\nu} \, d\ell
\end{aligned} \tag{2.42}$$

For optimality this expression should vanish for all possible perturbations $\delta\gamma$. The edge segment parametrizations $\gamma_1, \ldots, \gamma_N$ must therefore satisfy the ordinary differential (Euler) equations

$$\varphi_n \kappa_n - \alpha_n \frac{d^2\kappa_n}{d\ell^2} + \blacktriangle \varrho_n = 0 \qquad \text{on } \Sigma, \qquad n = 1, \ldots, N \tag{2.43a}$$

with the boundary conditions

$$\sum_{t=0}^{1}(-1)^t \sum_{n\in N_{jt}} \left[e_{\tau n}(t)\varphi_n(t) - e_{\nu n}(t)\alpha_n \frac{d\kappa_n}{d\ell}(t)\right] = 0 \qquad j = 1,\ldots J \tag{2.43b}$$

$$\alpha_n\kappa_n(t) = 0 \qquad t = 0,1, \qquad n = 1,\ldots,N \tag{2.43c}$$

Nonzero Curvedness and/or Shape Cost. Consider the case when $\varkappa > 0$ and/or $\iota > 0$. Since the edge segment parametrizations are (by (2.9) implicitly) assumed to be regular, it follows that $\Lambda_n > 0$, and hence that $\alpha_n > 0$, $n = 1,\ldots,N$. From (2.43c) we then see that

$$\kappa_n(0) = \kappa_n(1) = 0 \qquad n = 1,\ldots,N$$

whence

$$\varphi_n(0) = \varphi_n(1) = \beta_n \doteq \lambda + \iota \mathrm{K}(\gamma_n) > 0 \qquad n = 1,\ldots,N$$

This results in the following optimality conditions for the image segmentation

$$\varphi_n\kappa_n - \alpha_n \frac{d^2\kappa_n}{d\ell^2} + \blacktriangle\varrho_n = 0 \qquad \text{on } \Sigma, \qquad n = 1,\ldots,N \tag{2.44a}$$

$$\sum_{t=0}^{1}(-1)^t \sum_{n\in N_{jt}} \left[e_{\tau n}(t)\beta_n - e_{\nu n}(t)\alpha_n \frac{d\kappa_n}{d\ell}(t)\right] = 0 \qquad j = 1,\ldots,J \tag{2.44b}$$

$$\kappa_n(0) = \kappa_n(1) = 0 \qquad n = 1,\ldots,N \tag{2.44c}$$

Unfortunately the system (2.44) does not seem to be of any use for finding an optimal image segmentation. There are in fact several factors contributing to the hopelessness of any such attempt. First of all, the "coefficients" α_n, β_n and φ_n in (2.44a) and (2.44b) depend on the values of the functionals Λ and K evaluated at the unknown edge segmentation parametrization γ_n. Secondly, even if $\Lambda(\gamma_n)$ and $\mathrm{K}(\gamma_n)$ were known, the ordinary differential equation (2.44a) would be nonlinear due to the $\kappa_n(\sigma)^2$-terms in (2.41). A third complication is that the image cost density difference $\blacktriangle\varrho_n$ in (2.44a) is defined in terms of the unknown parametrization γ_n. Finally and most importantly $\blacktriangle\varrho_n$ depends also on the estimated image function z which in turn is defined on a different domain for each different image segmentation, and therefore depends on all the edge segment parametrizations $\gamma_1,\ldots,\gamma_N$ in a very subtle and inconvenient manner.

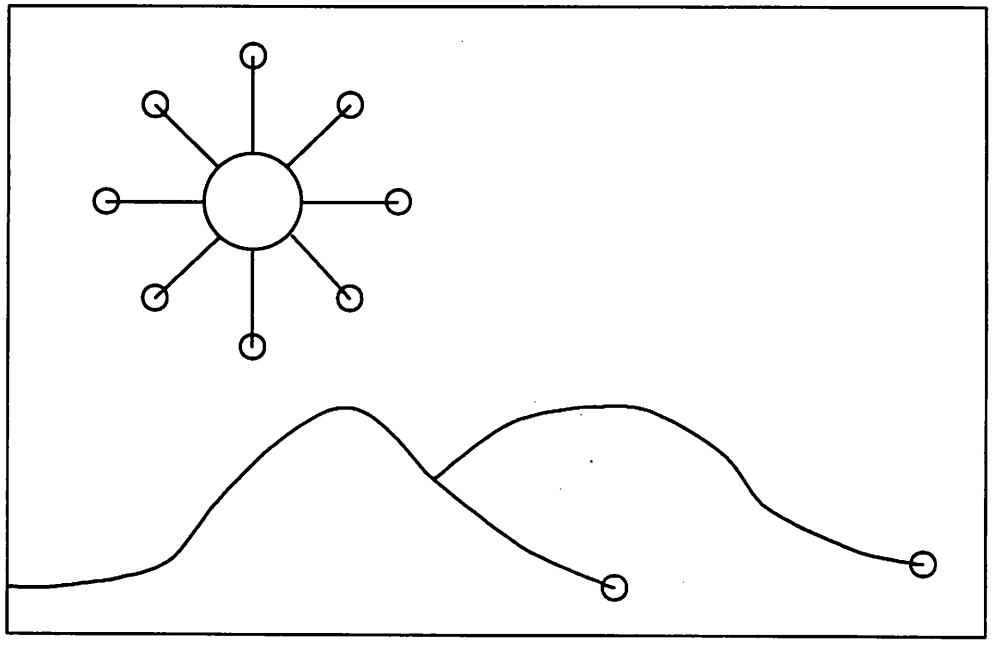


Figure 2.9: Free endpoints (in circles).

While the conditions (2.44) cannot be solved directly for the optimal edges, they do provide immediate information about some of the properties of the edges that will result if the edge detection problem gets solved, regardless of how this is being done. The ordinary differential equations (2.44a) relate the local shape of the optimal edge segments to the local image cost density, and thereby, if yet in a vague sense, to the original image function ζ. The boundary conditions (2.44b) restrict the possible ways in which the optimal edge segments can meet at the junctions. The boundary conditions (2.44c) finally tell us that all the optimal edge segments have zero curvature at their endpoints.

A disturbing but important consequence of (2.44b) is that the optimality conditions cannot be satisfied by any interconnection that allows single endpoint junctions, In other words, edge segments with *free endpoints* as those highlighted in figure 2.9 are ruled out. Indeed, for such a junction the double sum in (2.44b) contains only one term. Since the unit vectors $e_{\tau n}$ and $e_{\nu n}$ are orthogonal, and $\beta_n > 0$, this implies that (2.44b) is violated. A possible "conclusion" would be that interconnections of this kind are always suboptimal, but this goes against all intuition. In fact, if the variation (2.42) were used to compute an appropriate edge segment adjustment for lowering the edge cost, the resulting update

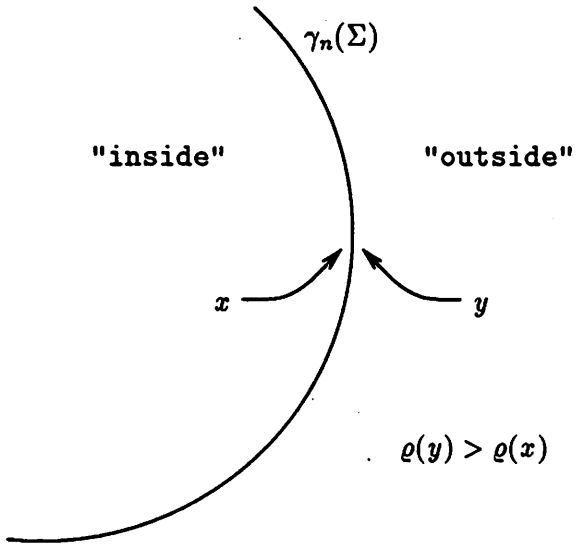


Figure 2.10: Relationship between the local shape of an optimal edge segment $\gamma_n(\Sigma)$ and the local image cost density ϱ.

scheme would keep on shortening a straight edge segment with a free endpoint until it disappeared. Since examples can be contrived for which such a "solution" could impossibly be optimal, we must conclude that our earlier simplifying assumptions regarding the "well-behavedness" of the image cost with respect to perturbations of the image segmentation are not valid for the interconnections in question. As the boundary of the continuity set is far from smooth at a free endpoint of an image segment, this should not be too surprising.

Pure Existence and Arc Length Cost. If $\varkappa = \iota = 0$, the optimality conditions (2.43) reduce to

$$\lambda\kappa_n + \blacktriangle\varrho_n = 0 \qquad \text{on } \Sigma, \qquad n = 1, \ldots, N \tag{2.45a}$$

$$\sum_{t=0}^{1}(-1)^t \sum_{n\in N_{jt}} e_{\tau n}(t) = 0 \qquad j = 1, \ldots, J \tag{2.45b}$$

Although these equations are much simpler than (2.44), they share some of the same fundamental complications discussed above, and can therefore not be solved directly either.

The meaning of the Euler equations (2.45a) are in this case easier to grasp; wherever an edge segment curves, the image cost density immediately "outside" the curve has to exceed that on the immediate "inside", as shown in figure 2.10. In places where the

optimal edges curve a lot, the estimated image function can thus be expected to fit the original image function better and/or to be smoother on the "inside" of the curve than on the "outside".

The geometric meaning of the boundary conditions (2.45b) is also transparent; if all the edge segments are imagined to pull on their endpoints in the local tangential direction with equal force, the edge segments, whose endpoints are forced to coincide at a junction, must meet at such angles so that the junction remains stationary. The optimality conditions thus in particular disallow edge segments with free endpoints, which once again has to be blamed on the simplifying assumptions.

Control Vertex Conditions

Suppose that the edges are represented by splines as discussed in section 2.1, and that the edge cost is of the form

$$\mathcal{E}_N(\gamma) \doteq \sum_{n=1}^{N} [\nu + \varpi \Pi(v_{n0}, \ldots v_{n,M_n+2})]$$

where $\nu \geq 0$ and $\varpi > 0$. From (2.28), (2.29) and (2.34) we then have that

$$\begin{aligned}
&\delta_v[\mathcal{E}_N(\gamma) + \mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)] = \\
&\quad = \sum_{n=1}^{N} \varpi \delta_v \Pi(v_{n0}, \ldots, v_{n,M_n+2}) + \delta_v[\mathcal{D}_{C_\gamma}(z,\zeta) + \mathcal{S}_{C_\gamma}(z)] \\
&\quad = \sum_{n\in N_0} \sum_{m=1}^{M_n} \left[\varpi \left(\frac{v_{nm} - v_{n,m-1}}{\|v_{nm} - v_{n,m-1}\|} + \frac{v_{nm} - v_{n,m+1}}{\|v_{nm} - v_{n,m+1}\|} \right)^T \right. \\
&\qquad \left. - \sum_{r=0}^{3} \sum_{s=0}^{3} \int_0^1 \blacktriangle \varrho_n((m-r) \bmod M_n + \sigma) b_r(\sigma) \dot{b}_s(\sigma)\, d\sigma v_{n,(m-r) \bmod M_n+s}^T R_\times \right] \\
&\qquad \cdot \delta v_{nm} \\
&\qquad + \sum_{n\in N_1} \left(\left[\varpi \left(\frac{v_{n2} - v_{n3}}{\|v_{n2} - v_{n3}\|} \right)^T \right. \right. \\
&\qquad\quad \left. \left. - \sum_{m=0}^{2} \sum_{r=0}^{m} \sum_{s=0}^{3} \int_0^1 \blacktriangle \varrho_n(m-r+\sigma) b_r(\sigma) \dot{b}_s(\sigma)\, d\sigma v_{n,m-r+s}^T R_\times \right] \delta v_{n2} \right. \\
&\qquad\quad + \sum_{m=3}^{M_n-1} \left[\varpi \left(\frac{v_{nm} - v_{n,m-1}}{\|v_{nm} - v_{n,m-1}\|} + \frac{v_{nm} - v_{n,m+1}}{\|v_{nm} - v_{n,m+1}\|} \right)^T \right. \\
&\qquad\quad \left. \left. - \sum_{r=0}^{3} \sum_{s=0}^{3} \int_0^1 \blacktriangle \varrho_n(m-r+\sigma) b_r(\sigma) \dot{b}_s(\sigma)\, d\sigma v_{n,m-r+s}^T R_\times \right] \delta v_{nm} \right.
\end{aligned}$$

$$+\left[\varpi\left(\frac{v_{n,M_n}-v_{n,M_n-1}}{\|v_{n,M_n}-v_{n,M_n-1}\|}\right)^T - \sum_{m=M_n}^{M_n+2}\sum_{r=m-M_n+1}^{3}\sum_{s=0}^{3}\int_0^1 \blacktriangle\varrho_n(m-r+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma v_{n,m-r+s}^T R_\times\right] \cdot \delta v_{n,M_n}\Bigg)$$

The control vertex interdependence imposed by the spline end conditions was eliminated already in the derivations of (2.28) and (2.29). In order to eliminate the interdependence due to the interconnection constraints we have to express the *end vertices* v_{n2}, v_{n,M_n}, $n = 1,\ldots,N$, of the open splines in terms of the junctions $w_1,\ldots,w_J$. This procedure leads to the following expression for the total cost variation with respect to the control vertices.

$$\delta_v[\mathcal{E}_N(\gamma)+\mathcal{D}_{C_\gamma}(z,\zeta)+\mathcal{S}_{C_\gamma}(z)] = \sum_{n=1}^{N}\sum_{m=1+2o_n}^{M_n-o_n} g_{nm}^T\delta v_{nm} + \sum_{j=1}^{J} g_j^T\delta w_j \tag{2.46}$$

where

$$g_{nm} \doteq \doteq \left(\frac{v_{nm}-v_{n,m-1}}{\|v_{nm}-v_{n,m-1}\|}+\frac{v_{nm}-v_{n,m+1}}{\|v_{nm}-v_{n,m+1}\|}\right)\varpi + R_\times\sum_{r=0}^{3}\sum_{s=0}^{3} v_{n,(m-r)\bmod M_n+s} \cdot\int_0^1 \blacktriangle\varrho_n((m-r) \bmod M_n+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma \tag{2.47a}$$

and

$$g_j \doteq \doteq \sum_{n\in N_{j0}}\left[\frac{v_{n2}-v_{n3}}{\|v_{n2}-v_{n3}\|}\varpi + R_\times\sum_{m=0}^{2}\sum_{r=0}^{m}\sum_{s=0}^{3} v_{n,m-r+s}\int_0^1 \blacktriangle\varrho_n(m-r+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma\right] + \sum_{n\in N_{j1}}\left[\frac{v_{n,M_n}-v_{n,M_n-1}}{\|v_{n,M_n}-v_{n,M_n-1}\|}\varpi + R_\times\sum_{m=M_n}^{M_n+2}\sum_{r=m-M_n+1}^{3}\sum_{s=0}^{3} v_{n,m-r+s} \cdot\int_0^1 \blacktriangle\varrho_n(m-r+\sigma)b_r(\sigma)\dot{b}_s(\sigma)\,d\sigma\right] \tag{2.47b}$$

For optimality the expression in (2.46) must vanish for all possible perturbations of the control vertices, which in turn means that all the coefficient vectors g_{nm} and g_j in (2.47) must be equal to zero. These vectors are of course nothing but the (2×1 block) components of the gradient of the total cost with respect to the independent control vertices. Setting the expressions in (2.47a) equal to zero gives the optimality conditions for the *intermediate vertices*, that is all those control vertices that are not end vertices of open splines. The optimality conditions for the end vertices or equivalently for the junctions are similarly obtained from (2.47b).

Even though the expressions in (2.47) are algebraic in the control vertices, and therefore simpler to deal with than the differential Euler equations (2.43a), the problem with the image cost density difference $\blacktriangle\varrho_n$ remains. The optimal control vertices can therefore not be found buy simply setting the total cost gradient to zero and solving the resulting set of equations. In this case the optimality conditions are moreover a bit too complicated to lend themselves to geometric interpretations. The reward for deriving the expression (2.46) is that the total cost gradient, (which is actually more conveniently represented as in (2.46) rather than by a vector requiring some artificial ordering of all the independent control vertices,) can be used for updating the edges so as to reduce the total cost by means of a *finite dimensional* gradient method. This topic will be treated in chapter 4.

Chapter 3

Existence of Optimal Edges

One of the central questions, that arises from the discussion in the previous chapter, is whether the total cost function attains its greatest lower bound for some discontinuity set formed by the edges and some estimated image function. In this chapter we present an affirmative answer to this question for a certain class of cost functions, by proving the existence of an optimal discontinuity set and an optimal estimated image function, which minimize the total cost.

3.1 Introduction

As in the previous chapter we assume, that we are given an original image function $\zeta : B \to \mathbf{R}$, and thereby implicitly, that the image domain B is a connected bounded open set in $\mathbf{R}^2$. Again we represent the edges by a finite number of continuous curves, parametrized by the functions $\gamma_1, \ldots, \gamma_N \in C(\Sigma)^2$,[★] defined on some compact interval $\Sigma \subseteq \mathbf{R}$, and define the *image segmentation*

$$\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C(\Sigma)^{2N} \tag{3.1}$$

the *discontinuity set*

$$D_\gamma \doteq \bigcup_{n=1}^{N} \gamma_n(\Sigma) \subseteq \mathbf{R}^2 \tag{3.2}$$

and the corresponding *continuity set*

$$C_\gamma \doteq B \setminus D_\gamma \tag{3.3}$$

[★] For any set $\Omega \subseteq \mathbf{R}^n$, $C(\Omega) \doteq \mathbf{C}^0(\Omega) \doteq \{f : \Omega \to \mathbf{R} : f \text{ is continuous}\}$.

As before we consider a *total cost function* of the form

$$c(N,\gamma,z) = c_N(\gamma) + c_{C_\gamma}(z) \tag{3.4}$$

where c_N and c_{C_γ} are the nonnegative real valued *edge* and *image cost functions* associated with D_γ and C_γ respectively, and $z : C_\gamma \to \mathbf{R}$ is the estimated image function, often referred to as just the image function. The difficulty of the existence proof does of course depend on the specific forms of the functions c_N and c_{C_γ} as well as on the domain, in which we allow (N,γ,z) to lie.

For the edge cost given by the functions $c_N : S_N \subseteq C(\Sigma)^{2N} \to \overline{\mathbf{R}_+}$, $N \in \mathbf{N}_0$, we will assume that

1. For each $N \in \mathbf{N}_0$ the function c_N is lower semicontinuous with respect to some topology $\mathcal{T}_N$, no weaker than the $C(\Sigma)^{2N}$-topology (induced by the norm defined in (2.1) with $l = 0$ and $K = 2N$).

2. $\inf_{\gamma \in S_N} c_N(\gamma) \longrightarrow \infty$ as $N \longrightarrow \infty$

It is easy to check, that all the examples of edge cost functions in the previous chapter satisfy this condition for some appropriate choice of the topologies $\mathcal{T}_N$, $N \in \mathbf{N}_0$.

For the image cost c_{C_γ} however, we will restrict our attention to the first example in the previous chapter, that is we will assume that

$$c_{C_\gamma}(z) = \int_{C_\gamma} \left[(z-\zeta)^2 + \mu\|\nabla z^T\|^2\right] dx \qquad z : C_\gamma \to \mathbf{R} \tag{3.5}$$

The main reason for this restriction is, that the cost functional given in (3.5) is well studied in the mathematical literature. Another reason is, that it allows our techniques to handle discontinuity sets, which are sufficiently nonsmooth to represent corners formed by intersecting edges. Among the image cost functions discussed in the previous chapter, the one given in (3.5) is also the most interesting from a practical point of view. It results, as least with the methods we have used, and which are described in chapter 4, in the simplest and fastest software implementations. The same is very likely to be true for other methods and hardware implementations as well.

In order to describe the domain of the total cost function correctly, we need to review and build up notation for a few concepts from real and functional analysis.

Since we will only consider real image functions, we will only consider real vector spaces. If X is a vector space, we denote by X^* its algebraic dual space, consisting of all

linear functionals on X. If X and Y are topological vector spaces, we denote by $\mathcal{L}(X,Y)$ the vector space of bounded linear maps from X to Y. The space $\mathcal{L}(X,\mathbf{R})$ is as usual referred to as the dual space of X, and denoted by X'. If $(X,|\cdot|_X)$ is a seminormed vector space, and $(Y,\|\cdot\|_Y)$ is a normed vector space, $\mathcal{L}(X,Y)$ is also a normed vector space with the norm

$$\|f\|_{\mathcal{L}(X,Y)} \doteq \sup_{\substack{x\in X\\ |x|\leq 1}} \|f(x)\|_Y \qquad f\in\mathcal{L}(X,Y) \tag{3.6}$$

In particular the vector space X' is normed with norm

$$\|f\|_{X'} \doteq \sup_{\substack{x\in X\\ |x|\leq 1}} |f(x)| \qquad f\in X' \tag{3.7}$$

For any set S in a space X, we denote by $\complement S$ the complement $X\setminus S$ of S, and by $\overline{S}$ and S° the closure and interior respectively of S in X. For a function $f:\Omega\subseteq\mathbf{R}^n\to\mathbf{R}^l$ we then define the *support* of f to be the set

$$\underline{f} \doteq \Omega\cap\overline{f^{-1}(\complement\{0\})} \tag{3.8}$$

that is $\underline{f}$ is the closure in its domain of the set, where it does not vanish. (Some authors define the support of a function slightly differently.) If U is an open set, and $K\subseteq U$ is compact, we write $K\subset\subset U$.

For open sets $\Omega\subseteq\mathbf{R}^n$, we denote by $C^\infty(\Omega)$ the space of real valued functions on Ω, which are continuously differentiable of all orders, and by $C_0^\infty(\Omega)$ the subspace

$$C_0^\infty(\Omega) \doteq \{f\in C^\infty(\Omega) : \underline{f}\subset\subset\Omega\} \tag{3.9}$$

Functions in $C_0^\infty(\Omega)$ we refer to as test functions (on Ω), and linear functionals on $C_0^\infty(\Omega)$, that is elements of $C_0^\infty(\Omega)^*$ we refer to as distributions. We will not find it necessary, to consider the commonly used smaller space $\mathcal{D}(\Omega)'$ of Schwartz distributions.

The pointwise partial derivative of a function $f:\Omega\to\mathbf{R}$ with respect to the kth variable x_k at a point $x\in\Omega$ we denote by $(\partial/\partial x_k)f(x)$. If $(\partial/\partial x_k)f(x)$ exists a.e., this defines a function $(\partial/\partial x_k)f$ on Ω. Thus $\partial/\partial x_k$ is a linear operator on $C_0^\infty(\Omega)$. We will denote this operator by D_k. If α is an (n-dimensional) multi-index, that is $\alpha=\langle\alpha_k\rangle_{k=1}^n\in\mathbf{N}_0^n$, of "magnitude" $|\alpha|\doteq\sum_{k=1}^n\alpha_k$, we also define the $|\alpha|$th order differential operator

$$D^\alpha \doteq \prod_{k=1}^n D_k^{\alpha_k} \tag{3.10}$$

If the product of two functions $f, g : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}$ is integrable (or nonnegative), we define

$$\langle f, g \rangle_\Omega \doteq \int_\Omega f g \, dx \tag{3.11}$$

We recall, that $\langle \cdot, \cdot \rangle_\Omega$ defines an inner product and thereby a norm $\| \cdot \|_{L_2(\Omega)}$ on the space $L_2(\Omega)$ of square integrable real valued functions on Ω, and that $C_0^\infty(\Omega)$ is dense in $L_2(\Omega)$ in the topology induced by this norm. We also recall, that $L_2(\Omega)$ is a Hilbert space. Hence its dual $L_2(\Omega)'$ can be identified with $L_2(\Omega)$ itself via (3.11) by the Riesz representation theorem. If $\Omega \subseteq \mathbf{R}^n$ is open, and $f \in L_1^{\text{loc}}(\Omega)$, the space of locally integrable real valued functions on Ω, we identify f with the distribution $C_0^\infty(\Omega) \to \mathbf{R} : \varphi \mapsto \langle f, \varphi \rangle_\Omega$, and define $D^\alpha f \in C_0^\infty(\Omega)^*$ by

$$D^\alpha f(\varphi) \doteq |-1|^{|\alpha|} \langle f, D^\alpha \varphi \rangle_\Omega \qquad \varphi \in C_0^\infty(\Omega) \tag{3.12}$$

It will often be the case, that there exists a function $g \in L_1^{\text{loc}}(\Omega)$, such that

$$D^\alpha f(\varphi) = \langle g, \varphi \rangle_\Omega \qquad \forall \varphi \in C_0^\infty(\Omega) \tag{3.13}$$

We will then identify $D^\alpha f$ with g. Finally for open sets $\Omega \subseteq \mathbf{R}^n$ and $l \in \mathbf{N}_0$ we define the Sobolev space

$$\mathcal{H}^l(\Omega) \doteq \{ f \in L_2(\Omega) : D^\alpha f \in L_2(\Omega), \quad \alpha \in \mathbf{N}_0^n, \quad |\alpha| \leq l \} \tag{3.14}$$

On $\mathcal{H}^l(\Omega)$ we also define the inner product

$$\langle f, g \rangle_{\Omega, l} \doteq \sum_{\substack{|\alpha| \leq l \\ \alpha \in \mathbf{N}_0^n}} \int_\Omega D^\alpha f D^\alpha g \, dx \qquad f, g \in \mathcal{H}^l(\Omega) \tag{3.15}$$

and the corresponding norm

$$\| f \|_{\mathcal{H}^l(\Omega)} \doteq \sqrt{\langle f, f \rangle_{\Omega, l}} \qquad f \in \mathcal{H}^l(\Omega) \tag{3.16}$$

The inner product spaces so defined have the following important property.

Theorem 3.1.1 *The Sobolev space $\mathcal{H}^l(\Omega)$ is a separable Hilbert space.*

Proof: The completeness of $L_2(\Omega)$ implies, that $\mathcal{H}^l(\Omega)$ is also complete and hence a Hilbert space. For the proof that $\mathcal{H}^l(\Omega)$ is separable we refer to [55, p47]. ■

It turns out, that the appropriate domain for the total cost function $c(N,\gamma,z)$ is given by

$$\mathcal{D}_{\mathcal{K}} \doteq \{(N,\gamma,z) : z \in \mathcal{H}^1(C_\gamma), \gamma \in K_N \in \mathcal{K}, N \in \mathbb{N}_0\} \tag{3.17}$$

where K_N is compact in $(S_N, \mathcal{T}_N)$, $N \in \mathbb{N}_0$. We will show that there exists a collection $\mathcal{K} \doteq \{K_N\}_{N\in\mathbb{N}_0}$ of such compact sets of image segmentations, such that $\bigcup_{N\in\mathbb{N}_0} K_N$ is large enough to describe the edges in most images, and that the total cost function $c(N,\gamma,z)$ attains its minimum on the resulting domain $\mathcal{D}_{\mathcal{K}}$.

3.2 Outline of Existence Proof

The proof, that the total cost function attains its minimum on the domain $\mathcal{D}_{\mathcal{K}}$ for a sufficiently large collection $\mathcal{K}$ of sets of image segmentations is quite long. We have therefore decomposed it into a number of parts, each of which is presented in a separate section of this chapter. To help the reader in getting the overall picture, in this section we present a brief outline of the whole proof section by section. This will hopefully motivate each of these sections, and clarify their interrelationships.

We begin by proving, that for each open subset Ω of the image domain B, there exists a *unique* image function $z_\Omega \in \mathcal{H}^1(\Omega)$, which minimizes the image cost c_Ω on Ω. Thus for each $N \in \mathbb{N}_0$ we can define the *optimal N-(edge-)segment image cost function*

$$\check{c}_N : C(\Sigma)^{2N} \to \overline{\mathbb{R}_+} : \gamma \mapsto c_{C_\gamma}(z_{C_\gamma}) \tag{3.18}$$

The main idea is then to select for each $N \in \mathbb{N}_0$ a sufficiently large compact set K_N in $(S_N, \mathcal{T}_N)$, and show, that $\check{c}_N|K_N$ is lower semicontinuous. It then follows in a straight forward manner from the assumptions on the edge cost, that the total cost function attains its minimum on $\mathcal{D}_{\mathcal{K}}$.

In section 3.3 we review some Hilbert space methods for elliptical problems, and use these methods to show the existence of a unique *optimal image function* z_Ω, which minimizes the image cost $c_\Omega(z)$ for any given open set $\Omega \subseteq B$, or equivalently for fixed edges. Thus with each open set $\Omega \subseteq B$ there is an associated *optimal image cost* $c_\Omega(z_\Omega)$. This result is important, because it allows us to define the optimal N-segment image cost $\check{c}_N$, $N \in \mathbb{N}_0$, as a function of the image segmentation $\gamma \in C(\Sigma)^{2N}$. It is also of importance in our later efforts to show, that this function $\check{c}_N$ is lower semicontinuous.

In section 3.4 we define the notion of a *Lipschitz chart*, and derive some of the properties of such charts. These charts are homeomorphisms, which will later be used to form atlases on neighborhoods of boundaries of subsets of the image domain.

In section 3.5 we define the *Lipschitz property* for domains in $\mathbf{R}^2$, and show, that the "natural" atlas on a neighborhood of the boundary of a domain with this property, also called a *Lipschitz domain*, consists of Lipschitz charts.

In section 3.6 we prove a number of simple results about restrictions and trivial extensions of functions.

In section 3.7 we use the results from the sections 3.4 – 3.6 to construct an extension operator $P_\Omega \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))$ for a Lipschitz domain Ω, and find an explicit bound on the norm of P_Ω in terms of an atlas of Lipschitz charts on a neighborhood of its boundary $\partial\Omega$.

In section 3.8 we define the concept of *admissible image segments*, and present a sufficient condition for subsets of the image domain to belong to this category. We then find a lower bound for any upper bound of the optimal image costs for a certain class of interior set approximations of an admissible image domain in terms of the optimal image cost for the admissible image domain itself.

In section 3.9 we define, what we mean by an image segmentation being *admissible*. Using the result from section 3.8 we then show, that the optimal N-segment image cost $\check{c}_N$ is lower semicontinuous on the set of admissible image segmentations.

In section 3.10 we use the edge cost assumptions given in the previous section and the lower semicontinuity result from section 3.9 to prove the existence of optimal edges, or equivalently of an optimal image segmentation with an optimal number of edge segments, which minimizes the total cost function over the entire image domain. As mentioned earlier, this minimization is done over a restricted domain $\mathcal{D}_\mathcal{K}$ of image segmentations.

Section 3.10 essentially completes our existence proof, except that it does not specify the domain $\mathcal{D}_\mathcal{K}$. In section 3.11 we therefore present a nontrivial example of a collection $\mathcal{K}$ of image segmentation domains, which we claim, is rich enough to describe the edges in most images, and moreover results in a total cost function domain $\mathcal{D}_\mathcal{K}$, which satisfies the conditions required by the existence proof in section 3.10.

3.3 Optimal Images for Fixed Edges

In this section we consider the simplified problem of proving the existence of an optimal image function, which minimizes the image cost for a given open subset of the image domain. Given an image domain $B \subseteq \mathbf{R}^2$ and an original image function $\zeta \in L_2(B)$ we define for open sets $\Omega \subseteq B$, the *image cost function*

$$c_\Omega : \mathcal{H}^1(\Omega) \to \overline{\mathbf{R}_+} : z \mapsto \int_\Omega \left[(z-\zeta)^2 + \mu \sum_{k=1}^{2} (D_k z)^2 \right] dx \qquad \mu > 0 \tag{3.19}$$

Our task then is to prove, that $\operatorname{argmin}\{c_\Omega(z) : z \in \mathcal{H}^1(\Omega)\}$ exists. For this purpose we need the concept of directional derivative of functionals, which supports a rigorous version of the variational calculus ideas, we applied in the previous chapter.

Definition 3.3.1 *Let X be a vector space. A functional $f : X \to \mathbf{R}$ is said to be* differentiable in the sense of Gâteaux, *or simply* G-differentiable *at $x \in X$, if there exists a functional $f^{(1)}(x) \in X'$, such that*

$$\lim_{t \downarrow 0} \frac{f(x+ty) - f(x)}{t} = f^{(1)}(x)(y) \qquad \forall y \in X$$

In this case $f^{(1)}(x)$ is called the G-differential *of f at x.*

Since $\zeta \in L_2(B)$, it is now easy to establish the following fact.

Fact 3.3.2 *The image cost c_Ω is G-differentiable everywhere in $\mathcal{H}^1(\Omega)$, and its G-differential is given by*

$$c_\Omega^{(1)}(z) : \mathcal{H}^1(\Omega) \to \mathbf{R} : y \mapsto 2 \int_\Omega \left(zy + \mu \sum_{k=1}^{2} D_k z D_k y \right) dx - 2\langle \zeta | \Omega, y \rangle_\Omega \qquad \forall z \in \mathcal{H}^1(\Omega)$$

Proof: Let $z, y \in \mathcal{H}^1(\Omega)$. Then

$$\begin{aligned}
&\frac{c_\Omega(z+ty) - c_\Omega(z)}{t} = \\
&= \frac{1}{t} \int_\Omega \left((z+ty-\zeta)^2 + \mu \sum_{k=1}^{2} [D_k(z+ty)]^2 - (z-\zeta)^2 - \mu \sum_{k=1}^{2} (D_k z)^2 \right) dx \\
&= 2 \int_\Omega \left(zy + \mu \sum_{k=1}^{2} D_k z D_k y - \zeta y \right) dx + t \int_\Omega \left[y^2 + \mu \sum_{k=1}^{2} (D_k y)^2 \right] dx \\
&\longrightarrow 2 \int_\Omega \left(zy + \mu \sum_{k=1}^{2} D_k z D_k y \right) dx - \langle \zeta | \Omega, y \rangle_\Omega \\
&= c_\Omega^{(1)}(z)(y) \qquad \text{as } t \downarrow 0
\end{aligned}$$

with $c_\Omega^{(1)}$ given as above. Clearly $c_\Omega^{(1)}(z)$ is linear, and by the Schwarz inequality

$$\begin{aligned}|c_\Omega^{(1)}(z)(y)| \leq \\ &\leq 2(1 \vee \mu)\|z\|_{\mathcal{H}^1(\Omega)}\|y\|_{\mathcal{H}^1(\Omega)} + 2\|\zeta|\Omega\|_{L_2(\Omega)}\|y\|_{L_2(\Omega)} \\ &\leq 2\left[(1 \vee \mu)\|z\|_{\mathcal{H}^1(\Omega)} + \|\zeta|\Omega\|_{L_2(\Omega)}\right]\|y\|_{\mathcal{H}^1(\Omega)}\end{aligned}$$

Thus $c_\Omega^{(1)}(z) \in \mathcal{H}^1(\Omega)'$. ■

Motivated by the form of the G-differential $c_\Omega^{(1)}$, for open sets $\Omega \subseteq B$ we now define the bilinear form

$$a_\Omega : \mathcal{H}^1(\Omega)^2 \to \mathbf{R} : (z, y) \mapsto \int_\Omega \left(zy + \mu \sum_{k=1}^{2} D_k z D_k y \right) dx \tag{3.20}$$

and note that

$$c_\Omega(z) = a_\Omega(z, z) - 2\langle \zeta|\Omega, z\rangle_\Omega + \|\zeta|\Omega\|^2_{L_2(\Omega)} \qquad \forall z \in \mathcal{H}^1(\Omega) \tag{3.21}$$

and

$$c_\Omega^{(1)}(z)(y) = 2a_\Omega(z, y) - 2\langle \zeta|\Omega, y\rangle_\Omega \tag{3.22}$$

We then continue with a couple of definitions and facts concerning $c_\Omega^{(1)}$.

Definition 3.3.3 *Let X be a vector space. A function $f : X \to X'$ is said to be* monotone, *if*

$$[f(x) - f(y)](x - y) \geq 0 \qquad \forall x, y \in X$$

Fact 3.3.4 *The G-differential $c_\Omega^{(1)} : \mathcal{H}^1(\Omega) \to \mathcal{H}^1(\Omega)'$ of the image cost is monotone.*

Proof: Let $z, y \in \mathcal{H}^1(\Omega)$. Then by (3.22) and the bilinearity of a_Ω we have

$$\left[c_\Omega^{(1)}(z) - c_\Omega^{(1)}(y)\right](z - y) = 2a_\Omega(z, z - y) - 2a_\Omega(y, z - y) = 2a_\Omega(z - y, z - y) \geq 0$$

■

Definition 3.3.5 *Let X be a normed vector space. A function $f : X \to X'$ is said to be* coercive, *if*

$$\frac{f(x)(x)}{\|x\|_X} \longrightarrow \infty \qquad \text{as } \|x\|_X \longrightarrow \infty$$

Fact 3.3.6 *The G-differential $c_\Omega^{(1)} : \mathcal{H}^1(\Omega)'$ is coercive.*

Proof: For $z \in \mathcal{H}^1(\Omega)$ by (3.22) we have

$$\begin{aligned}\frac{c_\Omega^{(1)}(z)(z)}{\|z\|_{\mathcal{H}^1(\Omega)}} &= \\ &= \frac{2a_\Omega(z,z) - 2\langle \zeta|\Omega, z\rangle_\Omega}{\|z\|_{\mathcal{H}^1(\Omega)}} \\ &\geq 2(1 \wedge \mu)\|z\|_{\mathcal{H}^1(\Omega)} - 2\|\zeta|\Omega\|_{L_2(\Omega)} \longrightarrow \infty \qquad \text{as } \|z\|_{\mathcal{H}^1(\Omega)} \longrightarrow \infty\end{aligned}$$

■

We are now in the position to make use of the following theorem, of which a proof can be found in [56, p157].

Theorem 3.3.7 *Let X be a separable Hilbert space, and let the functional $f : X \to \mathbf{R}$ be G-differentiable on X. Assume its G-differential $f^{(1)}$ is monotone and coercive. Then*

$$\{x \in X : f(y) \geq f(x) \quad \forall y \in X\} = \{x \in X : f^{(1)}(x)(y) = 0 \quad \forall y \in X\}$$

From theorem 3.1.1 and the facts 3.3.2, 3.3.4 and 3.3.6 we see, that theorem 3.3.7 above applies to the functional $c_\Omega : \mathcal{H}^1(\Omega) \to \mathbf{R}$. Hence from (3.22) we conclude, that the set of functions in $\mathcal{H}^1(\Omega)$, which minimize c_Ω is given by

$$\begin{aligned}&\{z \in \mathcal{H}^1(\Omega) : c_\Omega(y) \geq c_\Omega(z) \quad \forall y \in \mathcal{H}^1(\Omega)\} = \\ &\quad = \{z \in \mathcal{H}^1(\Omega) : a_\Omega(z,y) = \langle \zeta|\Omega, y\rangle_\Omega \quad \forall y \in \mathcal{H}^1(\Omega)\} \qquad (3.23)\end{aligned}$$

Our goal is now to show, that this set contains exactly one function $z_\Omega \in \mathcal{H}^1(\Omega)$. In order to do so, we need a few more definitions and facts regarding a_Ω and $\zeta|\Omega$.

Definition 3.3.8 *Let X be a Hilbert space with norm $\|\cdot\|_X$. A bilinear form $a : X^2 \to \mathbf{R}$ is said to be X-*coercive*, if there exists a constant $c > 0$, such that $|a(x,x)| \geq c\|x\|_X^2 \; \forall x \in X$. The constant c will then be referred to as the* coercivity *of the bilinear form a.*

Fact 3.3.9 *The bilinear form $a_\Omega : \mathcal{H}^1(\Omega)^2 \to \mathbf{R}$ is $\mathcal{H}^1(\Omega)$-coercive with coercivity $1 \wedge \mu$ and continuous (with respect to the product topology).*

Proof: The coercivity claim follows immediately from the definition of a_Ω and definition 3.3.8. To prove, that a_Ω is continuous, let $z, y, z_1, y_1 \in \mathcal{H}^1(\Omega)$. Then by the bilinearity of a_Ω we have

$$\begin{aligned}
&|a_\Omega(z_1, y_1) - a_\Omega(z, y)| \leq \\
&\quad \leq |a_\Omega(z_1 - z, y_1 - y)| + |a_\Omega(z_1 - z, y)| + |a_\Omega(z, y_1 - y)| \\
&\quad \leq (1 \vee \mu) \\
&\qquad \cdot \Big(\|z_1 - z\|_{\mathcal{H}^1(\Omega)} \|y_1 - y\|_{\mathcal{H}^1(\Omega)} + \|z_1 - z\|_{\mathcal{H}^1(\Omega)} \|y\|_{\mathcal{H}^1(\Omega)} \\
&\qquad\quad + \|z\|_{\mathcal{H}^1(\Omega)} \|y_1 - y\|_{\mathcal{H}^1(\Omega)} \Big) \\
&\quad \longrightarrow 0 \qquad \text{as } (z_1, y_1) \overset{\mathcal{H}^1(\Omega)^2}{\longrightarrow} (z, y)
\end{aligned}$$

■

Since $\zeta \in L_2(B)$, for open sets $\Omega \subseteq B$ we can define the functional

$$f_{\zeta|\Omega} : \mathcal{H}^1(\Omega) \to \mathbf{R} : z \mapsto \langle \zeta|\Omega, z\rangle_\Omega \tag{3.24}$$

Fact 3.3.10 *The functional $f_{\zeta|\Omega} \in \mathcal{H}^1(\Omega)'$ and $\|f_{\zeta|\Omega}\|_{\mathcal{H}^1(\Omega)'} \leq \|\zeta\|_{L_2(B)}$.*

Proof: Clearly $f_{\zeta|\Omega}$ is linear. Moreover for $z \in \mathcal{H}^1(\Omega)$ we have

$$|f_{\zeta|\Omega}(z)| \leq \|\zeta|\Omega\|_{L_2(\Omega)} \|z\|_{L_2(\Omega)} \leq \|\zeta\|_{L_2(B)} \|z\|_{\mathcal{H}^1(\Omega)}$$

■

With these preparations we are now ready to apply the following theorem, of which a proof can be found in [56, p54].

Theorem 3.3.11 *Let X be a Hilbert space with norm $\|\cdot\|_X$, and let $a : X^2 \to \mathbf{R}$ be an X-coercive continuous bilinear form with coercivity $c > 0$. Then for each $f \in X'$ there exists a unique $x \in X$, such that $a(x, y) = f(y)\ \forall y \in X$. Furthermore*

$$\|x\|_X \leq \frac{\|f\|_{X'}}{c}$$

From theorem 3.1.1 and the facts 3.3.6 and 3.3.10 we see, that theorem 3.3.11 applies to the bilinear form a_Ω and the functional $f_{\zeta|\Omega}$. Thus by (3.23) we have the following important conclusion.

Theorem 3.3.12 (Existence of a Unique Optimal Image Function) *Let $\zeta \in L_2(B)$ be an original image function, and let $\Omega \subseteq B$ be an open set. Furthermore let the image cost function $c_\Omega : \mathcal{H}^1(\Omega) \to \overline{\mathbf{R}_+}$ and the bilinear form $a_\Omega : \mathcal{H}^1(\Omega)^2 \to \mathbf{R}$ be defined as in (3.19) and (3.20) respectively. Then there exists a unique* optimal image function $z_\Omega \in \mathcal{H}^1(\Omega)$, *which minimizes c_Ω. Moreover*

$$a_\Omega(z_\Omega, z) = \langle \zeta|\Omega, z\rangle_\Omega \qquad \forall z \in \mathcal{H}^1(\Omega)$$

and

$$\|z_\Omega\|_{\mathcal{H}^1(\Omega)} \leq \frac{\|\zeta\|_{L_2(B)}}{1 \wedge \mu}$$

For the cost $c_\Omega(z_\Omega)$, referred to as the *optimal image cost over* Ω, this also implies

Corollary 3.3.13 *Let B, ζ, Ω and c_Ω be as in theorem 3.3.12. Then the optimal image cost over Ω is given by*

$$c_\Omega(z_\Omega) = \|\zeta|\Omega\|^2_{L_2(\Omega)} - \langle \zeta|\Omega, z_\Omega\rangle_\Omega$$

Proof: From theorem 3.3.12 we see that

$$a_\Omega(z_\Omega, z_\Omega) = \langle \zeta|\Omega, z_\Omega\rangle_\Omega$$

The corollary then follows from (3.21). ■

Another important consequence of theorem 3.3.12 is, that it shows, that the optimal image cost over the entire continuity set is a function of this set alone. We will make use of this idea in section 3.9, where we consider the optimal image cost over the continuity set as a function of the edge segments forming the corresponding discontinuity set.

3.4 Lipschitz Charts

Having demonstrated the existence of a unique optimal image function yielding an optimal image cost for any given edges, the next obvious question is, whether we can find some edges, that is a discontinuity set, which minimizes this optimal cost. To answer that question we will have to be more specific about, which discontinuity sets are to be considered. Basically we will require, that the components of the resulting continuity set have boundaries, which are locally described by the graphs of Lipschitz continuous functions.

This requirement, which does not even imply, that the components of the continuity set have piecewise smooth boundaries, is yet sufficiently powerful to support the necessary analysis of the image functions near these boundaries as well as of the boundaries themselves. The *Lipschitz chart* is the basic technical tool in this context. In this section we present its definition along with some related results, that are needed later.

Definition 3.4.1 *We say, that a homeomorphism* $\Phi : Q \subseteq \mathbf{R}^n \to U \to \mathbf{R}^n$ *is a* Lipschitz chart, *if both* Φ *and its inverse* Φ^{-1} *are Lipschitz continuous and differentiable a.e. (almost everywhere).*

For any function $f : V \subseteq \mathbf{R}^n \to W \subseteq \mathbf{R}^l$, whose first partial derivatives exist at $x \in V$, we denote by $J_f(x)$ the Jacobian matrix of f at x. If $J_f(x)$ exists a.e., this defines an equivalence class of matrix valued functions, which we denote by J_f. For any square matrix we also define $|J| \doteq |\det J|$. We then have the following extension of the regular change of variables formula for multiple integrals.

Proposition 3.4.2 *Let* $\Phi : Q \to U$ *be a Lipschitz chart. Then*

$$\int_U f\,dy = \int_Q (f \circ \Phi)|J_\Phi|\,dx \qquad \forall f \in L_1(U)$$

and obviously by symmetry

$$\int_Q f\,dx = \int_U (f \circ \Phi^{-1})|J_{\Phi^{-1}}|\,dy \qquad \forall f \in L_1(Q)$$

This fact can be proven, by modifying the proof of [57, Theorem 8.26, p185]. It differs from that theorem, in that the differential of Φ is allowed to *not* exist on a set of measure zero. It differs from the "regular change of variables formula" seen in most books, in that Φ is *not* required to be a C^1-diffeomorphism.

In order to use proposition 3.4.2 for our Lipschitz charts, we will need some bounds on the Jacobians involved as well as expressions for the distributional derivatives of certain Lipschitz continuous functions.

Proposition 3.4.3 *Let* $\Phi : V \subseteq \mathbf{R}^n \to W \subseteq \mathbf{R}^n$ *be Lipschitz continuous with Lipschitz constant* L *and differentiable a.e. Then*

(i) $\|J_\Phi(x)\| \leq L$ *a.e. in* V

(ii) $|J_\Phi(x)| \leq L^n$ *a.e. in* V

Proof: For all $x \in V$ for which $J_\Phi(x)$ exists, we have

$$\begin{aligned}
\|J_\Phi(x)\|^2 &= \\
&= \sup_{\|y\| \le t} \|J_\Phi(x)\frac{y}{t}\|^2 \\
&= \frac{1}{t^2} \sup_{\|y\| \le t} \|\Phi(x+y) - \Phi(x) + o(\|y\|)\|^2 \\
&= \frac{1}{t^2} \sup_{\|y\| \le t} \left[\|\Phi(x+y) - \Phi(x)\|^2 + 2\|\Phi(x+y) - \Phi(x)\|o(\|y\|) + o(\|y\|)^2\right] \\
&\le \frac{1}{t^2} \left[L^2t^2 + 2Lt\,o(t) + o(t)^2\right] \\
&= L^2 + O(t) \qquad \forall t > 0
\end{aligned}$$

Hence (i) follows. Moreover

$$|J_\Phi(x)|^2 = |J_\Phi(x)^T J_\Phi(x)| = \prod_{k=1}^{n} \lambda_k \left[J_\Phi(x)^T J_\Phi(x)\right] \le \|J_\Phi(x)\|^{2n}$$

from which (ii) follows. ■

We say that a function $f : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}^l$ is *absolutely continuous on a line* $\Lambda \subseteq \mathbf{R}^n$, if $\Lambda \cap \Omega \neq \varnothing$, and f (that is each of its components) is absolutely continuous on every closed interval in $\Lambda \cap \Omega$. For such functions the *distributional* partial derivatives equal their corresponding *pointwise* partial derivatives, according to the following theorem, which can be found in [58, p61].

Theorem 3.4.4 *Let Ω be an open set in $\mathbf{R}^n$, and let $f \in L_1^{\text{loc}}(\Omega)$. Assume that f is absolutely continuous on almost all lines parallel to the x_k-axis in $\mathbf{R}^n$, and that its pointwise partial derivative $\partial f/\partial x_k$, (which exists a.e.) $\in L_p(\Omega)$, $p \ge 0$. Then its corresponding distributional partial derivative $D_k f$ is given by*

$$D_k f = \frac{\partial f}{\partial x_k} \qquad \text{a.e.}$$

Corollary 3.4.5 *Let Ω be a bounded open set in $\mathbf{R}^n$, and assume, that the function $f : \Omega \to \mathbf{R}^l$ is Lipschitz continuous. Then $D_k f = \partial f/\partial x_k$ a.e., $k = 1, \ldots n$.*

Proof: Being continuous, $f \in L_1^{\text{loc}}(\Omega)$. Since f is Lipschitz continuous, it is absolutely continuous along *all* lines (parallel to the coordinate axes) in $\mathbf{R}^n$, and its pointwise partial derivatives are bounded and exist a.e. ■

The conclusions of the previous discussion are important, because they lead up to the following result, which basically asserts, that composition with a Lipschitz chart (or its inverse) can be viewed as a bounded linear operator, whose norm depends only on the Lipschitz constants of the Lipschitz chart and its inverse.

Proposition 3.4.6 *Let $\Phi : Q \subseteq \mathbf{R}^n \to U \subseteq \mathbf{R}^n$ be a Lipschitz chart, and let L_Φ and $L_{\Phi^{-1}}$ be the Lipschitz constants of Φ and Φ^{-1} respectively. If $f \in \mathcal{H}^1(U)$ is Lipschitz continuous, then $f \circ \Phi \in \mathcal{H}^1(Q)$ and*

$$\|f \circ \Phi\|_{\mathcal{H}^1(Q)} \leq \sqrt{(1 + L_\Phi^2) L_{\Phi^{-1}}^n}\, \|f\|_{\mathcal{H}^1(U)}$$

Proof: Applying propositions 3.4.2 and 3.4.3 to Φ^{-1} we have

$$\|f \circ \Phi\|_{L_2(Q)}^2 = \int_Q |f \circ \Phi|^2\, dx = \int_U |f|^2 |J_{\Phi^{-1}}|\, dy \leq L_{\Phi^{-1}}^n \|f\|_{L_2(U)}^2$$

Since f and Φ are Lipschitz continuous, so is $f \circ \Phi$. By corollary 3.4.5 we thus have

$$\begin{aligned}
&\sum_{\substack{\alpha \in \mathbf{N}_0^n \\ |\alpha| = 1}} \|D^\alpha (f \circ \Phi)\|_{L_2(Q)}^2 = \\
&\quad = \int_Q \|\nabla (f \circ \Phi)^T\|^2\, dx \\
&\quad = \int_Q \|J_\Phi^T (\nabla f \circ \Phi)^T\|^2\, dx \\
&\quad \leq \int_Q L_\Phi^2 \|(\nabla f \circ \Phi)^T\|^2\, dx \\
&\quad = L_\Phi^2 \int_U \|\nabla f^T\|^2 |J_{\Phi^{-1}}|\, dy \\
&\quad \leq L_\Phi^2 L_{\Phi^{-1}}^n \sum_{\substack{\alpha \in \mathbf{N}_0^n \\ |\alpha| = 1}} \|D^\alpha f\|_{L_2(U)}
\end{aligned}$$

Hence

$$\|f \circ \Phi\|_{\mathcal{H}^1(Q)}^2 \leq (1 + L_\Phi^2) L_{\Phi^{-1}}^n \|f\|_{\mathcal{H}^1(U)}^2$$

■

3.5 Lipschitz Domains

The discussion in the previous section about Lipschitz charts was motivated by a desired property of the components of the continuity set, namely that these have boundaries,

which are locally described by the graphs of Lipschitz continuous functions. In this section we give the precise definition of this desired property, which we naturally shall call the *Lipschitz property.* We then show, how this property relates to the Lipschitz charts in the previous section, and explain the reason for the usefulness of its concept.

In manipulating functions $z \in \mathcal{H}^1(\Omega)$ close to the boundary of an open set $\Omega \subseteq \mathbf{R}^2$, we will frequently map a patch of Ω onto a patch of a half-space in $\mathbf{R}^2$. For this purpose we define the three sets $\mathbf{R}^2_{\pm} \doteq \mathbf{R} \times \mathbf{R}_{\pm}$ and $\mathbf{R}^2_0 \doteq \mathbf{R} \times \{0\}$, where as before $\mathbf{R}_{\pm} \doteq \pm]0, \infty[$. We then have the following.

Definition 3.5.1 *We say, that a bounded open domain $\Omega \subseteq \mathbf{R}^2$ is a* Lipschitz domain, *if there exists a finite collection $\{T_m\}_{m=1}^M$ of (rigid) coordinate transformations, a collection $\{\phi_m : \Delta_m \doteq]a_m, b_m[\to \mathbf{R}\}_{m=1}^M$ of corresponding Lipschitz continuous functions and a number $d > 0$, such that the maps*

$$\begin{aligned} \Phi_m \quad &: \quad Q_m \doteq \Delta_m \times]-d, d[\to U_m \doteq \Phi_m(Q_m) \\ &: \quad x \mapsto T_m(x_1, \phi_m(x_1) + x_2) \qquad m = 1, \ldots, M \end{aligned}$$

satisfy the following conditions:

(i) $\Phi_m(\mathbf{R}^2_+ \cap Q_m) \subseteq \Omega \qquad m = 1, \ldots, M$

(ii) $\Phi_m(\mathbf{R}^2_- \cap Q_m) \subseteq \complement\overline{\Omega} \qquad m = 1, \ldots, M$

(iii) $\bigcup_{m=1}^{M} \Phi_m(\mathbf{R}^2_0 \cap Q_m) = \partial\Omega$

Here as well as in the mathematical literature Lipschitz domains are also referred to as domains having the *Lipschitz property* or domains of *class* $C^{0,1}$.

The maps $\Phi_1, \ldots \Phi_M$ in the definition above are indeed Lipschitz charts. This fact follows from a special case of the following proposition, which will be useful in later sections as well.

Proposition 3.5.2 *Consider two Lipschitz continuous functions $\phi, \chi :]a, b[\to \mathbf{R}$. If $\overline{\chi(]a, b[)} \subset\subset \mathbf{R}_+$, then the map*

$$\Phi :]a, b[\times \mathbf{R} \to]a, b[\times \mathbf{R} : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} x_1 \\ \chi(x_1)x_2 + \phi(x_1) \end{bmatrix}$$

has an inverse

$$\Phi^{-1} : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} x_1 \\ \frac{x_2 - \phi(x_1)}{\chi(x_1)} \end{bmatrix}$$

Moreover, Φ and Φ^{-1} are differentiable a.e. and satisfy the local Lipschitz conditions

$$\left.\begin{array}{l} \|\Phi(y) - \Phi(x)\| \\ \|\Phi^{-1}(y) - \Phi^{-1}(x)\| \end{array}\right\} \leq L(\|x\|)\|y - x\| \qquad \forall x, y \in]a, b[\times\mathbf{R}$$

where the "local Lipschitz constant" $L : \overline{\mathbf{R}_+} \to \overline{\mathbf{R}_+}$ is an increasing function that can be fully specified in terms of any constant exceeding the Lipschitz constants of ϕ and χ as well as the three constants $\sup_{x_1 \in]a,b[} \phi(x_1)$, $\sup_{x_1 \in]a,b[} \chi(x_1)$ and $1/\inf_{x_1 \in]a,b[} \chi(x_1)$.

Proof: It is easy to verify that Φ^{-1} as given above is a well-defined inverse of Φ. By straight forward calculation we have

$$\begin{aligned} &\Phi(x + y) - \Phi(x) = \\ &\quad = \begin{bmatrix} y_1 \\ [\chi(x_1 + y_1) - \chi(x_1)]\, x_2 + \chi(x_1 + y_1) y_2 + \phi(x_1 + y_1) - \phi(x_1) \end{bmatrix} \end{aligned} \tag{3.25}$$

Since ϕ and χ are Lipschitz continuous, they are absolutely continuous, and hence differentiable a.e. It thus follows that

$$\begin{aligned} \Phi(x + y) - \Phi(x) - \begin{bmatrix} 1 & 0 \\ \chi^{(1)}(x_1) x_2 + \phi^{(1)}(x_1) & \chi(x_1) \end{bmatrix} y = \\ = \begin{bmatrix} 0 \\ o(y_1) x_2 + [\chi(x_1 + y_1) - \chi(x_1)]\, y_2 + o(y_1) \end{bmatrix} \\ = o(\|y\|) \end{aligned}$$

for almost all $x_1 \in]a, b[\;\; \forall x_2 \in \mathbf{R}$. Hence Φ is differentiable a.e. Let L_ϕ and L_χ be the Lipschitz constants of ϕ and χ respectively, and let $M_\chi \doteq \sup_{x_1 \in]a,b[} \chi(x_1)$. From (3.25) we then see that

$$\begin{aligned} &\|\Phi(x + y) - \Phi(x)\|^2 \leq \\ &\quad \leq |y_1|^2 + (L_\chi |y_1||x_2| + M_\chi |y_2| + L_\phi |y_1|)^2 \\ &\quad \leq \left[1 + (L_\chi \|x\| + M_\chi + L_\phi)^2\right] \|y\|^2 \qquad \forall x, x + y \in]a, b[\times\mathbf{R} \end{aligned}$$

which shows, that Φ satisfies the local Lipschitz condition. Finally we note that if we substitute $\phi(x_1)$ for $-\phi(x_1)/\chi(x_1)$ and $\chi(x_1)$, for $1/\chi(x_1)$ then Φ^{-1} takes the original

form of Φ. Since the maps $x_1 \mapsto -\phi(x_1)/\chi(x_1)$ and $x_1 \mapsto 1/\chi(x_1)$ are Lipschitz continuous with Lipschitz constants $(L_\phi M_\chi + M_\phi L_\chi)/m_\chi^2$ and L_χ/m_χ^2 respectively where $M_\phi \doteq \sup_{x_1 \in]a,b[} |\phi(x_1)| < \infty$ and $m_\chi \doteq \inf_{x_1 \in]a,b[} \chi(x_1) > 0$, the proposition then follows. ■

Since any (rigid) coordinate transformation $T : \mathbf{R}^2 \to \mathbf{R}^2$ is an affine isometry, and hence a diffeomorphism of class C^∞, composition with a coordinate transformation does neither affect the properties of invertibility and differentiability, nor the values of Lipschitz constants. It is therefore evident from proposition 3.5.2, that the maps $\Phi_1, \ldots, \Phi_M$ in definition 3.5.1 are Lipschitz charts according to definition 3.4.1. In particular this implies, that any map Φ meeting the conditions of $\Phi_1, \ldots, \Phi_M$ in definition 3.5.1 is open. Since Ω is bounded, and $\partial\Omega$ therefore compact, this means, that definition 3.5.1 is equivalent to the condition, that for each $x \in \partial\Omega$ there exist a coordinate transformation T_x, a Lipschitz continuous function $\phi_x : \Delta_x \doteq]a_x, b_x[\to \mathbf{R}$ and a number $d_x > 0$, such that the map

$$\Phi_x : Q_x \doteq \Delta_x \times]-d_x, d_x[\to \mathbf{R}^2 : y \mapsto T_x(y_1, \phi_x(y_1) + y_2)$$

satisfies

(i) $\Phi_x(\mathbf{R}^2_+ \cap Q_x) \subseteq \Omega$

(ii) $\Phi_x(\mathbf{R}^2_- \cap Q_x) \subseteq \complement\overline{\Omega}$

(iii) $x \in \Phi_x(\mathbf{R}^2_0 \cap Q_x) \subseteq \partial\Omega$

For verification of condition (iii) above we will find it useful to consider the notions of (function) graphs and congruence. We therefore introduce the following notation. For any function $f : X \to Y$ we denote by $\mathcal{F}_f$ its graph $\{(x, f(x)) : x \in X\} \subseteq X \times Y$. Two sets $U, V \subseteq \mathbf{R}^n$ are said to be congruent (to each other), and we write $U \cong V$, if there exists a rigid coordinate transformation $T : \mathbf{R}^n \to \mathbf{R}^n$, such that $V = T(U)$. Using this notation the second relation in condition (iii) can be written: $\mathcal{F}_\phi \cong \Gamma \subseteq \partial\Omega$.

For any set $\Omega \subseteq \mathbf{R}^n$, we denote by $C_0^\infty(\mathbf{R}^n)|\Omega$ the space of functions in $C_0^\infty(\mathbf{R}^n)$ restricted to Ω, that is $C_0^\infty(\mathbf{R}^n)|\Omega \doteq \{f|\Omega : f \in C_0^\infty(\mathbf{R}^n)\}$. For Sobolev spaces of functions on Lipschitz domains we then have the following important result, of which a proof can be found in [58, p67].

Theorem 3.5.3 *If $\Omega \subseteq \mathbf{R}^2$ is of class $C^{0,1}$, then $C_0^\infty(\mathbf{R}^2)|\Omega$ is dense in $\mathcal{H}^l(\Omega)$ for all $l \in \mathbf{N}_0$.*

This theorem allows us to approximate the distributions in $\mathcal{H}^l(\Omega)$ by the extremely well behaved functions in $C_0^\infty(\mathbf{R}^2)$. Since these approximations are crucial in our analysis, as well as for various regularity results in the literature, the concept of the Lipschitz property defined earlier is well-motivated.

3.6 Restrictions and Trivial Extensions

Restrictions and extensions of functions are going to be important on several occasions in the following sections. In the treatment of the nonrectangularly shaped components of the continuity set we will need to restrict functions to the domains and the ranges of the Lipschitz charts, as well as to extend functions, which are merely defined on these sets to the entire image domain B. The decomposition of the continuity set by the edges is one such example. Another one occurs, when we consider perturbations of the edges. Then each one of the components of the continuity set will be decomposed into two subcomponents, one which "encloses" the perturbation, and one which is "untouched" by the perturbation. In each of these cases it will further be necessary, that the restricted or extended function remains in the same type of Sobolev space. More precisely, if $W \subseteq V \subseteq \mathbf{R}^n$, we want to be able to think of elements in $\mathcal{H}^l(W)$ as elements in $\mathcal{H}^l(V)$ and vice versa. (This is to say, that we are interested in the embedding of $\mathcal{H}^l(W)$ in $\mathcal{H}^l(V)$ and the projection of $\mathcal{H}^l(V)$ on $\mathcal{H}^l(W)$.) In this section we collect the simpler results of this kind, which will be necessary for the continuation. In the next section we will consider a more complicated case of a nontrivial extension.

For any function $f : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}^l$ we define its *trivial extension* $\tilde{f}$, (also written $f^\sim$), by

$$\tilde{f}(x) \doteq \begin{cases} f(x) & \text{if } x \in \Omega \\ 0 & \text{if } x \in \complement\Omega \end{cases}$$

The next proposition lists a number of simple useful facts about the supports, defined as in section 3.3 of various functions.

Proposition 3.6.1 *For any functions $f, g : V \subseteq \mathbf{R}^l \to \mathbf{R}^k$, $\Phi : U \subseteq \mathbf{R}^n \to V$ and any set $W \subseteq \mathbf{R}^l$ the following are true:*

(i) $f^{-1}(\complement\{0\}) \subseteq \underline{f} = V \cap \underline{\tilde{f}} \subseteq \tilde{f} = \overline{f^{-1}(\complement\{0\})} \subseteq \overline{V}$

(ii) $\underline{\widetilde{f|W}} \subseteq \overline{W} \cap \underline{\tilde{f}} \subseteq \overline{W} \cap \overline{V}$

(iii) $\underline{\tilde{f}|W} = W \cap \underline{\widetilde{f|W}} \subseteq W \cap \underline{\tilde{f}} \subseteq W \cap \overline{V}$

(iv) $\underline{f|W} = V \cap W \cap \underline{\widetilde{f|W}} = V \cap \underline{\tilde{f}|W} \subseteq W \cap \underline{f} \subseteq W \cap V$

(v) $\underline{\widetilde{fg}} \subseteq \underline{\tilde{f}} \cap \underline{\tilde{g}}$

(vi) $\underline{fg} \subseteq \underline{f} \cap \underline{g}$

(vii) $\underline{\widetilde{f \circ \Phi}} \subseteq \overline{\Phi^{-1}(\underline{f})}$

Proof: Claim (i) follows straight from the definition of $\underline{f}$ together with the observation that

$$\underline{\tilde{f}} \doteq \mathbf{R}^l \cap \overline{\tilde{f}^{-1}(\mathbb{C}\{0\})} = \overline{f^{-1}(\mathbb{C}\{0\})}$$

For $W \subseteq \mathbf{R}^l$ we then have

$$\underline{\widetilde{f|W}} = \overline{W \cap f^{-1}(\mathbb{C}\{0\})} \subseteq \overline{W} \cap \underline{\tilde{f}} \subseteq \overline{W} \cap \overline{V}$$

Thus

$$\underline{\tilde{f}|W} = W \cap \underline{\widetilde{f|W}}$$

and

$$\underline{f|W} = V \cap W \cap \underline{\widetilde{f|W}} \subseteq V \cap W \cap \underline{\tilde{f}} = W \cap \underline{f}$$

Hence (ii), (iii) and (iv) follow. Moreover

$$\underline{\widetilde{fg}} = \overline{f^{-1}(\mathbb{C}\{0\}) \cap g^{-1}(\mathbb{C}\{0\})} \subseteq \underline{\tilde{f}} \cap \underline{\tilde{g}}$$

so

$$\underline{fg} = V \cap \underline{\widetilde{fg}} \subseteq V \cap \underline{\tilde{f}} \cap \underline{\tilde{g}} = \underline{f} \cap \underline{g}$$

which proves (v) and (vi). Finally

$$\underline{\widetilde{f \circ \Phi}} = \overline{\Phi^{-1}(f^{-1}(\mathbb{C}\{0\}))} \subseteq \overline{\Phi^{-1}(\underline{f})}$$

establishes (vii). ■

For functions of compact support we also have the following useful characterizations.

Proposition 3.6.2 *Let $V \subseteq \mathbf{R}^n$ be an open set, and consider a function $f : V \to \mathbf{R}^l$. Then the following are equivalent:*

(i) $\exists$ *a set* K, *such that* $f^{-1}(\mathbb{C}\{0\}) \subseteq K \subset\subset V$

(ii) $\underline{f} = \underline{\tilde{f}} \subset\subset V$

(iii) $\underline{\widetilde{f|W}} \subset\subset V \quad \forall$ *open sets* $W \supseteq V$

(iv) $\underline{\tilde{f}|W} \subset\subset V \quad \forall$ *open sets* $W \supseteq V$

Proof: Using the claims (i) and (iii) in proposition 3.6.1 we show, that (i) $\Rightarrow$ (ii), (ii) $\Rightarrow$ (iii) & (iv), (iii) $\Rightarrow$ (i) and (iv) $\Rightarrow$ (i): Suppose (i) is true. Since K is compact

$$\tilde{f} = \overline{f^{-1}(\mathbb{C}\{0\})} \subseteq K \subseteq V$$

and thus

$$\underline{f} = V \cap \underline{\tilde{f}} = \tilde{f}$$

Since $\tilde{f}$ is closed and $\tilde{f} \subseteq K \subset\subset V$, we therefore conclude, that (i) $\Rightarrow$ (ii). Next suppose (ii) is true. Then for $W \supseteq V$

$$\underline{\widetilde{f|W}} = \overline{W \cap f^{-1}(\mathbb{C}\{0\})} = \tilde{f} \subset\subset V$$

and thus

$$\underline{\tilde{f}|W} = W \cap \underline{\widetilde{f|W}} = \tilde{f} \subset\subset V$$

Hence (ii) $\Rightarrow$ (iii) & (iv). Finally since $\widetilde{f|V} = \tilde{f}$, (iii) $\Rightarrow$ (i), and since $\tilde{f}|V = f$, (iv) $\Rightarrow$ (i). ■

It is worth while noticing, that if $\underline{f} \subset\subset V$ or $\underline{\tilde{f}} \subset\subset V$, then (i) is satisfied, and hence $\underline{f} = \underline{\tilde{f}}$.

The next couple of results are useful, when we consider restrictions of functions in Sobolev spaces.

Proposition 3.6.3 *Let* V *and* W *be open sets in* $\mathbb{R}^n$, *and let* $\alpha \in \mathbb{N}_0^n$. *If* $f, D^\alpha f \in L_1^{\mathrm{loc}}(V)$, *then* $D^\alpha(f|W) = (D^\alpha f)|W$.

Proof: Let $\varphi \in C_0^\infty(V \cap W)$. By proposition 3.6.2 (ii) and (iv) $\underline{\tilde{\varphi}|V} \subset\subset V \cap W$. Moreover $\tilde{\varphi}|V \equiv \varphi$ on $V \cap W$ and $\tilde{\varphi}|V \equiv 0$ on $V \setminus \underline{\tilde{\varphi}|V} \supset V \setminus W$, with $V \cap W$ and $V \setminus \underline{\tilde{\varphi}|V}$ both open. Thus $\tilde{\varphi}|V \in C_0^\infty(V)$. If f and $D^\alpha f$ are locally integrable

$$D^\alpha(f|W)(\varphi) \doteq |-1|^{|\alpha|}\langle f|W, D^\alpha\varphi\rangle_{V\cap W} = |-1|^{|\alpha|}\int_{V\cap W} f D^\alpha\varphi \, dx$$

and

$$(D^\alpha f)|W\,(\varphi) \doteq \langle (D^\alpha f)|W, \varphi\rangle_{V\cap W} = \int_{V\cap W} D^\alpha f\varphi\,dx$$

are both well-defined. Furthermore by the properties of $\tilde{\varphi}|V$

$$\begin{aligned} |-1|^{|\alpha|}\int_{V\cap W} fD^\alpha\varphi\,dx &= \\ &= |-1|^{|\alpha|}\int_V fD^\alpha(\tilde{\varphi}|V)\,dx \\ &= \langle D^\alpha f, \tilde{\varphi}|V\rangle_V \\ &= \int_V D^\alpha f\tilde{\varphi}|V\,dx \\ &= \int_{V\cap W} D^\alpha f\varphi\,dx \end{aligned}$$

■

We remark, that this proposition is not quite as obvious, as it might seem to be at first sight. One can for example easily find examples of functions $f \in L_1^{\mathrm{loc}}(V)$, for which $D^\alpha f$ is not even a function, in which case $(D^\alpha f)|W$ does not make much sense.

Proposition 3.6.3 is useful, because it allows us to replace the functional $D^\alpha(f|W)$ by the function $(D^\alpha f)|W$ in computations. The proof of the next result is a good example thereof.

Proposition 3.6.4 *Let V and W be open sets in $\mathbf{R}^n$, and let $f \in \mathcal{H}^l(V)$. Then $f|W \in \mathcal{H}^l(V\cap W)$ and $\|f|W\|_{\mathcal{H}^l(V\cap W)} \leq \|f\|_{\mathcal{H}^l(V)}$.*

Proof: Since $f \in \mathcal{H}^l(V)$, it follows, that $D^\alpha f \in L_2(V) \subseteq L_1^{\mathrm{loc}}(V)\ \ \forall\alpha \in \mathbf{N}_0^n$ with $|\alpha| \leq l$. Hence by proposition 3.6.3

$$\|D^\alpha(f|W)\|_{L_2(V\cap W)} = \|(D^\alpha f)|W\|_{L_2(V\cap W)} \leq \|D^\alpha f\|_{L_2(V)} \qquad \forall\alpha \in \mathbf{N}_0^n \text{ with } |\alpha| = 1$$

■

We end this section with a few results for (restrictions of) trivial extensions of functions in Sobolev spaces. For the proof of the first of these results we need the following theorem, which can be found in [56, p28].

Theorem 3.6.5 *Let V be an open set in $\mathbf{R}^n$. For every set $K \subset\subset V$ there exists an open neighborhood $W_K \supseteq V$ of K and a function $\psi_K \in C_0^\infty(V)$, such that $\psi_K(V) \subseteq [0,1]$ and $\psi_K(W_K) = \{1\}$.*

Theorem 3.6.6 *Let V and W be open sets in $\mathbf{R}^n$. If $f \in \mathcal{H}^l(V \cap W)$ and $\underline{\tilde{f}} \subset\subset V$, then $\tilde{f}|W \in \mathcal{H}^l(W)$ and $\|\tilde{f}|W\|_{\mathcal{H}^l(W)} = \|f\|_{\mathcal{H}^l(V\cap W)}$.*

Proof: Let $\varphi \in C_0^\infty(W)$, and let $\alpha \in \mathbf{N}_0^n$ with $|\alpha| \le l$. Since $\underline{\tilde{f}}$ and $\underline{\tilde{\varphi}}$ are compact, $\underline{\tilde{f}} \cap \underline{\tilde{\varphi}} \subset\subset V \cap W$. Thus by theorem 3.6.5 $\exists$ an open neighborhood $U \subset V \cap W$ of $\underline{\tilde{f}} \cap \underline{\tilde{\varphi}}$ and a function $\psi \in C_0^\infty(V \cap W)$, such that $\psi(V \cap W) \subseteq [0,1]$ and $\psi(U) = \{1\}$. Thus $fD^\alpha\varphi\psi \equiv fD^\alpha\varphi$ on U. Since $f \equiv 0$ on $(V \cap W) \setminus \underline{\tilde{f}}$ and $\varphi \equiv 0$ on the open set $(V \cap W) \setminus \underline{\tilde{\varphi}}$, we also have that $fD^\alpha\varphi \equiv 0$ on $(V \cap W) \setminus (\underline{\tilde{f}} \cap \underline{\tilde{\varphi}}) \supseteq (V \cap W) \setminus U$. Hence $fD^\alpha\varphi\psi \equiv fD^\alpha\varphi$ on $V \cap W$, and we obtain

$$\begin{aligned}
\left|\langle D^\alpha(\tilde{f}|W), \varphi\rangle_W\right| &= \\
&= \left|\int_W \tilde{f} D^\alpha\varphi \, dx\right| \\
&= \left|\int_{V\cap W} f D^\alpha\varphi \, dx\right| \\
&= \left|\int_{V\cap W} f D^\alpha\varphi\psi \, dx\right| \\
&= \left|\int_{V\cap W} f D^\alpha(\varphi\psi)\, dx - \sum_{\substack{\beta\in\mathbf{N}_0^n\setminus\{0\} \\ |\beta|\le|\alpha|}} \int_{V\cap W} f D^{\alpha-\beta}\varphi D^\beta\psi \, dx\right| \qquad (3.26)
\end{aligned}$$

Since $fD^{\alpha-\beta}\varphi \equiv 0$ on $(V \cap W) \setminus U$ and $\psi \equiv 1$ on the open set U, the sum over β vanishes. Moreover by proposition 3.6.1 (v)

$$\overline{(\varphi|V\ \psi)^{-1}(\mathbb{C}\{0\})} = \underline{(\varphi|V\ \psi)^\sim} \subseteq \underline{\tilde{\psi}} = \underline{\psi} \subset\subset V \cap W$$

Hence by proposition 3.6.2 $\underline{\varphi|V\ \psi} \subset\subset V \cap W$, and therefore $\varphi|V\ \psi \in C_0^\infty(V \cap W)$. From (3.26) we then see that

$$\left|\langle D^\alpha(\tilde{f}|W), \varphi\rangle_W\right| = |\langle D^\alpha f, \varphi|V\ \psi\rangle_{V\cap W}| \le \|D^\alpha f\|_{L_2(V\cap W)}\|\varphi|V\ \psi\|_{L_2(V\cap W)}$$

Since $D^\alpha f \in L_2(V \cap W)$ and

$$\|\varphi|V\ \psi\|^2_{L_2(V\cap W)} = \int_{V\cap W} |\varphi|^2|\psi|^2\, dx \le \int_W |\varphi|^2\, dx = \|\varphi\|^2_{L_2(W)}$$

this shows, that $D^\alpha(\tilde{f}|W) \in L_2(W)$, and that $\|D^\alpha(\tilde{f}|W)\|_{L_2(W)} \le \|D^\alpha f\|_{L_2(V\cap W)}$ $\forall \alpha \in \mathbf{N}_0^n$ with $|\alpha| \le l$. Hence $\tilde{f}|W \in \mathcal{H}^l(W)$ and $\|\tilde{f}|W\|_{\mathcal{H}^l(W)} \le \|f\|_{\mathcal{H}^l(V\cap W)}$. Since $f = (\tilde{f}|W)|V$, the theorem then follows from proposition 3.6.4. ■

Corollary 3.6.7 *Let $V \subseteq W$ be two open sets in $\mathbb{R}^n$. If $f \in \mathcal{H}^l(V)$ and $\underline{f} \subset\subset V$, then $\tilde{f}|W \in \mathcal{H}^l(W)$ and $\|\tilde{f}|W\|_{\mathcal{H}^l(W)} = \|f\|_{\mathcal{H}^l(V)}$.*

Proof: By proposition 3.6.2 $\underline{\tilde{f}} = \underline{f} \subset\subset V = V \cap W$. Hence the corollary follows from the previous theorem. ■

Proposition 3.6.8 *Let $\mathcal{V}$ be the collection of all the components of an open set $W \subseteq \mathbb{R}^n$, and let $f_V \in \mathcal{H}^l(V)$, $V \in \mathcal{V}$, for some fixed $l \in \mathbb{N}_0$. Assume that*

$$\sum_{V \in \mathcal{V}} \|f_V\|^2_{\mathcal{H}^l(V)} < \infty \tag{3.27}$$

Define the function

$$f \doteq \sum_{V \in \mathcal{V}} \tilde{f}_V|W$$

Then $f \in \mathcal{H}^l(W)$ and

$$\|f\|_{\mathcal{H}^l(W)} \leq \sqrt{\sum_{V \in \mathcal{V}} \|f_V\|_{\mathcal{H}^l(V)}}$$

Proof: Since W is open, so are its components. Thus $\mathcal{H}^l(V)$ is well-defined $\forall V \in \mathcal{V}$. Let $\varphi \in C_0^\infty(W)$. We claim, that $\varphi|V \in C_0^\infty(V)$ $\forall V \in \mathcal{V}$. To see this, let $V \in \mathcal{V}$. Since $\varphi|V \equiv \varphi$ on the open set V, we have $D^\alpha(\varphi|V) = (D^\alpha\varphi)|V$ $\forall \alpha \in \mathbb{N}_0^n$, and thus $\varphi|V \in CinV$. Let $\mathcal{U}$ be an open covering of $V \cap \underline{\varphi}$. Then $\mathcal{U} \cup (\mathcal{V} \setminus \{V\})$ is an open covering of the compact set $\underline{\varphi}$. Hence $\exists$ a finite collection $\mathcal{W} \subseteq \mathcal{U}$ of open sets, such that $\mathcal{W} \cup (\mathcal{V} \setminus \{V\})$ is an open covering of $\underline{\varphi}$. Since $U \cap V \cap \underline{\varphi} = \emptyset$ $\forall U \in \mathcal{V} \setminus \{V\}$, this implies, that $\mathcal{W}$ is a finite open covering of $V \cap \underline{\varphi}$, which shows, that $V \cap \underline{\varphi}$ is compact. Thus

$$(\varphi|V)^{-1}(\mathbb{C}\{0\}) = V \cap \varphi^{-1}(\mathbb{C}\{0\}) \subseteq V \cap \underline{\varphi} \subset\subset V$$

By proposition 3.6.2 it then follows, that $\underline{\varphi|V} \subset\subset V$, which proves the claim. Next we observe, that $\tilde{f}_V|W \in L_2(W)$ $\forall V \in \mathcal{V}$, and therefore $f \in L_2(W) \subseteq L_1^{\mathrm{loc}}(W)$. Hence $D^\alpha f \in C_0^\infty(W)^*$ $\forall \alpha \in \mathbb{N}_0^n$, and if $|\alpha| \leq l$, by the Cauchy-Schwarz inequalities we have

$$\begin{aligned} |D^\alpha f(\varphi)| &= \\ &= \left| \int_W f D^\alpha \varphi \, dx \right| \\ &= \left| \sum_{V \in \mathcal{V}} \int_V f_V D^\alpha(\varphi|V) \, dx \right| \end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{V\in\mathcal{V}} |\langle D^\alpha f_V, \varphi|V\rangle_V| \\
&\leq \sum_{V\in\mathcal{V}} \|D^\alpha f_V\|_{L_2(V)} \|\varphi|V\|_{L_2(V)} \\
&\leq \sqrt{\sum_{V\in\mathcal{V}} \|D^\alpha f_V\|^2_{L_2(V)}} \sqrt{\sum_{V\in\mathcal{V}} \|\varphi|V\|^2_{L_2(V)}} \\
&= \sqrt{\sum_{V\in\mathcal{V}} \|D^\alpha f_V\|^2_{L_2(V)}} \|\varphi\|_{L_2(W)}
\end{aligned}
$$

From (3.27) it then follows, that $D^\alpha f \in L_2(W)$ and

$$
\|D^\alpha f\|^2_{L_2(W)} \leq \sum_{V\in\mathcal{V}} \|D^\alpha f_V\|^2_{L_2(V)} \qquad \forall \alpha \in \mathbb{N}_0^n \text{ with } |\alpha| \leq l
$$

Thus $f \in \mathcal{H}^l(W)$, and by changing order of summation we obtain

$$
\|f\|^2_{\mathcal{H}^l(W)} = \sum_{\substack{\alpha\in\mathbb{N}_0^n \\ |\alpha|\leq l}} \|D^\alpha f\|^2_{L_2(W)} \leq \sum_{\substack{\alpha\in\mathbb{N}_0^n \\ |\alpha|\leq l}} \sum_{V\in\mathcal{V}} \|D^\alpha f_V\|^2_{L_2(V)} = \sum_{V\in\mathcal{V}} \|f_V\|^2_{\mathcal{H}^l(V)}
$$

■

From the proof above and the propositions 3.6.3 and 3.6.4 it is easy to show, that the map

$$
\Pi : \mathcal{H}^l(W) \to \left\{ \langle f_V\rangle_{V\in\mathcal{V}} \in \prod_{V\in\mathcal{V}} \mathcal{H}^l(V) : \sum_{V\in\mathcal{V}} \|f_V\|^2_{\mathcal{H}^l(V)} < \infty \right\} : f \mapsto \langle f|V\rangle_{V\in\mathcal{V}}
$$

is an isometric isomorphism with inverse

$$
\Pi^{-1} : \langle f_V\rangle_{V\in\mathcal{V}} \mapsto \sum_{V\in\mathcal{V}} \tilde{f}_V|V
$$

but this is more information, than we will need.

3.7 Extension Operators on Lipschitz Domains

In the previous section we examined some trivial extensions of functions in $\mathcal{H}^l(V)$ to some larger domain W. We gave conditions, under which these extensions are functions in $\mathcal{H}^l(W)$, and found bounds on the $\mathcal{H}^l(W)$-norm of the extension in terms of the $\mathcal{H}^l(V)$-norm of the original function. However, it was always the case, that the function $f \in \mathcal{H}^l(V)$ under consideration vanished on a neighborhood of $\partial V \setminus \partial W$ (in W). If this condition is not satisfied, it is obvious, that a trivial extension in general does indeed not define a function in $\mathcal{H}^l(W)$. Thus in the general case a more sophisticated extension method is needed.

Definition 3.7.1 *Let $V \subseteq \mathbf{R}^n$ be an open set. We say, that $g \in \mathcal{H}^l(\mathbf{R}^n)$ is an* extension *of $f \in \mathcal{H}^l(V)$, if $g = f$ a.e. on V. We say, that an operator $P_V \in \mathcal{L}(\mathcal{H}^l(V), \mathcal{H}^l(\mathbf{R}^n))$ is an* extension operator, *if $P_V(f)$ is an extension of f $\forall f \in \mathcal{H}^l(V)$.*

The existence of an extension operator P_V of this kind for Lipschitz domains is well known in the mathematical literature. See for example [55, 58, 56], where the extension operators due to Calderón and Nikol'skiĭ for fixed domains of class $C^{0,1}$ can be found. For our purpose however, we will need to construct a whole collection $\mathcal{P} \doteq \{P_{\Omega_h}\}_{h \in]0,H[}$ of such extension operators for a corresponding collection $\{\Omega_h\}_{h \in]0,H[}$ of domains, and in such a way, that the operators in $\mathcal{P}$ are uniformly bounded. For this reason we will go through the procedure of constructing an extension operator $P_\Omega \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))$ for a bounded open set $\Omega \subseteq \mathbf{R}^2$. This will give us an explicit upper bound on $\|P_\Omega\|_{\mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))}$, from which the necessary uniform boundedness can be determined.

The first step is to consider the simplified case, when the domain is just the half-space $\mathbf{R}^2_+$. A function on this domain can easily be extended to $\mathbf{R}^2$, by simply "reflecting" it in the boundary $\mathbf{R}^2_0$. Thus we define the maps

$$R_\pm : \mathbf{R}^2 \to \mathbf{R}^2 : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \begin{bmatrix} x_1 \\ \pm x_2 \end{bmatrix} \tag{3.28}$$

and then an operator on $\mathcal{H}^1(\mathbf{R}^2_+)$ by

$$P_{\mathbf{R}^2_+}(f)(x) \doteq f(R_\pm(x)) \qquad x \in \mathbf{R}^2_\pm \tag{3.29}$$

Since $\mathbf{R}^2_0$ has zero measure in $\mathbf{R}^2$, it is immediately clear, that $P_{\mathbf{R}^2_+}(f) \in L_2(\mathbf{R}^2)$, but more than so, as the next two lemmas will show, $P_{\mathbf{R}^2_+}$ is indeed an extension operator in the sense of definition 3.7.1.

Lemma 3.7.2 *$C_0^\infty(\mathbf{R}^2)|\mathbf{R}^2_+$ is dense in $\mathcal{H}^l(\mathbf{R}^2_+)$.*

A proof of this lemma can be found in [56, p46].

Lemma 3.7.3 *Let the operator $P_{\mathbf{R}^2_+} : \mathcal{H}^1(\mathbf{R}^2_+) \to L_2(\mathbf{R}^2)$ be defined as in (3.29). Then $P_{\mathbf{R}^2_+} \in \mathcal{L}(\mathcal{H}^1(\mathbf{R}^2_+), \mathcal{H}^1(\mathbf{R}^2))$ with norm $\|P_{\mathbf{R}^2_+}\|_{\mathcal{L}(\mathcal{H}^1(\mathbf{R}^2_+), \mathcal{H}^1(\mathbf{R}^2))} = \sqrt{2}$ and*

$$\underline{P_{\mathbf{R}^2_+}(f)} = \underline{\tilde{f}} \cup R_-(\underline{\tilde{f}}) \qquad \forall f \in \mathcal{H}^1(\mathbf{R}^2_+)$$

Furthermore if f is Lipschitz continuous with Lipschitz constant L, the same is true for $P_{\mathbf{R}^2_+}(f)$.

Proof: For the proof, that $P_{\mathbf{R}^2_+} \in \mathcal{L}(\mathcal{H}^1(\mathbf{R}^2_+), \mathcal{H}^1(\mathbf{R}^2))$ we refer to [56, p46]. Let $f \in C_0^\infty(\mathbf{R}^2)|\mathbf{R}^2_+$. Then for $\alpha \in \mathbf{N}_0^2$ we have

$$\begin{aligned}
&\|D^\alpha P_{\mathbf{R}^2_+}(f)\|^2_{L_2(\mathbf{R}^2)} = \\
&\quad = \int_{\mathbf{R}^2_-} |D^\alpha (f \circ R_-)|^2 \, dx + \int_{\mathbf{R}^2_+} |D^\alpha (f \circ R_+)|^2 \, dx \\
&\quad = \int_{\mathbf{R}^2_+} |D^\alpha f|^2 \, dx + \int_{\mathbf{R}^2_+} |D^\alpha f|^2 \, dx \\
&\quad = 2\|D^\alpha f\|^2_{L_2(\mathbf{R}^2_+)}
\end{aligned}$$

Hence $\|P_{\mathbf{R}^2_+}(f)\|_{\mathcal{H}^1(\mathbf{R}^2)} = \sqrt{2}\|f\|_{\mathcal{H}^1(\mathbf{R}^2_+)}$, and thus by lemma 3.7.2

$$\|P_{\mathbf{R}^2_+}\|_{\mathcal{L}(\mathcal{H}^1(\mathbf{R}^2_+), \mathcal{H}^1(\mathbf{R}^2))} = \sqrt{2}$$

Since R_+ is the identity map, and R_- is a homeomorphism equal to its inverse, by proposition 3.6.1 we further have that

$$\begin{aligned}
&\underline{P_{\mathbf{R}^2_+}(f)} = \\
&\quad = \overline{\left[\mathbf{R}^2_+ \cap R_+^{-1}(f^{-1}(\mathbb{C}\{0\}))\right] \cup \left[\mathbf{R}^2_- \cap R_-^{-1}(f^{-1}(\mathbb{C}\{0\}))\right]} \\
&\quad = \overline{f^{-1}(\mathbb{C}\{0\})} \cup \overline{R_-^{-1}(f^{-1}(\mathbb{C}\{0\}))} \\
&\quad = \underline{\tilde{f}} \cup R_-(\underline{\tilde{f}}) \qquad \forall f \in \mathcal{H}^1(\mathbf{R}^2_+)
\end{aligned}$$

Finally if $f \in \mathcal{H}^1(\mathbf{R}^2_+)$ is Lipschitz continuous with Lipschitz constant L, we have

$$\begin{aligned}
&|P_{\mathbf{R}^2_+}(f)(y) - P_{\mathbf{R}^2_+}(f)(x)| = \\
&\quad = |f(y_1, |y_2|) - f(x_1, |x_2|)| \\
&\quad \le L\sqrt{|y_1 - x_1|^2 + ||y_2| - |x_2||^2} \\
&\quad \le L\sqrt{|y_1 - x_1|^2 + |y_2 - x_2|^2} \\
&\quad = L\|y - x\| \qquad \forall x, y \in \mathbf{R}^2
\end{aligned}$$

■

Next we consider a (bounded open) domain $\Omega \subseteq B$ of class $C^{0,1}$. Let $\mathcal{A} \doteq \{\Phi_m : Q_m \to U_m\}_{m=1}^M$ be an atlas of Lipschitz charts satisfying the conditions (i)–(iii) in definition 3.5.1, and let U_0 be an open set with the property that $\Omega \setminus \bigcup_{m=1}^M U_m \subseteq U_0 \subseteq \Omega$. †

†This requirement is always satisfied for $U_0 = \Omega$. The proof in appendix A, however, relies on the possibility of choosing $U_0 \subsetneq \Omega$.

Then $\mathcal{U} \doteq \{U\}_{m=0}^{M}$ is an open covering of sets in $\mathbf{R}^2$ of the compact set $\overline{\Omega}$. Hence there exists a *C^∞-partition of unity for $\overline{\Omega}$ subordinate to $\mathcal{U}$*, that is a collection $\Psi \doteq \{\psi_m\}_{m=0}^{M}$ of functions with the following properties:

(i) $\psi_m \in C_0^\infty(\mathbf{R}^2), \ m = 0, \ldots, M$

(ii) $\psi_m(\mathbf{R}^2) \subseteq [0,1], \quad m = 0, \ldots, M$

(iii) $\underline{\psi_m} \subset\subset U_m, \quad m = 0, \ldots, M$

(iv) $\sum_{m=0}^{M} \psi_m(x) = 1 \quad \forall x \in \Omega$

A proof of this fact can be found in for example [59, p63]. Using this C^∞-partition of unity together with theorem 3.5.3 the construction of an extension operator $P_\Omega \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))$ can be reduced to the problem of extending functions of the form $\psi_m|\Omega\, z$, $m = 0, \ldots, M$, where $z \in C_0^\infty(\mathbf{R}^2)|\Omega \subseteq \mathcal{H}^1(\Omega)$ to $\mathbf{R}^2$ in such a way, that the extension belongs to $\mathcal{H}^1(\mathbf{R}^2)$. For convenience we define the constant

$$M_\Psi \doteq \bigvee_{m=0}^{M} \bigvee_{l=0}^{1} \sup_{x \in \mathbf{R}^2} \|\psi_m^{(l)}(x)\| \tag{3.30}$$

and note that $M_\Psi < \infty$. We then have the following useful fact about the products $\psi_m|\Omega\, z$, $m = 0, \ldots, M$.

Fact 3.7.4 *If $\psi \in \Psi$ and $z \in C_0^\infty(\mathbf{R}^2)|\Omega$, then $\psi|\Omega\, z \in \mathcal{H}^1(\Omega)$ and $\|\psi|\Omega\, z\|_{\mathcal{H}^1(\Omega)} \leq \sqrt{5} M_\Psi \|z\|_{\mathcal{H}^1(\Omega)}$.*

Proof: Since

$$\|\psi|\Omega\, z\|_{L_2(\Omega)}^2 = \int_\Omega |\psi z|^2 \, dx \leq M_\Psi^2 \|z\|_{L_2(\Omega)}^2$$

and

$$\begin{aligned}
&\|D_k(\psi|\Omega\, z)\|_{L_2(\Omega)}^2 = \\
&\quad = \int_\Omega |D_k\psi z + \psi D_k z|^2 \, dx \\
&\quad \leq \int_\Omega 2M_\Psi^2 \left(|z|^2 + |D_k z|^2\right) dx \\
&\quad = 2M_\Psi^2 \left(\|z\|_{L_2(\Omega)}^2 + \|D_k z\|_{L_2(\Omega)}^2\right) \qquad k = 1, 2
\end{aligned}$$

we have

$$\|\psi|\Omega\, z\|^2_{\mathcal{H}^1(\Omega)} \leq M_\Psi^2 \left(5\|z\|^2_{L_2(\Omega)} + 2\sum_{k=1}^{2} \|D_k z\|^2_{L_2(\Omega)}\right) \leq 5M_\Psi^2 \|z\|^2_{\mathcal{H}^1(\Omega)}$$

■

The case, when $m = 0$, is easy. As $\psi_0|\Omega\, z$ vanishes on a neighborhood of $\partial\Omega$, one should expect the trivial extension to be good enough. This is, as we shall see next, indeed the case. We therefore define the linear operator P_0 on $\mathcal{H}^1(\Omega)$ by

$$P_0(z) \doteq \psi_0 \tilde{z} \qquad z \in \mathcal{H}^1(\Omega) \tag{3.31}$$

Fact 3.7.5 *$P_0 \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))$ and $\|P_0\|_{\mathcal{L}(\mathcal{H}^1(\Omega),\mathcal{H}^1(\mathbf{R}^2))} \leq \sqrt{5} M_\Psi$.*

Proof: Let $z \in C_0^\infty(\mathbf{R}^2)|\Omega$. Then by fact 3.7.4 above $\psi_0|\Omega\, z \in \mathcal{H}^1(\Omega)$ and

$$\|\psi_0|\Omega\, z\|_{\mathcal{H}^1(\Omega)} \leq \sqrt{5} M_\Psi \|z\|_{\mathcal{H}^1(\Omega)} \tag{3.32}$$

Furthermore

$$(\psi_0|\Omega\, z)^{-1}(\mathbb{C}\{0\}) \subseteq \psi_0^{-1}(\mathbb{C}\{0\}) \subseteq \underline{\psi_0} \subset\subset U_0 \subseteq \Omega$$

so by proposition 3.6.2 $\underline{\psi_0|\Omega\, z} \subset\subset \Omega$. Hence by corollary 3.6.7 $P_0(z) = (\psi_0|\Omega\, z)^\sim \in \mathcal{H}^1(\mathbf{R}^2)$ and

$$\|P_0(z)\|_{\mathcal{H}^1(\mathbf{R}^2)} = \|\psi_0|\Omega\, z\|_{\mathcal{H}^1(\Omega)} \tag{3.33}$$

Since Ω is of class $C^{0,1}$, $C_0^\infty(\mathbf{R}^2)|\Omega$ is dense in $\mathcal{H}^1(\Omega)$ by theorem 3.5.3. The fact thus follows by (3.32), (3.33) and the linearity of P_0. ■

It remains to consider $m = 1, \ldots, M$. In this case a nontrivial extension and a bit more work are required. For $m = 1, \ldots, M$ we therefore define the sets $Q_{m+} \doteq \mathbf{R}^2_+ \cap Q_m$ and $U_{m+} \doteq \Omega \cap U_m$ and a linear operator P_m on $\mathcal{H}^1(\Omega)$ by

$$P_m(z) \doteq \left[P_{\mathbf{R}^2_+}\left([(\psi_m \tilde{z})|U_{m+} \circ \Phi_m|Q_{m+}]^\sim | \mathbf{R}^2_+\right) \middle| Q_m \circ \Phi_m^{-1}\right]^\sim \qquad z \in \mathcal{H}^1(\Omega) \tag{3.34}$$

For convenience we also define the constant $L_{\mathcal{A}}$ to be the maximum of the Lipschitz constants of all the charts $\Phi_1, \ldots, \Phi_M$ and their inverses $\Phi_1^{-1}, \ldots, \Phi_M^{-1}$. We then have the following fact.

Fact 3.7.6 *For each* $m \in \{1, \ldots, M\}$ $P_m \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))$ *and*

$$\|P_m\|_{\mathcal{L}(\mathcal{H}^1(\Omega),\mathcal{H}^1(\mathbf{R}^2))} \leq \sqrt{10}(1 + L_{\mathcal{A}}^2)L_{\mathcal{A}}^2 M_\Psi \tag{3.35}$$

Proof: Throughout this proof we drop the subscript m. Let $z \in C_0^\infty(\mathbf{R}^2)|\Omega$, and let

$$f_1 \doteq (\psi \tilde{z})|U_+ \tag{3.36}$$

Since $U_+ \subseteq \Omega$, $f_1 = (\psi|\Omega\, z)|U_+$. Hence from fact 3.7.4 and proposition 3.6.4 we see that $f_1 \in \mathcal{H}^1(U_+)$ and

$$\|f_1\|_{\mathcal{H}^1(U_+)} \leq \sqrt{5} M_\Psi \|z\|_{\mathcal{H}^1(\Omega)} \tag{3.37}$$

Since $\psi \in C_0^\infty(\mathbf{R}^2)$ and $z \in C_0^\infty(\mathbf{R}^2)|\Omega$, f_1 is also Lipschitz continuous. Let

$$f_2 \doteq f_1 \circ \Phi|Q_+ \tag{3.38}$$

It then follows from proposition 3.4.6, that $f_2 \in \mathcal{H}^1(Q_+)$ and

$$\|f_2\|_{\mathcal{H}^1(Q_+)} \leq \sqrt{1 + L_{\mathcal{A}}^2} L_{\mathcal{A}} \|f_1\|_{\mathcal{H}^1(U_+)} \tag{3.39}$$

Next from proposition 3.6.1 (iv) we note that $\underline{f_1} \subseteq \underline{\psi}$. Since $\underline{\psi} \subset\subset U$ and Φ is a homeomorphism, from claim (vii) of the same proposition it thus follows that

$$\underline{\tilde{f}_2} \subseteq \overline{\Phi^{-1}(\underline{f_1})} \subseteq \overline{\Phi^{-1}(\underline{\psi})} = \Phi^{-1}(\underline{\psi}) \subset\subset \Phi^{-1}(U) = Q$$

Let

$$f_3 \doteq \tilde{f}_2|\mathbf{R}_+^2 \tag{3.40}$$

From theorem 3.6.6 we then have that $f_3 \in \mathcal{H}^1(\mathbf{R}_+^2)$ and

$$\|f_3\|_{\mathcal{H}^1(\mathbf{R}_+^2)} = \|f_2\|_{\mathcal{H}^1(Q_+)} \tag{3.41}$$

Since f_1 and Φ are Lipschitz continuous, so is f_2 and hence f_3. Moreover from proposition 3.6.1 (iii) we see that $\underline{\tilde{f}_3} \subseteq \underline{\tilde{f}_2} \subseteq Q$. Since $\underline{\tilde{f}_3}$ is closed, this implies that $\underline{\tilde{f}_3} \subset\subset Q$. By the symmetry of Q and the continuity of R_-, defined in (3.28), we then also have that $R_-(\underline{\tilde{f}_3}) \subset\subset R_-(Q) = Q$. Thus $\underline{\tilde{f}_3} \cup R_-(\underline{\tilde{f}_3}) \subset\subset Q$. Let

$$f_4 \doteq P_{\mathbf{R}_+^2}(f_3) \tag{3.42}$$

From lemma 3.7.3 it then follows, that $f_4 \in \mathcal{H}^1(\mathbf{R}^2)$ is Lipschitz continuous with $\underline{f_4} \subset\subset Q$ and

$$\|f_4\|_{\mathcal{H}^1(\mathbf{R}^2)} \leq \sqrt{2} \|f_3\|_{\mathcal{H}^1(\mathbf{R}_+^2)} \tag{3.43}$$

Next let

$$f_5 \doteq f_4|Q \tag{3.44}$$

Then clearly f_5 is Lipschitz continuous and moreover by proposition 3.6.4 $f_5 \in \mathcal{H}^1(Q)$ and

$$\|f_5\|_{\mathcal{H}^1(Q)} \leq \|f_4\|_{\mathcal{H}^1(\mathbf{R}^2)} \tag{3.45}$$

Then letting

$$f_6 \doteq f_5 \circ \Phi^{-1} \tag{3.46}$$

we find by proposition 3.4.6, that $f_6 \in \mathcal{H}^1(U)$ and

$$\|f_6\|_{\mathcal{H}^1(U)} \leq \sqrt{1 + L_{\mathcal{A}}^2} L_{\mathcal{A}} \|f_5\|_{\mathcal{H}^1(Q)} \tag{3.47}$$

Furthermore proposition 3.6.1 (iv) implies that $\underline{f_5} \subseteq \underline{f_4}$. Since $\underline{f_4} \subset\subset Q$ and Φ is continuous, then by claim (vii) of the same proposition

$$f_6^{-1}(\mathbb{C}\{0\}) \subseteq \underline{f_6} \subseteq \overline{\Phi(\underline{f_5})} \subseteq \overline{\Phi(\underline{f_4})} = \Phi(\underline{f_4}) \subset\subset \Phi(Q) = U$$

Hence by proposition 3.6.2 $\underline{f_6} \subset\subset U$. From corollary 3.6.7 it therefore follows, that $\tilde{f}_6 \in \mathcal{H}^1(\mathbf{R}^2)$ and

$$\|\tilde{f}_6\|_{\mathcal{H}^1(\mathbf{R}^2)} = \|f_6\|_{\mathcal{H}^1(U)} \tag{3.48}$$

Finally by examining (3.36), (3.38), (3.40), (3.42), (3.44) and (3.46) we see that

$$\tilde{f}_6 = \left[P_{\mathbf{R}^2_+}\left([(\psi\tilde{z})|U_+ \circ \Phi|Q_+]^{\sim}|\mathbf{R}^2_+\right)\Big|Q \circ \Phi^{-1}\right]^{\sim} = P(z) \tag{3.49}$$

Thus from (3.37), (3.39), (3.41), (3.43), (3.45), (3.47) and (3.48) we obtain

$$\|P(z)\|_{\mathcal{H}^1(\mathbf{R}^2)} \leq \sqrt{10}(1 + L_{\mathcal{A}}^2)L_{\mathcal{A}}^2 M_\Psi \|z\|_{\mathcal{H}^1(\Omega)}$$

Since P is linear and Ω is of class $C^{0,1}$, the fact then follows by theorem 3.5.3. ■

We conclude this section with the following important result.

Theorem 3.7.7 *Let $\Omega \subseteq \mathbf{R}^2$ be a (bounded open) Lipschitz domain and let $\mathcal{A} \doteq \{\Phi_m : Q_m \to U_m\}_{m=1}^M$ be an atlas of Lipschitz charts satisfying the conditions (i)–(iii) in definition 3.5.1. Let $U_0 \subseteq \Omega$ be an open set such that $\{U_m\}_{m=0}^M$ is an open covering of $\overline{\Omega}$, and let $\Psi \doteq \{\psi_m\}_{m=0}^M$ be a C^∞-partition of unity for $\overline{\Omega}$ subordinate to $\{U_m\}_{m=0}^M$. Then there exists an extension operator $P_\Omega \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))$ with norm*

$$\|P_\Omega\|_{\mathcal{L}(\mathcal{H}^1(\Omega),\mathcal{H}^1(\mathbf{R}^2))} \leq (M+1)\sqrt{10}(1 + L_{\mathcal{A}}^2)L_{\mathcal{A}}^2 M_\Psi$$

where $L_{\mathcal{A}}$ is the maximum of the Lipschitz constants for the charts in $\mathcal{A}$ and their inverses, and

$$M_\Psi \doteq \bigvee_{m=0}^{M} \bigvee_{l=0}^{1} \sup_{x \in \mathbf{R}^2} \|\psi_m^{(l)}(x)\|$$

Proof: Let $P_0, \ldots, P_M$ be defined as in (3.31) and (3.34) above, and define

$$P_\Omega \doteq \sum_{m=0}^{M} P_m \tag{3.50}$$

From the facts 3.7.5 and 3.7.6 we know, that $P_0, \ldots, P_M \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))$ with norms

$$\|P_m\|_{\mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))} \leq \sqrt{10}(1 + L_{\mathcal{A}}^2) L_{\mathcal{A}}^2 M_\Psi \qquad m = 0, \ldots, M$$

Hence $P_\Omega \in \mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))$ and

$$\|P_\Omega\|_{\mathcal{L}(\mathcal{H}^1(\Omega), \mathcal{H}^1(\mathbf{R}^2))} \leq (M+1)\sqrt{10}(1 + L_{\mathcal{A}}^2) L_{\mathcal{A}}^2 M_\Psi$$

It remains to show that $P_\Omega(z)$ is an extension of z $\forall z \in \mathcal{H}^1(\Omega)$. First we observe, that by (3.31)

$$P_0(z)(x) = \psi_0(x) z(x) \qquad \forall x \in \Omega \tag{3.51}$$

Next for $m = 1, \ldots, M$ we once again drop the subscript m, and use the notation in the proof of fact 3.7.6. Since $f_6 \in \mathcal{H}^1(U)$ and ψ vanishes outside $\underline{\psi} \subseteq U$, from (3.48) we see that

$$P(z)(x) = \tilde{f}_6(x) = 0 = \psi(x) z(x) \qquad \forall x \in \Omega \setminus U \tag{3.52}$$

For $x \in \Omega \cap U = U_+$ on the contrary $\Phi^{-1}(x) \in Q_+ = \mathbf{R}^2_+ \cap Q$. Thus using (3.36), (3.38), (3.40), (3.42), (3.44), (3.46) and (3.48) we get

$$\begin{aligned}
P(z)(x) &= \\
&= \tilde{f}_6(x) \\
&= f_6(x) \\
&= f_5(\Phi^{-1}(x)) \\
&= f_4|Q\,(\Phi^{-1}(x)) \\
&= f_4(\Phi^{-1}(x)) \\
&= P_{\mathbf{R}^2_+}(f_3)(\Phi^{-1}(x)) \\
&= f_3(\Phi^{-1}(x))
\end{aligned}$$

$$
\begin{aligned}
&= \tilde{f}_2|\mathbf{R}^2_+ (\Phi^{-1}(x)) \\
&= f_2(\Phi^{-1}(x)) \\
&= f_1(\Phi|Q_+ (\Phi^{-1}(x))) \\
&= f_1(\Phi(\Phi^{-1}(x))) \\
&= f_1(x) \\
&= (\psi\tilde{z})|U_+ (x) \\
&= \psi(x)\tilde{z}(x) \\
&= \psi(x)z(x) \qquad \forall x \in \Omega \cap U
\end{aligned} \tag{3.53}
$$

From (3.51), (3.52) and (3.53) it is now clear that

$$P_m(z)(x) = \psi_m(x)z(x) \qquad \forall x \in \Omega \qquad m = 0, \ldots, M$$

Hence by (3.50) and the property (iv) of the C^∞-partition of unity Ψ we have

$$P_\Omega(z)(x) = \sum_{m=0}^{M} \psi_m(x)z(x) = z(x) \qquad \forall x \in \Omega$$

■

The important aspect to note about this theorem is, that the constant upper bound on the norm of the extension operator P_Ω *only* depends on the upper bound of the Lipschitz constants of the Lipschitz charts and their inverses, and on the C^∞-partition of unity Ψ. This fact will, as we shall see in appendix A, make it possible to define a uniformly bounded collection of extension operators for a whole corresponding collection of interior set approximations of a given Lipschitz domain. The importance of such a collection of extension operators will be illustrated in the next section.

3.8 Admissible Image Segments

The defining properties of the Lipschitz domains introduced in section 3.5 are, as we have seen, appropriate for generating an extension operator. This is of course also the case for other tasks, which can easily be reduced to a local problem, where the Lipschitz chart serves as a convenient tool. Some proofs of regularity for example, fall into this category. However, for the analysis of cost sensitivity with respect to boundary perturbations,

which we soon will be facing, the situation is different. Here the Hilbert space results from section 3.3 will be adequate and lead to simple solutions for a class of domains characterized by certain global properties. Exactly which the global properties of these domains, referred to as *admissible image segments*, should be, are determined by the necessity, that the optimal image cost over such a domain is lower semicontinuous with respect to perturbations of its boundary.

Consider an open subset G of an image domain B of an original image function $\zeta \in L_2(B)$, and let $F \subseteq G$. If F is chosen, so that $\overline{F} \subseteq G$, the distance between F and the boundary ∂G of G is strictly positive, and hence $F \subseteq G'$ for any set G' obtained from G by a sufficiently small perturbation of ∂G. It then follows, that the optimal image cost $c_{G'}(z_{G'})$ over G' is bounded below by the optimal image cost $c_F(z_F)$ over the interior set F. Hence the difference $c_G(z_G) - c_{G'}(z_{G'})$ between the optimal image costs over G and G' is bounded above by $c_G(z_G) - c_F(z_F)$. The desired lower semicontinuity then follows, provided that F can be chosen, so that $c_F(z_F)$ is an arbitrarily good approximation of $c_G(z_G)$.

It turns out that, if the optimal image function z_F over F is extendible by means of an extension operator of the kind, that we discussed in the previous section, and F is a "good" interior set approximation of G, that is the set $G \setminus F$ is of small measure, then $c_F(z_F)$ is indeed a good approximation of $c_G(z_G)$. Moreover the smaller the norm of the extension operator, and the smaller the measure of the difference set $G \setminus F$, the better this approximation will be. We would therefore like an admissible image segment to be defined as follows:

Definition 3.8.1 *Let $B \subseteq \mathbf{R}^2$ be an image domain. We say, that a (bounded) open set $G \subseteq B$ is an* admissible image segment *(of B), if there exists a collection $\mathcal{F}_G$ of open sets with the following properties:*

(i) $\overline{F} \subseteq G \ \forall F \in \mathcal{F}_G$

(ii) $\exists$ *a uniformly bounded collection* $\mathcal{P}_{\mathcal{F}_G} \doteq \{P_F \in \mathcal{L}(\mathcal{H}^1(F), \mathcal{H}^1(\mathbf{R}^2))\}_{F \in \mathcal{F}_G}$ *of extension operators.*

(iii) $\inf_{F \in \mathcal{F}_G} m(G \setminus F) = 0$

An obvious question at this point is, whether such admissible image segments exist, and if so, whether one can find sufficient (local) conditions on a given subset of B,

which ensure, that this subset is an admissible image segment of B. The following theorem, which we prove in appendix A, gives an answer to that question.

Theorem 3.8.2 *All subsets of class $C^{0,1}$ of an image domain B are admissible image segments of B.*

Although this theorem fails to be conclusive for domains with cusps, occurring for example in images of cylinders with circular cross sections, such domains or a large subclass thereof, we believe, are still likely to be admissible image segments. Besides that, it seems for most purposes reasonable to assume, that domains with cusps can be sufficiently well approximated by domains of class $C^{0,1}$. Thus the class of admissible image segments seems to be large enough, to be useful for modeling of the "true" segments in a real image.

We now continue with the cost sensitivity problem outlined above.

Lemma 3.8.3 (Key Approximation) *Let $\zeta \in L_2(B)$ be an original image function, and let the image cost function $c_\Omega : \mathcal{H}^1(\Omega) \to \overline{\mathbf{R}_+}$ be defined for open sets $\Omega \subseteq B$ as in (3.19). Furthermore let G be an admissible image segment of B, and let $\mathcal{F}_G$ be a collection of open sets with the properties (i) – (iii) in definition 3.8.1. Then*

$$\sup_{F \in \mathcal{F}_G} c_F(z_F) \geq c_G(z_G)$$

Proof: Let $F \in \mathcal{F}_G$, and let $E \doteq G \setminus F$. Then by corollary 3.3.13

$$\begin{aligned} c_G(z_G) - c_F(z_F) &= \\ &= \|\zeta|E\|^2_{L_2(E)} - \langle \zeta|G, z_G\rangle_G + \langle \zeta|F, z_F\rangle_F \\ &= \langle \zeta|E, (\zeta - z_G)|E\rangle_E - \langle \zeta|F, z_G|F - z_F\rangle_F \end{aligned} \tag{3.54}$$

In order to bound this expression above, we let $w \doteq P_F(z_F)|G$, where $P_F \in \mathcal{P}_{\mathcal{F}_G}$ in definition 3.8.1 (ii). By proposition 3.6.4 $w \in \mathcal{H}^1(G)$ and

$$\|w\|_{\mathcal{H}^1(G)} \leq \|P_F\|_{\mathcal{L}(\mathcal{H}^1(F), \mathcal{H}^1(\mathbf{R}^2))} \|z_F\|_{\mathcal{H}^1(F)}$$

Thus by theorem 3.3.12 and the uniform boundedness of $\mathcal{P}_{\mathcal{F}_G}$ we have that

$$\|w\|_{\mathcal{H}^1(G)} \leq M \doteq \sup_{F \in \mathcal{F}_G} \|P_F\|_{\mathcal{L}(\mathcal{H}^1(F), \mathcal{H}^1(\mathbf{R}^2))} \cdot \frac{\|\zeta\|_{L_2(B)}}{1 \wedge \mu} < \infty \tag{3.55}$$

where M, so defined, is independent of F (and hence of w). To simplify some later expressions, we also note, that the constant M bounds a couple of other quantities as well.

Indeed from proposition 3.6.4 it is clear, that the norm of any extension operator ($\in \mathcal{P}_{\mathcal{F}_G}$) is bounded below by unity. Hence

$$\|\zeta|G\|_{L_2(G)} \leq M \tag{3.56}$$

and by theorem 3.3.12

$$\|z_G\|_{\mathcal{H}^1(G)} \leq M \tag{3.57}$$

Now let the bilinear form $a_\Omega : \mathcal{H}^1(\Omega)^2 \to \mathbf{R}$ be defined for open sets $\Omega \subseteq B$ as in (3.20). Since $F \subseteq G$ is open, and $z_G \in \mathcal{H}^1(G)$, proposition 3.6.4 shows, that $z_G|F \in \mathcal{H}^1(F)$. Hence by theorem 3.3.12

$$a_F(w|F, z_G|F) = a_F(z_F, z_G|F) = \langle \zeta|F, z_G|F\rangle_F \tag{3.58}$$

Likewise as $w \in \mathcal{H}^1(G)$ we have

$$a_G(w, z_G) = \langle \zeta|G, w\rangle_G = \langle \zeta|F, z_F\rangle_F + \langle \zeta|E, w|E\rangle_E \tag{3.59}$$

Next from proposition 3.6.3 we note that

$$D_k(f|F) = (D_k f)|F \qquad k = 1,2 \qquad \forall f \in \mathcal{H}^1(G)$$

Thus

$$a_G(w, z_G) - a_F(w|F, z_G|F) = \int_E \left(w z_G + \mu \sum_{k=1}^{2} D_k w D_k z_G \right) dx \tag{3.60}$$

From (3.58), (3.59) and (3.60) we now obtain

$$\langle \zeta|F, z_G|F - z_F\rangle_F = \langle \zeta|E, w|E\rangle_E - \int_E \left(w z_G + \mu \sum_{k=1}^{2} D_k w D_k z_G \right) dx \tag{3.61}$$

Substituting (3.61) in (3.54), the Cauchy-Schwarz inequalities and the bounds (3.55) – (3.57) then yield

$$\begin{aligned} & c_G(z_G) - c_F(z_F) = \\ & \quad = \langle \zeta|E, (\zeta - z_G - w)|E\rangle_E + \int_E \left(w z_G + \mu \sum_{k=1}^{2} D_k w D_k z_G \right) dx \\ & \quad \leq \|\zeta|E\|_{L_2(E)} \|(\zeta - z_G - w)|E\|_{L_2(E)} \\ & \qquad + (1 \vee \mu) \sum_{\substack{\alpha \in \mathbf{N}_0^2 \\ |\alpha| \leq 1}} \|(D^\alpha w)|E\|_{L_2(E)} \|(D^\alpha z_G)|E\|_{L_2(E)} \end{aligned}$$

$$\leq \|\zeta|E\|_{L_2(E)}\|\zeta|G - z_G - w\|_{L_2(G)} + (1 \vee \mu)\|w\|_{\mathcal{H}^1(G)} \sqrt{\sum_{\substack{\alpha \in \mathbb{N}_0^2 \\ |\alpha| \leq 1}} \|(D^\alpha z_G)|E\|^2_{L_2(E)}}$$

$$\leq 3M\|\zeta|E\|_{L_2(E)} + (1 \vee \mu)M \sqrt{\sum_{\substack{\alpha \in \mathbb{N}_0^2 \\ |\alpha| \leq 1}} \|(D^\alpha z_G)|E\|^2_{L_2(E)}}$$

Since $\zeta \in L_2(B)$, $z_G \in \mathcal{H}^1(G)$ and $\inf_{F \in \mathcal{F}_G} m(G \setminus F) = 0$, it then follows by the Dominated Convergence Theorem, that

$$\inf_{F \in \mathcal{F}_G} [c_G(z_G) - c_F(z_F)] \leq 0$$

■

The primary value of the lemma above is, that it captures the dependence of the image cost functional on the boundary of its domain, without explicit reference to any of the properties of this boundary or the optimal image function, which minimizes the image cost. A more naïve approach would have been, to use various regularity properties of the optimal image function to translate small perturbations in the boundary of the domain into small perturbations of the boundary conditions of the Euler equation associated with the image cost functional over some interior set of the original domain. From this it might have been possible to show, that the resulting perturbation of the solution of this Euler equation, that is the optimal image function over the interior set, is small in some sense, and that this in turns implies a small perturbation of the optimal cost over the whole domain. This approach would unavoidably lead to a number of technicalities and possibly extra assumptions, which the proof of the lemma above in a remarkable way avoids, by finding a more direct estimate of the perturbation of the optimal image cost in terms only of the measure of the set, in which the boundary perturbation is taking place. Most of the work in the earlier sections of this chapter has been aimed at deriving results used in the proof of the lemma above, or at finding conditions, under which this lemma can be applied. In addition the work in the sections to come largely depends on this same lemma. It is therefore fair to say, that lemma 3.8.3 is the heart of the entire proof of existence of optimal edges in the sense of the total cost functions considered in this chapter.

3.9 Admissible Image Segmentations

Up till this point our analysis has yet not produced any results specifically about the optimal image cost over the entire image domain, or to be more precise, over the continuity set. As a consequence we have not yet been able to make any claims about the existence of an optimal continuity set or equivalently a collection of optimal edge segments specifying such a set. For this development we need to return to our earlier idea of parametrizing the space of continuity sets by image segmentations as in (3.1) – (3.3).

It was indicated already in section 3.3, that the existence of a unique optimal image function $z_\Omega \in \mathcal{H}^1(\Omega)$ for each open subset Ω of the image domain allows us to consider the optimal image cost over the whole continuity set as a function of the edge segments forming the corresponding discontinuity set. In this section we make this claim more precise by defining the *optimal N-segment image cost*, $N \in \mathbb{N}_0$, as a function of the image segmentation. We then show, that this function is lower semicontinuous on a certain class of image segmentations, which we will refer to as *admissible image segmentations.*

Following the discussion in section 3.1 we let Σ denote the compact interval $[0,1]$. By an *N-segment image segmentation* we then mean an $\mathbf{R}^{2N}$-valued function

$$\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C(\Sigma)^{2N} \qquad N \in \mathbb{N}_0$$

where each one of the functions $\gamma_n \in C(\Sigma)^2$, $n = 1, \ldots, N$ parametrizes a curve $\gamma_n(\Sigma)$, representing and *edge(-segment).* The degenerate case, when $N = 0$, that is when no edges are present, has been included only for later consistency in our notation. In this case it is of course immaterial, what the function γ is. If we define $\mathbf{R}^0$ to be the trivial real vector space $\{0\}$, it follows, that $C(\Sigma)^0$ consists of the single constant function $\gamma^0 : \Sigma \to \mathbf{R}^0 : \sigma \mapsto 0$, referred to as the *trivial image segmentation.* The *discontinuity set* D_γ of a given N-segment image segmentation $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T$ is now defined to be the union of the N edge segments associated with γ, that is

$$D_\gamma \doteq \bigcup_{n=1}^{N} \gamma_n(\Sigma) \tag{3.62}$$

(interpreted as usual as $\varnothing$ if $N = 0$). For a given image domain B we also define the corresponding *continuity set* C_γ of γ to be the "edge-free" portion of B, that is

$$C_\gamma \doteq B \setminus D_\gamma \tag{3.63}$$

If $N = 0$, obviously $C_\gamma = B$.

Fact 3.9.1 *The continuity set C_γ is open in $\mathbf{R}^2$.*

Proof: If $N = 0$, then $C_\gamma = B$ is the entire image domain, and hence open by definition. If $N \in \mathbf{N}$, then $\gamma_1, \ldots, \gamma_N$ are continuous functions on the compact domain Σ. Thus D_γ is compact in $\mathbf{R}^2$, and since B is open, this implies, that $C_\gamma = B \setminus D_\gamma$ is open. ■

This trivial fact is important, because it allows us to consider image functions $z \in \mathcal{H}^1(C_\gamma)$ without further restrictions on the image segmentation γ, a condition, that was used implicitly in the discussion about the domain of the total cost function in section 3.1. Accordingly we define the *optimal N-segment image cost function*

$$\check{c}_N : C(\Sigma)^{2N} \to \overline{\mathbf{R}_+} : \gamma \mapsto c_{C_\gamma}(z_{C_\gamma}) \qquad N \in \mathbf{N}_0 \tag{3.64}$$

where the continuity set C_γ is defined as in (3.62) and (3.63) above, and the image cost c_{C_γ} is defined as in (3.19).

In spite of this desirable property of the continuity set, the results in the previous section about admissible image segments suggest, that we restrict attention to a corresponding class of admissible image segmentations.

Definition 3.9.2 *An* image segmentation $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C(\Sigma)^{2N}$, $N \in \mathbf{N}_0$, *of an image domain B is said to be* admissible, *if each of the components of the continuity set $C_\gamma \doteq B \setminus \bigcup_{n=1}^N \gamma_n(\Sigma)$ of γ is an admissible image segment of B.*

Restricting the image segmentations to belong to this admissible class obviously rules out edges with free endpoints (in the image domain), that is edges of the kind shown in figure 2.9. This restriction is probably the most serious limitation of the existence proof presented in this chapter. We argue however, that edges with free endpoints are rare, and that their value for image segmentation purposes, one of the prime motivations for edge detection, is limited, as they do not give rise to any new components of the continuity set.

Our next goal is now to show, that the optimal N-segment image cost $\check{c}_N$ is lower semicontinuous on the set of admissible image segmentations. For the proof of that we need the following fact.

Proposition 3.9.3 *Let $\zeta \in L_2(B)$ be an original image function, and let the image cost function c_Ω be defined for open sets $\Omega \subseteq B$ as in (3.19). Let $C \subseteq B$ be an open set, and let $z \in \mathcal{H}^1(C)$. Then*

$$c_C(z) \geq c_F(z|F) \qquad \forall \text{ open sets } F \subseteq C$$

and

$$c_C(z) = \sum_{G \in \mathcal{G}} c_G(z|G)$$

where $\mathcal{G}$ is any disjoint open covering of C consisting of subsets of C.

Proof: Let F be an open subset of C. Then by proposition 3.6.3

$$D_k(z|F) = (D_k z)|F \qquad k = 1, 2 \tag{3.65}$$

and by proposition 3.6.4 $z|F \in \mathcal{H}^1(F)$. Hence $c_F(z|F)$ is well-defined, and

$$\begin{aligned} c_C(z) &= \\ &= \int_C \left[(z-\zeta)^2 + \mu \sum_{k=1}^{2} (D_k z)^2 \right] dx \\ &= c_F(z|F) + \int_{C \setminus F} \left[(z-\zeta)^2 + \mu \sum_{k=1}^{2} (D_k z)^2 \right] dx \\ &\geq c_F(z|F) \end{aligned}$$

Since the members of $\mathcal{G}$ are pairwise disjoint and open in $\mathbf{R}^2$, and hence at most countably many, and since the integral kernel $(z-\zeta)^2 + \mu \sum_{k=1}^{2} (D_k z)^2$ is positive, by the Monotone Convergence Theorem and (3.65) we have that

$$c_C(z) = \sum_{G \in \mathcal{G}} \int_G \left[(z-\zeta)^2 + \mu \sum_{k=1}^{2} (D_k z)^2 \right] dx = \sum_{G \in \mathcal{G}} c_G(z|G)$$

■

Theorem 3.9.4 *Let $\zeta \in L_2(B)$ be an original image function. Then the optimal N-segment image cost $\check{c}_N : C(\Sigma)^{2N} \to \overline{\mathbf{R}_+}$, $N \in \mathbf{N}_0$, defined as in (3.64), is lower semicontinuous on the set of admissible image segmentations of B.*

Proof: Since $C(\Sigma)^0$ only contains the single trivial (admissible) image segmentation $\Sigma \to \mathbf{R}^0$, corresponding to the empty discontinuity set, $\check{c}_0$ is trivially lower semicontinuous. Next let γ be an admissible N-segment image segmentation of the image domain B, $N \in \mathbf{N}$, and let $\mathcal{G}$ be the collection of all the components of the continuity set C_γ of γ. Clearly $\mathcal{G}$ is countable, and

$$\sum_{G \in \mathcal{G}} \|\zeta|G\|^2_{L_2(G)} = \|\zeta|C_\gamma\|^2_{L_2(C_\gamma)} = \|\zeta\|^2_{L_2(B)} < \infty$$

Thus given $\epsilon > 0$, $\exists$ a finite collection $\{G_j\}_{j=1}^J \subseteq \mathcal{G}$, such that

$$\sum_{G \in \mathcal{G}_0} c_G(0) = \sum_{G \in \mathcal{G}_0} \|\zeta|G\|^2_{L_2(G)} < \frac{\epsilon}{2} \tag{3.66}$$

where $\mathcal{G}_0 \doteq \mathcal{G} \setminus \{G_j\}_{j=1}^J$. Since $G_1, \ldots, G_J$ are admissible image segments of B, $\exists$ corresponding collections $\mathcal{F}_1, \ldots, \mathcal{F}_J$ of open subsets of $G_1, \ldots, G_J$ respectively with the properties (i) – (iii) listed in definition 3.8.1. Furthermore by lemma 3.8.3 we have that

$$\sup_{F \in \mathcal{F}_j} c_F(z_F) \geq c_{G_j}(z_{G_j}) \qquad j = 1, \ldots, J$$

Thus we can pick $F_j \in \mathcal{F}_j$, such that

$$c_{F_j}(z_{F_j}) > c_{G_j}(z_{G_j}) - \frac{\epsilon}{2J} \qquad j = 1, \ldots, J \tag{3.67}$$

Now let

$$F \doteq \bigcup_{j=1}^J F_j$$

Then by property (i) of the collections $\mathcal{F}_1, \ldots, \mathcal{F}_J$ and (3.63) above

$$\overline{F} = \bigcup_{j=1}^J \overline{F_j} \subseteq \bigcup_{j=1}^J G_j \subseteq C_\gamma = B \setminus D_\gamma \subseteq \complement D_\gamma$$

Since the discontinuity set D_γ of γ is compact, and $\overline{F}$ is closed, this implies, that D_γ and F are separated by a Euclidean distance $\delta > 0$. Thus for any image segmentation $\gamma' \in C(\Sigma)^{2N}$, such that $\|\gamma' - \gamma\|_{C(\Sigma)^{2N}} < \delta/\sqrt{2}$, we have that, $D_{\gamma'} \subseteq \complement \overline{F}$ and hence, as $F \subseteq B$, we see that, $F \subseteq B \cap \complement D_{\gamma'} = C_{\gamma'}$. By proposition 3.6.4 it then follows, that $z_{C_{\gamma'}}|F \in \mathcal{H}^1(F)$ and $z_{C_{\gamma'}}|F_j \in \mathcal{H}^1(F_j)$, $j = 1, \ldots, J$. Since $G_1, \ldots, G_J$ and hence $F_1, \ldots, F_J$ are pairwise disjoint, by (3.64), proposition 3.9.3, theorem 3.3.12 and (3.67) we then obtain

$$\begin{aligned}
\check{c}_N(\gamma') = & \\
&= c_{C_{\gamma'}}(z_{C_{\gamma'}}) \\
&\geq c_F(z_{C_{\gamma'}}|F) \\
&= \sum_{j=1}^J c_{F_j}(z_{C_{\gamma'}}|F_j) \\
&\geq \sum_{j=1}^J c_{F_j}(z_{F_j}) \\
&> \sum_{j=1}^J c_{G_j}(z_{G_j}) - \frac{\epsilon}{2}
\end{aligned} \tag{3.68}$$

Let

$$z \doteq \sum_{j=1}^{J} \widetilde{z_{G_j}} | C_\gamma$$

Then from proposition 3.6.8 we see, that $z \in \mathcal{H}^1(C_\gamma)$, and thus by the same token

$$\check{c}_N(\gamma) = c_{C_\gamma}(z_{C_\gamma}) \leq c_{C_\gamma}(z) = \sum_{G \in \mathcal{G}} c_G(z|G) = \sum_{j=1}^{J} c_{G_j}(z_{G_j}) + \sum_{G \in \mathcal{G}_0} c_G(0) \tag{3.69}$$

It now follows from (3.66), (3.68) and (3.69), that $\check{c}_N(\gamma') > \check{c}_N(\gamma) - \epsilon$, whenever $\|\gamma' - \gamma\|_{C(\Sigma)^{2N}} < \delta$. ■

The theorem above summarizes, what we need to know about the image cost function. We are now ready to prove the existence of an optimal image segmentation, which minimizes the total cost function.

3.10 Existence of Optimality

In this section we show, how the just demonstrated semicontinuity of the optimal N-segment image cost $\check{c}_N$ together with the edge cost assumptions stated in the introduction lead to the conclusion, that the total cost $c(N, \gamma, z)$ attains its minimum on a certain class of domains. In particular this implies, that there exists some optimal image segmentation $\gamma \in C(\Sigma)^{2N}$ for some optimal $N \in \mathbb{N}_0$, and thus optimal edges represented by the corresponding discontinuity set D_γ.

Typically a desirable edge cost function, such as those given in the examples of the previous chapter, is not well-defined on the entire space $C(\Sigma)^{2N}$. As mentioned in the introduction of this chapter we therefore consider a collection of edge cost functions

$$c_N : S_N \subseteq C(\Sigma)^{2N} \to \overline{\mathbb{R}_+} \qquad N \in \mathbb{N}_0 \tag{3.70}$$

defined on some appropriate domains S_N, $N \in \mathbb{N}_0$. In order to obtain a tractable description of the discontinuity set, it is essential, that the number of edge segments is not allowed to grow beyond all bounds. We therefore assume that

$$\inf_{\gamma \in S_N} c_N(\gamma) \longrightarrow \infty \qquad \text{as } N \longrightarrow \infty \tag{3.71}$$

As we noted earlier, the edge cost examples in the previous chapter all satisfy this condition.

Naturally the domain of the total cost function is forced to depend on the domains S_N, $N \in \mathbb{N}_0$, of the edge cost functions, and is therefore not just a simple Cartesian product space. Given an original image function $\zeta \in L_2(B)$, for any collection $\mathcal{K} \doteq \{K_N \subseteq C(\Sigma)^{2N}\}_{N \in \mathbb{N}_0}$, we therefore define the *total cost function domain (subordinate to $\mathcal{K}$)* by

$$\mathcal{D}_{\mathcal{K}} \doteq \{(N, \gamma, z) : z \in \mathcal{H}^1(C_\gamma), \gamma \in K_N, N \in \mathbb{N}_0\} \tag{3.72}$$

where the continuity set C_γ of γ is defined (in terms of the corresponding discontinuity set D_γ) as in (3.62) and (3.63), and the image cost function c_{C_γ} is defined as in (3.19). For the same original image function ζ the *total cost function* is then defined by

$$c : \mathcal{D}_{\mathcal{S}} \to \overline{\mathbb{R}_+} : (N, \gamma, z) \mapsto c_N(\gamma) + c_{C_\gamma}(z) \tag{3.73}$$

where

$$\mathcal{S} \doteq \{S_N\}_{N \in \mathbb{N}_0} \tag{3.74}$$

To be able to apply the semicontinuity result from the previous section, we need to restrict the image segmentations to be admissible. Furthermore for each $N \in \mathbb{N}_0$ we need to equip S_N with a topology no weaker than the inherited $C(\Sigma)^{2N}$-topology, and sufficiently strong, that the edge cost function c_N is lower semicontinuous. For this reason for $N \in \mathbb{N}_0$ we define $A_N \subseteq C(\Sigma)^{2N}$ to be the *set of admissible N-segment image segmentations*, and $\mathcal{T}_N$ to be the *topology on S_N generated by the function c_N and the $C(\Sigma)^{2N}$-topology.* With the optimal N-segment image cost $\check{c}_N : C(\Sigma)^{2N}$, $N \in \mathbb{N}_0$ defined as in (3.64), from (3.73) above we note that

$$c(N, \gamma, z_{C_\gamma}) = c_N(\gamma) + \check{c}_N(\gamma) \qquad \gamma \in S_N \tag{3.75}$$

We then have the following result asserting the existence of an optimal N-segment image segmentation for each $N \in \mathbb{N}_0$.

Fact 3.10.1 *For each $N \in \mathbb{N}_0$ the function $c_N + \check{c}_N | S_N$ attains its minimum on every nonempty $\mathcal{T}_N$-compact set $K \subseteq S_N \cap A_N$ of image segmentations.*

Proof: To prove that $c_N + \check{c}_N | S_N$ attains its minimum on a nonempty $\mathcal{T}_N$-compact set $K \subseteq S_N \cap A_N$, it is sufficient to show, that the two functions c_N and $\check{c}_N | S_N$ are lower semicontinuous on $S_N \cap A_N$ with respect to $\mathcal{T}_N$. By definition of $\mathcal{T}_N$, $c_N : (S_N, \mathcal{T}_N) \to \overline{\mathbb{R}_+}$ is continuous, and hence lower semicontinuous on $S_N \cap A_N$. From theorem 3.9.4 we know, that $\check{c}_N$ is lower semicontinuous on A_N with respect to the $C(\Sigma)^{2N}$-topology. Since $\mathcal{T}_N$ by

definition is stronger than this topology, it follows, that $\check{c}_N|S_N : (S_N, \mathcal{T}_N) \to \overline{\mathbb{R}_+}$ is lower semicontinuous on $S_N \cap A_N$ as well. ■

The fact above suggests, that we consider a restriction of the total cost function c to a subset $\mathcal{D}_\mathcal{K}$ of $\mathcal{D}_\mathcal{S}$. Thus we let

$$\mathcal{K} \doteq \{K_N\}_{N \in \mathbb{N}_0} \tag{3.76}$$

where for each $N \in \mathbb{N}_0$, $K_N \subseteq S_N \cap A_N$ is a $\mathcal{T}_N$-compact set of N-segment image segmentations. From fact 3.10.1 we then see, that for each $N \in \mathbb{N}_0$, such that $K_N \neq \emptyset$, there exists an optimal N-segment image segmentation $\gamma^N \in K_N$, which minimizes the function $c_N|K_N + \check{c}_N|K_N$. We can therefore define the *optimal total cost function* $\check{c} : \mathbb{N}_0 \to [0, \infty]$ by

$$\check{c}(N) \doteq \begin{cases} c_N(\gamma^N) + \check{c}_N(\gamma^N) & \text{if } K_N \neq \emptyset \\ \infty & \text{otherwise} \end{cases} \tag{3.77}$$

For this function we have the following result, asserting the existence of an optimal number of edge segments.

Fact 3.10.2 *The function $\check{c}$ attains its minimum. If $K_N \neq \emptyset$ for some $N \in \mathbb{N}_0$, this minimum is finite.*

Proof: If $K_N = \emptyset \ \forall N \in \mathbb{N}_0$, then $\check{c}(N) = \infty \ \forall N \in \mathbb{N}_0$, in which case $\check{c}$ trivially attains its minimum. Suppose instead, that $\exists \ \underline{N} \in \mathbb{N}_0$, such that $K_{\underline{N}} \neq \emptyset$. Then $\check{c}(\underline{N}) < \infty$. From the nonnegativity of the optimal N-segment image cost $\check{c}_N$ and the edge cost assumption (3.71) we see that $\check{c}(N) \geq c_N(\gamma^N) \longrightarrow \infty$ as $N \longrightarrow \infty$. Hence $\exists \ \overline{N} \in \mathbb{N}_0$, such that $\check{c}(N) > \check{c}(\underline{N}) \ \forall N > \overline{N}$. This implies that

$$\inf_{N \in \mathbb{N}_0} \check{c}(N) = \bigwedge_{N=0}^{\overline{N}} \check{c}(N) \leq \check{c}(\underline{N}) < \infty$$

Thus $\check{c}$ attains its minimum for some $\check{N} \in \{0, \ldots, \overline{N}\}$, and this minimum is finite. ■

For obvious reasons we refer to the number $\check{N}$, which minimizes the optimal total cost $\check{c}$, as the *optimal number of edge segments.*

We are now finally in the position of reaching the main goal of this chapter, that is we are ready to show, that the total cost function attains its minimum on every nonempty domain $\mathcal{D}_\mathcal{K}$ of the kind introduced above.

Theorem 3.10.3 *Let $\zeta \in L_2(B)$ be an original image function. Assume that the edge cost functions $c_N : S_N \subseteq C(\Sigma)^{2N} \to \overline{\mathbb{R}_+}$, $N \in \mathbb{N}_0$, where $\Sigma \doteq [0,1]$, satisfy the condition*

$$\inf_{\gamma \in S_N} c_N(\gamma) \longrightarrow \infty \qquad \text{as } N \longrightarrow \infty$$

Define the continuity sets

$$C_\gamma \doteq B \setminus \bigcup_{n=1}^{N} \gamma_n(\Sigma) \qquad \gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in \prod_{n=1}^{N} C(\Sigma)^2 \qquad N \in \mathbb{N}_0$$

and consider the total cost function

$$c : \mathcal{D}_S \to \overline{\mathbb{R}_+} : (N,\gamma,z) \mapsto c_N(\gamma) + \int_{C_\gamma} \left[(z-\zeta)^2 + \mu \sum_{k=1}^{2} (D_k z)^2 \right] dx$$
$$\mathcal{D}_S \doteq \{(N,\gamma,z) : z \in \mathcal{H}^1(C_\gamma), \gamma \in S_N, N \in \mathbb{N}_0\} \qquad \mu > 0 \tag{3.78}$$

Let $\mathcal{D}_K \doteq \{(N,\gamma,z) : z \in \mathcal{H}^1(C_\gamma), \gamma \in K_N, N \in \mathbb{N}_0\}$, where for each $N \in \mathbb{N}_0$, $K_N \subseteq S_N$ is a compact set of admissible image segmentations with respect to the topology generated by c_N and the $C(\Sigma)^{2N}$-topology. If $\mathcal{D}_K \neq \emptyset$, then $c|\mathcal{D}_K$ attains its minimum. In other words, there exists $(\check{N},\check{\gamma},\check{z})$, such that

$$c(N,\gamma,z) \geq c(\check{N},\check{\gamma},\check{z}) \qquad \forall (N,\gamma,z) \in \mathcal{D}_K$$

Furthermore, given $(\check{N},\check{\gamma})$, then $\check{z}$ is unique.

Proof: Assume that $\mathcal{D}_K \neq \emptyset$, and let $(N,\gamma,z) \in \mathcal{D}_K$. Then $K_N \neq \emptyset$. Thus by fact 3.10.2 $\exists\ \check{N} \in \mathbb{N}_0$, which minimizes the optimal total cost function $\check{c}$ defined in (3.77), and the minimum is finite. Hence $K_{\check{N}} \neq \emptyset$ and

$$\infty > c_N(\gamma^N) + \check{c}_N(\gamma^N) = \check{c}(N) \geq \check{c}(\check{N}) = c_{\check{N}}(\check{\gamma}) + \check{c}_{\check{N}}(\check{\gamma}) \tag{3.79}$$

where $\check{c}_N$ and $\check{c}_{\check{N}}$ are the optimal N- and $\check{N}$-segment image cost functions respectively defined in (3.64), and $\gamma^N \in K_N$ and $\check{\gamma} \doteq \gamma^{\check{N}} \in K_{\check{N}}$ are the image segmentations, which minimize the functions $c_N|K_N + \check{c}_N|K_N$ and $c_{\check{N}}|K_{\check{N}} + \check{c}_{\check{N}}|K_{\check{N}}$ respectively according to fact 3.10.1. Let $z_{C_\gamma} \in \mathcal{H}^1(C_\gamma)$ and $\check{z} \doteq z_{C_{\check{\gamma}}} \in \mathcal{H}^1(C_{\check{\gamma}})$ be the unique optimal image functions minimizing the image cost functions c_{C_γ} and $c_{C_{\check{\gamma}}}$ respectively defined in (3.19) according to theorem 3.3.12. Then by (3.78)

$$c(N,\gamma,z) = c_N(\gamma) + c_{C_\gamma}(z) \geq c_N(\gamma) + c_{C_\gamma}(z_{C_\gamma}) = c_N(\gamma) + \check{c}_N(\gamma) \tag{3.80}$$

and by (3.75)

$$c_{\check{N}}(\check{\gamma}) + \check{c}_{\check{N}}(\check{\gamma}) = c(\check{N}, \check{\gamma}, \check{z}) \tag{3.81}$$

The theorem now follows from the equations (3.79) – (3.81). ■

The theorem above shows, that for any total cost function of the form considered in this chapter, there exists an optimal image segmentation $\check{\gamma}$ consisting of an optimal number $\check{N}$ of edge segments and an optimal estimated image function $\check{z}$. Without the existence of this optimal triplet $(\check{N}, \check{\gamma}, \check{z})$ the discontinuity set D_γ, that is the edges themselves, could conceivably get wilder and wilder as the total cost $c(N, \gamma, z)$ gets arbitrarily close to its greatest lower bound. Theorem 3.10.3 is therefore most essential for the justification of the global cost function minimization approach to edge detection and localization, that we proposed in the previous chapter.

It should be pointed out, that we have not made any claims about the uniqueness of the optimal edges. For some of the total cost functions considered in theorem 3.10.3, it is in fact not too hard, although somewhat tedious, to find original image functions, for which the optimal discontinuity set is *not* unique. Thus general claims about uniqueness of the optimal edges cannot be made. Moreover, even in the event of a unique optimal discontinuity set, $(\check{N}, \check{\gamma})$ is in general not unique, due to the general noninjectivity of the map

$$\bigcup_{N \in \mathbf{N}_0} K_N \to \mathbf{2}^{\mathbf{R}^2} : \gamma \mapsto D_\gamma$$

where $\mathbf{2}^{\mathbf{R}^2}$ denotes the power set of $\mathbf{R}^2$. For example, reversing the parametrization of one of the edge segments $\gamma_1(\Sigma), \ldots, \gamma_N(\Sigma)$, or permuting any two of them, does not alter the resulting discontinuity set.

For edge detection purposes we are primarily interested in the optimal discontinuity set itself, so any nonuniqueness of $(\check{N}, \check{\gamma})$ due to multiple parametrizations of this set is an artifact, which can be neglected. The possibility of multiple optimal discontinuity sets on the other hand is not an artifact, but means, that the complete solution of our edge detection problem is indeed represented by the whole equivalence class of all the optimal discontinuity sets. One could for example aim at finding the whole equivalence class, and use another mechanism, such as higher level knowledge, to select the member of this class, which is best in some sense. Alternatively one could adopt the point of view, that the equivalence class of optimal discontinuity sets is sufficiently well represented by any of its

members, and accept any one of these as the solution to the edge detection problem. In practice the occurrence of multiple optimal discontinuity sets might be rare. On the other hand the discontinuity sets corresponding to the different local minima of the total cost might all be of value for later higher level processing steps, even if the optimal discontinuity set *is* unique.

For image recovery purposes we are not only interested in the equivalence class of optimal discontinuity sets, but also, and primarily, in the optimal estimated image function z_{C_γ} for each of the optimal continuity sets $C_\gamma = B \setminus D_\gamma$. Since by theorem 3.3.12 z_{C_γ} is unique for a given C_γ, the class of optimal estimated image functions is exactly as large as the equivalence class of optimal discontinuity sets. Again in practice it might be worth while considering the optimal estimated image functions corresponding to different local minima of the total cost.

The edge cost assumption (3.71) was used in the proof of theorem 3.10.3, to obtain an upper bound $\overline{N} \in \mathsf{N}_0$ on the number of edge segments an optimal image segmentation could have. This led us to the existence of an optimal number of edge segments $\check{N}$. If $\overline{N}$ is large, this kind of reasoning is of little practical value. What really matters for practical purposes however, is the existence of an optimal N-segment image segmentation γ^N and a corresponding optimal estimated image function $z_{C_{\gamma^N}}$ for each $N \in \mathsf{N}_0$, and this follows from our proof regardless of the existence of $\check{N}$.

The critical reader has of course observed, that the collection $\mathcal{K}$ of compact subsets of admissible image segmentations in S_N, $N \in \mathsf{N}_0$ yet is to be specified, and that the value of theorem 3.10.3 strongly depends on how rich such a collection can be found. Since $\mathcal{K}$ necessarily depends on $\mathcal{S}$, which in turn depends on the expressions defining the edge cost, we have in theorem 3.10.3 intentionally avoided being more specific about the selection of $\mathcal{K}$. This last step is the topic of the next section.

3.11 Image Segmentation Domains

In this final section we give an example of, how the collection $\mathcal{K}$ of compact domains of admissible image segmentations in theorem 3.10.3 can be chosen. By virtue of the wide range of edge cost functions allowed by the hypotheses of that theorem, it is hard, if not impossible, to specify meaningful such collections for all possible edge costs. Yet the example, that we present in this section, serves at least a couple of different purposes. First

of all we want to demonstrate the existence of at least one nontrivial collection $\mathcal{K}$, to which theorem 3.10.3 can be applied. Secondly we want to show, that the collection $\mathcal{K}$ can be specified by local conditions on the edge segments, suitable for computational verification.

3.11.1 Edge Cost Continuity

To be able to exhibit an example, which meets these goals, we will make the additional assumption, that the edge cost is, what we shall call, *C^l-continuous.*

Definition 3.11.1 *Let $l \in \mathbf{N}$. An* edge cost function $c_N : S_N \subseteq C(\Sigma)^{2N} \to \overline{\mathbf{R}_+}$ *is said to be* C^l-continuous, *if the following two conditions are satisfied:*

(i) $S_N \subseteq C^l(\Sigma)^{2N}$

(ii) c_N *is continuous with respect to the* $C^l(\Sigma)^{2N}$*-topology.*

If the edge cost is given by a collection $\{c_N\}_{N \in \mathbf{N}_0}$ *of* C^l*-continuous edge cost functions, we say, that the* edge cost *is* C^l*-continuous.*

As before we let $\Sigma \doteq [0,1]$. For a set $S \subseteq C^l(\Sigma)^{2N}$ of N-segment image segmentations, $l, N \in \mathbf{N}_0$, we define the *l-range of S* to be the set

$$R_{l,N}(S) \doteq \bigcup_{\gamma \in S} \prod_{i=0}^{l} \gamma^{(i)}(\Sigma) \subseteq \mathbf{R}^{2N(l+1)} \tag{3.82}$$

Fact 3.11.2 *Let $N \in \mathbf{N}$. If $S_N \subseteq C^l(\Sigma)^{2N}$, $R_{l,N}(S_N)$ is open (in $\mathbf{R}^{2N(l+1)}$), and $f : R_{l,N}(S_N) \to \overline{\mathbf{R}_+}$ is a continuous function, then the edge cost function*

$$c_N : S_N \to \overline{\mathbf{R}_+} : \gamma \mapsto \int_\Sigma f(\gamma^{(0)}(\sigma), \ldots, \gamma^{(l)}(\sigma))\, d\sigma$$

is C^l-continuous.

Proof: Let $\gamma \in S_N$. Then $\gamma^{(i)} \in C(\Sigma)^{2N}$, $i = 0, \ldots, l$. Thus the compactness of Σ implies, that $R_{l,N}(\{\gamma\})$ is compact. Since $R_{l,N}(S_N)$ is open in the locally compact space $\mathbf{R}^{2N(l+1)}$, $R_{l,N}(S_N)$ is locally compact as well, [60, p186]. Hence $\exists$ a compact neighborhood K of $R_{l,N}(\{\gamma\})$ in $R_{l,N}(S_N)$, [61, p168]. Thus f is uniformly continuous on K, from which it follows, that c_N is continuous at γ. ■

For an N-segment edge cost function defined as the integral over Σ of an algebraic expression of the image segmentation and its l first derivatives, as those in the examples of chapter 2,

it is natural to consider a domain $S_N \subseteq C^l(\Sigma)^{2N}$, whose l-range $R_{l,N}(S_N)$ is open. The fact above therefore shows, that the class of C^l-continuous edge costs is quite large and a very natural one to consider.

3.11.2 Compact Spaces of Admissible Image Segmentations

Given the C^l-continuity assumption above our main objective is to construct a collection $\mathcal{K} \doteq \{K_N\}_{N \in \mathbb{N}_0}$ (depending on l), which satisfies the conditions in theorem 3.10.3, and which is sufficiently rich, that the thereby induced collection of discontinuity sets, that is $\{D_\gamma\}_{\gamma \in K \in \mathcal{K}}$, contains a desirable edge detector output for every scene, that the original image function could possibly represent. Exactly what this means, depends of course on, what is considered to be desirable, and on in which environment the original image function is sampled. We will not discuss these important issues here, but rather limit ourselves, to pick $K_N \subseteq C^l(\Sigma)^{2N}$, $N \in \mathbb{N}_0$ as large as possible with a reasonable number of local constraints on the edge segments. To ensure, that the hypotheses of theorem 3.10.3 are satisfied, we will assume, that the edge cost given by $c_N : S_N \to \overline{\mathbb{R}_+}$, $N \in \mathbb{N}_0$ is C^l-continuous, and choose K_N, $N \in \mathbb{N}_0$ as follows: For each $N \in \mathbb{N}_0$ we first specify a relatively large compact subset of $C^l(\Sigma)^{2N}$. We then intersect this compact set with a closed set in $C^l(\Sigma)^{2N}$, such that the intersection is contained in the set A_N of all admissible N-segment image segmentations. To form K_N, the so obtained intersection is then, if necessary, intersected once more with a subset of S_N, which is closed in $C^l(\Sigma)^{2N}$. It is clear, that K_N so chosen is a compact possibly empty subset of $S_N \cap A_N$. Since $S_0 \subseteq A_0 = C^l(\Sigma)^0 = \{\gamma^0\}$ is trivially compact, we naturally let $K_0 \doteq S_0$. For $N \in \mathbb{N}$ considerably more work is required.

Compactness

In order to specify a large compact subset of $C^l(\Sigma)^{2N}$, we introduce a new family of function spaces. For $l \in \mathbb{N}_0$, $h \in]0,1]$ and $\Omega \subseteq \mathbb{R}^n$ we define $C^{l,h}(\Omega)$ to be the linear subspace of $C^l(\Omega)$ consisting of those functions, whose derivatives of order $\leq l$ are Hölder continuous with exponent h, that is

$$C^{l,h}(\Omega) \doteq \left\{ f \in C^l(\Omega) : \bigvee_{\substack{\alpha \in \mathbb{N}_0^n \\ |\alpha| \leq l}} \sup_{\substack{x,y \in \Omega \\ x \neq y}} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{\|x - y\|^h} < \infty \right\} \tag{3.83}$$

For compact sets $\Omega \subseteq \mathbf{R}^n$ we also define a norm on $C^{l,h}(\Omega)$ according to

$$\|f\|_{C^{l,h}(\Omega)} \doteq \|f\|_{C^l(\Omega)} + \bigvee_{\substack{\alpha \in \mathbf{N}_0^n \\ |\alpha| \leq l}} \sup_{\substack{x,y \in \Omega \\ x \neq y}} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{\|x - y\|^h} \qquad f \in C^{l,h}(\Omega) \tag{3.84}$$

The important property of $C^{l,h}(\Omega)$ has to do with the notion of compact embeddings. We say, that the normed vector space X is *embedded* in the normed vector space Y, and write $X \hookrightarrow Y$, if $X \subseteq Y$ and the identity map $\iota : X \to Y : x \mapsto x$ is continuous. If moreover every bounded set in X is precompact, that is has compact closure, in Y, we say, that X is *compactly embedded* in Y, and write $X \hookrightarrow\hookrightarrow Y$.

Theorem 3.11.3 *Let $l \in \mathbf{N}_0$ and $h \in]0,1]$. If $\Omega \subseteq \mathbf{R}^n$ is compact, then*

$$C^{l,h}(\Omega) \hookrightarrow\hookrightarrow C^l(\Omega)$$

A proof of this result can be found in [55, p11].

To make use of the embedding theorem above, for $l \in \mathbf{N}_0$, $h \in]0,1]$ and $r \geq 0$ we define the set

$$K_{l,h,r} \doteq \left\{ f \in C^{l,h}(\Sigma) : \|f\|_{C^l(\Sigma)} \vee \sup_{\substack{\sigma,\tau \in \Sigma \\ \sigma \neq \tau}} \frac{|f^{(l)}(\sigma) - f^{(l)}(\tau)|}{|\sigma - \tau|^h} \leq r \right\} \tag{3.85}$$

Fact 3.11.4 *Let $f \in K_{l,h,r}$, where $l \in \mathbf{N}_0$, $h \in]0,1]$ and $r \geq 0$. Then $\|f\|_{C^{l,h}(\Sigma)} \leq 2r$.*

Proof: If $l = 0$, it follows trivially from (3.85) that

$$\|f\|_{C^{l,h}(\Sigma)} = \|f\|_{C^0(\Sigma)} + \sup_{\substack{\sigma,\tau \in \Sigma \\ \sigma \neq \tau}} \frac{|f^{(1)}(\sigma) - f^{(1)}(\tau)|}{|\sigma - \tau|^h} \leq 2r$$

If $l \in \mathbf{N}$, for $i = 1, \ldots, l-1$ we have by the mean value theorem that

$$|f^{(i)}(\sigma) - f^{(i)}(\tau)| \leq \sup_{\varsigma \in \Sigma} |f^{(i+1)}(\varsigma)||\sigma - \tau| \leq \|f\|_{C^l(\Sigma)} |\sigma - \tau|^h \qquad \forall \sigma, \tau \in \Sigma$$

Hence by (3.85)

$$\|f\|_{C^{l,h}(\Sigma)} \leq \|f\|_{C^l(\Sigma)} + \left(\|f\|_{C^l(\Sigma)} \vee \sup_{\substack{\sigma,\tau \in \Sigma \\ \sigma \neq \tau}} \frac{|f^{(l)}(\sigma) - f^{(l)}(\tau)|}{|\sigma - \tau|^h} \right) \leq 2r$$

■

Fact 3.11.5 *Let $l \in \mathbb{N}_0$ and $n \in \mathbb{N}$. Then $K^n_{l,h,r}$ is compact in $C^l(\Sigma)^n$ for all $h \in]0,1]$ and $r \geq 0$.*

Proof: We recall, that the $C^l(\Sigma)^n$(-norm)-topology is identical to the product topology on $C^l(\Sigma)^n = \prod_{k=1}^n C^l(\Sigma)$, [60, p121]. By the Tychonoff theorem it is therefore sufficient to show, that $K_{l,h,r}$ is compact in $C^l(\Sigma)$. By fact 3.11.4 $K_{l,h,r}$ is bounded in $C^{l,h}(\Sigma)$. Since Σ is compact, theorem 3.11.3 then implies, that $K_{l,h,r}$ is precompact in $C^l(\Sigma)$. Suppose $f \in C^l(\Sigma) \setminus K_{l,h,r}$. Then either $\exists\ i \in \{0,\ldots,l\}$ and $\sigma \in \Sigma$, such that $|f^{(i)}(\sigma)| > r$, or $\exists\ \sigma,\tau \in \Sigma$, such that $|f^{(l)}(\sigma) - f^{(l)}(\tau)| > r|\sigma - \tau|^h$. In either case $C^l(\Sigma) \setminus K_{l,h,r}$ contains an open $C^l(\Sigma)$-ball centered at f. Indeed, in the first case

$$B_{C^l(\Sigma)}\left(f, |f^{(i)}(\sigma)| - r\right) \subseteq C^l(\Sigma) \setminus K_{l,h,r}$$

whereas in the second case

$$B_{C^l(\Sigma)}\left(f, \frac{|f^{(l)}(\sigma) - f^{(l)}(\tau)| - r|\sigma - \tau|^h}{2}\right) \subseteq C^l(\Sigma) \setminus K_{l,h,r}$$

Thus $K_{l,h,r}$ is closed, and hence compact in $C^l(\Sigma)$. ■

We remark, that $\lim_{h \downarrow 0} \lim_{r \uparrow \infty} K^{2N}_{l,h,r}$ is exactly the mildly restricted linear subspace of $C^l(\Sigma)^{2N}$, consisting of those image segmentations, whose lth derivative is Hölder continuous. Fact 3.11.5 therefore solves the problem of finding a large compact subset of $(K^{2N}_{l,h,r})$ of $C^l(\Sigma)^{2N}$ for $l \in \mathbb{N}_0$ and $N \in \mathbb{N}$. Obviously the "radius" r should be chosen large compared with the dimensions of the image domain. For the Hölder exponent on the other hand any choice $h \in]0,1]$ makes sense. The smaller $h > 0$, the less restrictive the smoothness constraint imposed by the Hölder condition.

Admissibility

In order to specify a closed subset of admissible image segmentations in $K^{2N}_{l,h,r}$, we will impose a number of closed constraints on the image segmentations in $C^l(\Sigma)^{2N} \supseteq K^{2N}_{l,h,r}$. These can be classified as constraints of *edge segment interconnection*, *regularity*, *image boundary intersection* and *edge segment intersection*. The interconnection constraint ensures, that the curves forming the discontinuity set do not, at any point, have the same component of the corresponding continuity set on both sides, or equivalently that none of the components of the continuity set lies on both sides of any portion of its boundary.

The regularity and various intersection constraints are necessary to avoid cusps, which are too sharp to be modeled as the graph of a Lipschitz continuous function. For an image segmentation $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C^l(\Sigma)^{2N}$, $l, N \in \mathbf{N}$, of an image domain $B \doteq]a,b[\times]c,d[$ the constraints above can be described in terms of four small strictly positive constants $\omega, \delta_0 > 0$ and $\delta_1, v \in]0,1[$ as follows:

Interconnection:

We will demand, that each endpoint of the edge segments $\gamma_n(\Sigma)$, $n = 1,\ldots,N$, either coincides with at least one other endpoint, or that it lies on the boundary ∂B of the image domain. Endpoints, which are constrained to coincide, are said to form a *node.* The endpoints, which are constrained to lie on the boundary ∂B, are also said to form a node. This special node, referred to as the *boundary node*, will be denoted by $\mathcal{N}_{\partial B}$. The boundary node is of course allowed to be empty, that is the boundary may be separated from all the edge segments. To distinguish constrained endpoint coincidences from accidental ones, we associate with each ordered pair

$$(n,s) \in E_N \doteq \{1,\ldots,N\} \times \{0,1\}$$

the endpoint $\gamma_n(s)$ of the edge segment $\gamma_n(\Sigma)$. Each node can then be identified with a subset of the *endpoint space* E_N, and any given edge segment interconnection with a directed graph $\mathcal{I}$, referred to as the (edge segment) interconnection graph, whose nodes just represent these subsets of E_N, and whose branches $\mathcal{B}(1),\ldots,\mathcal{B}(N)$ represent the directed edge segments parametrized by $\gamma_n : \Sigma \to \mathbf{R}^2$, $n = 1,\ldots,N$. Figure 3.1 shows the simple correspondence between an 8-segment image segmentation (without accidental endpoint coincidences) and its interconnection graph.

The nodes associated with any possible interconnection are obviously disjoint, and their union equals the entire endpoint space E_N. The nodes therefore partition E_N into equivalence classes of endpoints. The unique node, which contains the endpoint (n,s), will be denoted by $\mathcal{N}(n,s)$. As usual we write $(n,s) \sim (p,t)$ to indicate, that the endpoints (n,s) and (p,t) are equivalent, that is belong to the same node. If $(n,s) \sim (p,t) \in E_N \setminus \mathcal{N}_{\partial B}$ and $(n,s) \neq (p,t)$, we say, that (n,s) *joins* (p,t), and write $(n,s) \bowtie (p,t)$.

To describe which interconnection graphs correspond to admissible image segmentations, we introduce the following notion of *connectedness.* Consider a subset $\mathcal{J}$ of the

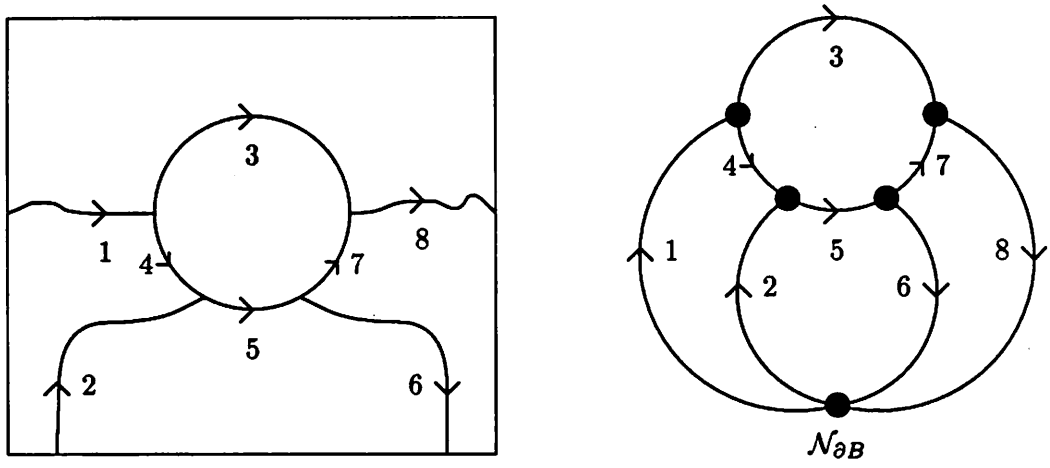


Figure 3.1: 8-segment image segmentation (a) and corresponding interconnection graph (b).

elements of an interconnection graph $\mathcal{I}$, that is a subset of the branches and nodes of $\mathcal{I}$. A branch $\mathcal{B}$ and a node $\mathcal{N}$ of $\mathcal{J}$ are said to be *directly connected* iff there exists an endpoint $(n, s) \in E_N$, such that $\mathcal{B} = \mathcal{B}(n)$ and $\mathcal{N} = \mathcal{N}(n, s)$. This defines a *relation of "direct connectedness"* on the elements of $\mathcal{J}$. Two possibly identical elements of $\mathcal{J}$ are then said to be *connected*, iff they are equivalent with respect to the equivalence relation generated by the relation of direct connectedness. We naturally refer to the equivalence classes of the so connected elements of $\mathcal{J}$ as the *connected components* of $\mathcal{J}$. If we draw the nodes as circles, and the branches as curve segments between these circles, not touching each other, these notions coincide with the intuitive concepts of connectedness and connected components (of the drawing). Figure 3.2 shows a graph $\mathcal{I}$ consisting of a single component and a subset of $\mathcal{I}$ with three components.

The interconnection constraint can now be expressed as follows:

(I1) The image segmentation γ satisfies constraints, which can be represented by a directed graph $\mathcal{I}$ as described above.

(I2) The number of connected components of $\mathcal{I}$ cannot be increased by the removal of any single branch or node, except for the boundary node.

The exception of the boundary node is essential in order to allow for curves, such as horizons, to cut through the image.

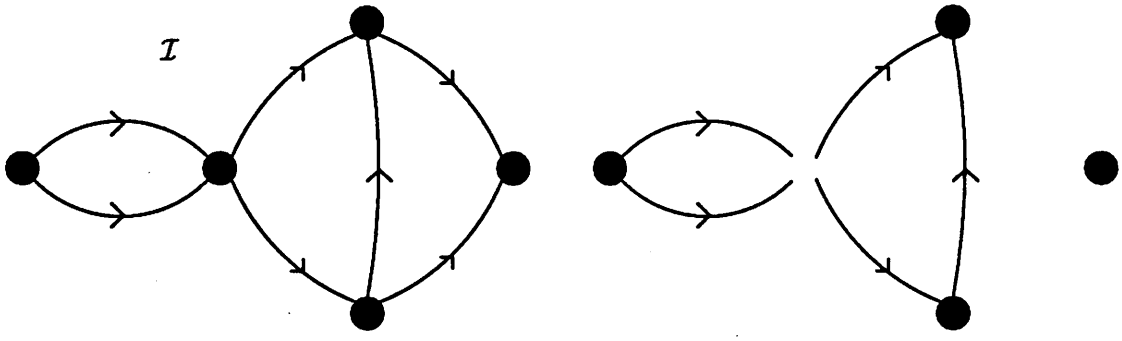


Figure 3.2: Connected components of subset of directed graph. (a) Directed graph $\mathcal{I}$ with one component. (b) Subset of $\mathcal{I}$ with three components.

Regularity:

(R1) $\|\dot{\gamma}_n(\sigma)\| \geq \omega \qquad \forall \sigma \in \Sigma \qquad n = 1, \ldots, N$

Image boundary intersection:

(B1) For $n = 1, \ldots, N$ and $\forall \sigma \in \Sigma$ at least one of the conditions (i) and (ii), and at least one of the conditions (iii) and (iv) below are satisfied.

(i) $a + \delta_0 \leq \gamma_{n1}(\sigma) \leq b - \delta_0$

(ii) $|\dot{\gamma}_{n2}(\sigma)| \leq (1 - \delta_1)\|\dot{\gamma}_n(\sigma)\|$

(iii) $c + \delta_0 \leq \gamma_{n2}(\sigma) \leq d - \delta_0$

(iv) $|\dot{\gamma}_{n1}(\sigma)| \leq (1 - \delta_1)\|\dot{\gamma}_n(\sigma)\|$

(B2) $\|\gamma_n(s) - x\| \geq \delta_0 \qquad \forall x \in \partial B \qquad \forall (n, s) \in E_N \setminus \mathcal{N}_{\partial B}$

Stated in words the first constraint (B1) just says, that any given point on the edge segment $\gamma_n(\Sigma)$ is at least a certain distance away from ∂B, or has a tangent, whose direction differs at least a certain amount from that of ∂B. The second constraint (B2) inhibits, as we shall see later, completely the possibility of intersection of edge segments right on the image boundary ∂B.

Edge segment intersection:

To describe this constraint without a long list of separate cases, we first define a number of subsets of Σ^2. Thus for $(n,s),(p,t) \in E_N$ we define the sets

$$\Upsilon_{np}(s,t) \doteq \begin{cases} \{(\sigma,\tau) \in \Sigma^2 : |\sigma - \tau| < v\} & \text{if } (n,s) = (p,t) \\ \{(\sigma,\tau) \in \Sigma^2 : |\sigma - s| + |\tau - t| < v\} & \text{if } (n,s) \bowtie (p,t) \\ \varnothing & \text{otherwise} \end{cases} \tag{3.86}$$

Obviously the sets $\Upsilon_{nn}(s,s)$, $s = 0,1$, $n = 1,\ldots,N$ are identical. This redundance is kept for compact notation. Next for $n,p \in \{1,\ldots,N\}$ we define the set

$$T_{np} \doteq \Sigma^2 \setminus \bigcup_{s=0}^{1} \bigcup_{t=0}^{1} \Upsilon_{np}(s,t) \tag{3.87}$$

The edge segment intersection constraint can now be expressed as follows:

(E1) $\|\gamma_n(\sigma) - \gamma_p(\tau)\| \geq \delta_0 \qquad \forall(\sigma,\tau) \in T_{np} \qquad n,p \in \{1,\ldots,N\}$

(E2) For $(n,s) \bowtie (p,t) \in E_N$

$$(-1)^{s+t}\dot{\gamma}_n(s)^T\dot{\gamma}_p(t) \leq (1-\delta_1)\|\dot{\gamma}_n(s)\|\|\dot{\gamma}_p(t)\|$$

Similarly to the image boundary constraint (B1) above, the constraint (E1) basically implies, that any two points on any of the edge segments $\gamma_n(\Sigma)$, $n = 1,\ldots,N$, are at least a certain distance apart. In this case however, the constraint has to be relaxed for points, which are forced to be close to each other, either by the interconnection constraints, or by the fact, that they belong to the same edge segment with a small difference between their respective parameter values σ and τ. These considerations are taken care of by excluding the sets $\bigcup_{s=0}^{1}\bigcup_{t=0}^{1}\Upsilon_{np}(s,t)$ from the sets T_{np}. For example, these relaxations permit the continuation of an edge segment by a second segment in almost any direction. They also permit the formation of a closed curve by a single edge segment with a corner of almost any angle at the coinciding endpoints. In particular smooth continuations and smooth closed curves are allowed.

Having completed the definitions of our interconnection, regularity and intersection constraints, our next step is to show, that these constraints are closed, that is that the subset of image segmentations in $C^l(\Sigma)^{2N}$, which satisfy these constraints, is closed in $C^l(\Sigma)^{2N}$. In appendix B we prove the following:

Fact 3.11.6 *For any $l, N \in \mathbf{N}$ the set of image segmentations in $C^l(\Sigma)^{2N}$, which satisfy the interconnection, regularity and intersection constraints (I1), (I2), (R1), (B1), (B2), (E1) and (E2) above is closed in $C^l(\Sigma)^{2N}$.*

For $l, N \in \mathbf{N}$, $h \in]0,1]$, $r, \omega, \delta_0 > 0$ and $\delta_1 \in]0,1[$ we define $C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$ to be the set of image segmentations in $K^{2N}_{l,h,r}$, which satisfy the interconnection, regularity and intersection constraints (I1), (I2), (R1), (B1), (B2), (E1) and (E2) with the given constants ω, δ_0, δ_1 and v for the image domain $B \doteq]a,b[\times]c,d[$. It follows immediately from the facts 3.11.5 and 3.11.6, that $C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$ is compact in $C^l(\Sigma)^{2N}$.

Our next step is to find conditions, under which the image segmentations in $C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$ are admissible. In appendix B we prove the following:

Fact 3.11.7 *Let $l, N \in \mathbf{N}$, $h \in]0,1]$, $r, \omega, \delta_0 > 0$, $\delta_1, v \in]0,1[$ and assume that*

$$v < \left(\frac{\delta_1 \omega}{2\sqrt{2}r}\right)^{\frac{1}{H}}$$

where $H = h$ if $l = 1$, and $H = 1$ otherwise. Then $C_{l,N}(h, r, \omega, \delta_0, \delta_1, v) \subseteq A_N$, that is the image segmentations in $C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$ are admissible.

3.11.3 Optimal Edge Results

With these preparations we are now ready to give explicit examples of collections of image segmentation domains, which satisfy the hypotheses regarding $\mathcal{K}$ of theorem 3.10.3.

Theorem 3.11.8 *Assume that the edge cost in theorem 3.10.3 is C^l-continuous for some $l \in \mathbf{N}$. Let the constants $h, r, \omega, \delta_0, \delta_1$ and v be as in fact 3.11.7, and for each $N \in \mathbf{N}$ let C_N be a closed subset of S_N. Then theorem 3.10.3 holds with*

$$K_0 \doteq S_0 \tag{3.88a}$$

$$K_N \doteq C_N \cap C_{l,N}(h, r, \omega, \delta_0, \delta_1, v) \qquad N \in \mathbf{N} \tag{3.88b}$$

Proof: We have to show, that the image segmentation domains K_N, $N \in \mathbf{N}_0$, satisfy the hypotheses of theorem 3.10.3. Since $C^l(\Sigma)^0 = \{\gamma^0\}$, where $\gamma^0 : \Sigma \to \mathbf{R}^0$ is the *unique* trivial image segmentation, is finite, $K_0 = S_0 \subseteq C^l(\Sigma)^0$ is trivially compact in $C^l(\Sigma)^0$. The continuity set C_{γ^0} associated with γ^0 is by definition the entire image domain B,

which being rectangular, is certainly of class $C^{0,1}$. Hence by theorem 3.8.2 γ^0, the only possible image segmentation in K_0, is admissible. Thus K_0 satisfies the hypotheses. Next let $N \in \mathbf{N}$. From (3.88) and fact 3.11.7 we see, that $K_N \subseteq S_N \cap A_N$. Since C_N is closed and $C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$ is compact in $C^l(\Sigma)^{2N}$, it also follows, that K_N is compact in $C^l(\Sigma)^{2N}$. Since the edge cost is C^l-continuous and $l > 0$, the $C^l(\Sigma)^{2N}$-topology is stronger than that generated by the edge cost function c_N and the $C(\Sigma)^{2N}$-topology. Thus K_N, $N \in \mathbf{N}$, satisfy the hypotheses as well. ■

Corollary 3.11.9 *Assume that the edge cost in theorem 3.10.3 is C^l-continuous for some $l \in \mathbf{N}$. Let the constants $h, r, \omega, \delta_0, \delta_1$ and v be as in fact 3.11.7. If $C_{l,N}(h, r, \omega, \delta_0, \delta_1, v) \subseteq S_N$, $\forall N \in \mathbf{N}$, then theorem 3.10.3 holds with*

$$\begin{aligned} K_0 &\doteq S_0 \\ K_N &\doteq C_{l,N}(h, r, \omega, \delta_0, \delta_1, v) \qquad N \in \mathbf{N} \end{aligned}$$

Proof: Let $C_N \doteq C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$, $N \in \mathbf{N}$, in theorem 3.11.8. ■

Normally one would assign some edge cost $c_0(\gamma^0) \in \overline{\mathbf{R}_+}$ (most naturally $c_0(\gamma^0) \doteq 0$) to the trivial image segmentation. This means, that $S_0 = \{\gamma^0\} \neq \emptyset$, and hence

$$\mathcal{D}_{\mathcal{K}} \supseteq \{0\} \times \{\gamma^0\} \times \mathcal{H}^1(B) \neq \emptyset$$

in theorem 3.10.3. Thus the case $\mathcal{D}_{\mathcal{K}} = \emptyset$ need *not* be considered.

The corollary above can readily be applied to edge costs composed of any of the general parametrized curve costs presented in chapter 2. In fact, for the edge cost examples in (2.19) and (2.18) we have the following two corollaries:

Corollary 3.11.10 (Existence of C^1-smooth Optimal Edges) *Let the constants h, r, $\omega, \delta_0, \delta_1$ and v be as in fact 3.11.7, and let $\zeta \in L_2(B)$ and the domain $\mathcal{D}_{\mathcal{K}}$ be as in theorem 3.10.3 with*

$$\begin{aligned} K_0 &\doteq C^1(\Sigma)^0 = \{\gamma^0\} \\ K_N &\doteq C_{1,N}(h, r, \omega, \delta_0, \delta_1, v) \qquad N \in \mathbf{N} \end{aligned}$$

If $\nu, \lambda \geq 0$ and $\nu + \lambda, \mu > 0$, then the total cost function

$$c : \mathcal{D}_{\mathcal{K}} \to \overline{\mathbf{R}_+} : (N, \gamma, z) \mapsto \sum_{n=1}^{N} [\nu + \lambda \Lambda(\gamma_n)] + \int_{C_\gamma} \left[(z - \zeta)^2 + \mu \sum_{k=1}^{2} (D_k z)^2 \right] dx$$

attains its minimum.

Proof: Let

$$S_N \doteq \left\{\gamma \in C^1(\Sigma)^{2N} : \|\dot{\gamma}_n(\sigma)\| > \frac{\omega}{2} \quad \forall \sigma \in \Sigma, \quad n = 1, \ldots, N\right\} \qquad N \in \mathbb{N}_0 \qquad (3.89)$$

and let

$$c_N : S_N \to \overline{\mathbb{R}_+} : \gamma = [\gamma_1^T \cdots \gamma_N^T]^T \mapsto \sum_{n=1}^{N} [\nu + \lambda \Lambda(\gamma_n)] \qquad N \in \mathbb{N}_0 \qquad (3.90)$$

Since $S_0 = \{\gamma^0\}$ is a singleton, c_0 is trivially C^1-continuous. Since $S_N \subseteq C^1(\Sigma)^{2N}$ and its 1-range $R_{1,N}(S_N)$ is open in $\mathbb{R}^4$ $\forall N \in \mathbb{N}$, it follows from fact 3.11.2, that c_N, $N \in \mathbb{N}$, are C^1-continuous as well. Moreover from (3.89) we see, that $C_{1,N}(h, r, \omega, \delta_0, \delta_1, v) \subseteq S_N$ $\forall N \in \mathbb{N}$. Hence the hypotheses of corollary 3.11.9 are satisfied. Next from (3.89) and (3.90) we note that

$$\inf_{\gamma \in S_N} c_N(\gamma) \geq \sum_{n=1}^{N} \left(\nu + \lambda \int_\Sigma \frac{\omega}{2} d\sigma\right) \geq (\nu + \lambda)\left(1 \wedge \frac{\omega}{2}\right) N \longrightarrow \infty \qquad \text{as } N \longrightarrow \infty$$

Since $S_0 \neq \emptyset$ implies, that $\mathcal{D}_\mathcal{K} \neq \emptyset$, the corollary therefore follows from theorem 3.10.3. ■

Corollary 3.11.11 (Existence of C^2-smooth Optimal Edges) *Let the constants h, r, $\omega, \delta_0, \delta_1$ and v be as in fact 3.11.7, and let $\zeta \in L_2(B)$ and the domain $\mathcal{D}_\mathcal{K}$ be as in theorem 3.10.3 with*

$$\begin{aligned} K_0 &\doteq C^2(\Sigma)^0 = \{\gamma^0\} \\ K_N &\doteq C_{2,N}(h, r, \omega, \delta_0, \delta_1, v) \qquad N \in \mathbb{N} \end{aligned}$$

If $\nu, \lambda, \varkappa, \iota \geq 0$ and $\nu + \lambda, \mu > 0$, then the total cost function $c : \mathcal{D}_\mathcal{K} \to \overline{\mathbb{R}_+}$ defined by

$$c(N, \gamma, z) \doteq \sum_{n=1}^{N} [\nu + \lambda \Lambda(\gamma_n) + \varkappa \mathrm{K}(\gamma_n) + \iota \Lambda(\gamma_n)\mathrm{K}(\gamma_n)] + \int_{C_\gamma} \left[(z - \zeta)^2 + \mu \sum_{k=1}^{2} (D_k z)^2\right] dx$$

attains its minimum.

Proof: This proof is identical to that of corollary 3.11.10, with the exception that the sets $C_{1,N}(h, r, \omega, \delta_0, \delta_1, v)$ and $R_{1,N}(S_N)$ are everywhere replaced by $C_{2,N}(h, r, \omega, \delta_0, \delta_1, v)$ and $R_{2,N}(S_N)$ respectively. ■

The four results above complete our discussion about the existence of the solution to the global edge detection problem posed in the previous chapter. It is worth noticing, that all these results remain valid if the image segmentation domains K_N, $N \in \mathbb{N}$, are restricted by further constraints, as long as these constraints are closed with respect to the $C^l(\Sigma)^{2N}$-topology for the value of l under consideration. One could for example demand, that the edge segments remain in the closure of the image domain, by adding the constraint, that $D_\gamma \subseteq \overline{B}$, or equivalently that

$$\gamma_n(\sigma) \in \overline{B} \qquad \forall \sigma \in \Sigma \qquad n = 1, \ldots, N \qquad \forall \gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in K_N \qquad \forall N \in \mathbb{N}$$

As another example, if $(n, s) \bowtie (p, t) \in E_N$, one could enforce a smooth continuation of the edge segment $\gamma_n(\Sigma)$ by the segment $\gamma_p(\Sigma)$, by imposing the additional constraint

$$(-1)^s \dot{\gamma}_n(s) + (-1)^t \dot{\gamma}_p(t) = 0$$

on the members of K_N.

3.11.4 Image Segmentation Space Parameters

As far as the parameters $h, r, \omega, \delta_0, \delta_1, v$ and in some sense l are concerned, we have been content with specifying ranges, which ensure, that the total cost function attains its minimum on the resulting domain $\mathcal{D}_\mathcal{K}$. For the implementation of an algorithm, honoring the hypotheses in our existence proof, there is still an issue of selecting appropriate values for these parameters. Since most of these parameters have simple geometric interpretations, this should not cause any problem. The exceptions, if any, are the parameters v and possibly h. For the choice of h there seems to be little to go by. Any value in the range $]0, 1]$ will work theoretically, and one might as well pick the numerically most tractable value, which is likely to be 1. For v the situation is different. If v is chosen too small compared with δ_0, the edge segment intersection constraint (E1) will rule out useful edge segments far from intersecting either themselves or each other, and in the extreme case the set $C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$ will even be empty. To avoid these undesirable conditions, we would like to demand that

$$|\sigma - \tau| = v \Rightarrow \|\gamma_n(\sigma) - \gamma_n(\tau)\| \geq \delta_0 \qquad n = 1, \ldots, N \tag{3.91}$$

We therefore first choose

$$v < \left(\frac{\delta_1 \omega}{2\sqrt{2}r}\right)^{\frac{1}{H}}$$

where $H = h$ if $l = 1$, and $H = 1$ otherwise, as suggested by fact 3.11.7. We then choose

$$\delta_0 \leq \frac{\omega v}{2}$$

Then (3.91) is satisfied. Indeed if $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$ and $|\sigma - \tau| = v$, $\sigma, \tau \in \Sigma$, then

$$\begin{aligned}
&\|\gamma_n(\sigma) - \gamma_n(\tau)\| \geq \\
&\quad \geq \left| \frac{\dot{\gamma}_n(\sigma)^T}{\|\dot{\gamma}_n(\sigma)\|} [\gamma_n(\sigma) - \gamma_n(\tau)] \right| \\
&\quad = \left| \int_\tau^\sigma \frac{\dot{\gamma}_n(\sigma)^T \dot{\gamma}_n(\varsigma)}{\|\dot{\gamma}_n(\sigma)\|} \, d\varsigma \right| \\
&\quad \geq \int_{\sigma \wedge \tau}^{\sigma \vee \tau} \left(\|\dot{\gamma}_n(\sigma)\| - \|\dot{\gamma}_n(\varsigma) - \dot{\gamma}_n(\sigma)\| \right) d\varsigma \\
&\quad \geq (\omega - \sqrt{2} r v^H) v \\
&\quad > \left(\omega - \frac{\delta_1 \omega}{2} \right) v \\
&\quad > \frac{\omega v}{2} \\
&\quad \geq \delta_0 \qquad n = 1, \ldots, N
\end{aligned}$$

The discussion above in part suggests the following selection rules for the constants $l, h, r, \omega, \delta_0, \delta_1$ and v:

1. Select acceptable values for the minimum edge segment arc length $\Lambda_\wedge > 0$, the maximum edge segment arc length $\Lambda_\vee \geq \Lambda_\wedge$ and the minimum intersection angle $\theta \in]0, \pi/2[$, that is the minimum angle between tangents of edge and/or image boundary segments at points, where these segments intersect.

2. Let l be the smallest integer, such that the edge cost is C^l-continuous.

3. Let $h \doteq 1$, unless less smooth edge segments is a necessity.

4. Let $\omega \doteq \Lambda_\wedge$.

5. Let $r \doteq \Lambda_\vee$.

6. Let $\delta_1 \doteq 1 - \cos\theta$.

7. Let $v \lessapprox \left(\frac{\delta_1 \omega}{2\sqrt{2} r} \right)^{\frac{1}{H}}$, where $H = h$ if $l = 1$, and $H = h$ otherwise.

8. Let $\delta_0 \doteq \frac{\omega v}{2}$.

The simple rules above do neither guarantee, that the edge segments can reside anywhere in the image domain $B =]a, b[\times]c, d[$, nor that they can acquire a certain acceptable maximum curvature. To satisfy these conditions, one would have to choose

$$r \doteq \Lambda_\vee \vee |a| \vee |b| \vee |c| \vee |d| \vee \kappa_\vee \omega^2$$

where $\kappa_\vee$ is the maximum curvature. However, any sensible choice of $\Lambda_\wedge$, $\Lambda_\vee$ and B would imply that

$$\Lambda_\wedge \ll \sqrt{(b-a)^2 + (d-c)^2} \leq \Lambda_\vee$$

and that $0 \in B$, in which case $\Lambda_\vee \geq |a| \vee |b| \vee |c| \vee |d|$, and the ratio of the resulting "minimum radius of curvature" and the "radius of the image domain" is given by the expression

$$\frac{1}{\kappa_\vee \sqrt{(b-a)^2 + (d-c)^2}} = \frac{\omega^2}{r\sqrt{(b-a)^2 + (d-c)^2}} \leq \frac{\Lambda_\wedge^2}{(b-a)^2 + (d-c)^2} \lll 1$$

Thus under normal circumstances these considerations can safely be neglected.

If specific and independent bounds on location, length, curvature, etc. of the edge segments are desired, the best approach would be to replace the compact set $K^{2N}_{l,h,r}$ in the development of this section by another compact set, which precisely reflects these bounds.

Another consideration neglected by the rules above is, that the selection of δ_0 imposes a minimum distance between nonintersecting edge segments.

However,

$$\delta_0 = \frac{\omega v}{2} < \frac{\omega^2}{4\sqrt{2}r} = \frac{\Lambda_\wedge^2}{4\sqrt{2}\Lambda_\vee}$$

Thus for any sensible choice of $\Lambda_\wedge$ and $\Lambda_\vee$ one has $\delta_0 \ll \Lambda_\wedge$, which should be satisfactory for most purposes.

Chapter 4

An Algorithm for Global Curve-Represented Edge Detection

In the previous two chapters we have presented a paradigm for curve-represented edge detection, and demonstrated the existence of a solution to the resulting optimization problem. So far, however, little has been said about how to find such a solution. In this chapter we present an algorithm intended for that purpose. We begin with a description of the basic strategy. At this level the algorithm is essentially the same for all the total costs that can be composed as weighted sums of the various cost functionals introduced in section 2.2. We then concentrate on the specific cost functional that was used in our computational experiments, and discuss our particular implementation of the algorithm in some detail. Finally we present some of our experimental results.

Many of the concepts and notations that appear in this chapter were introduced earlier, and will be used without repeating their definitions. Unless otherwise stated these definitions can be found in chapter 2.

4.1 General Procedure

One of the main problems with the paradigm presented in chapter 2 is undoubtedly that it is hard to find an image segmentation γ and an estimated image function z

that minimize the total cost $c_{N\zeta}(\gamma, z) \doteq \mathcal{E}_N(\gamma) + \mathcal{D}_{C_\gamma}(z, \zeta) + \mathcal{S}_{C_\gamma}(z)$ (for a given original image function ζ and a given edge segment interconnection of N edge segments). As we recall from section 2.4, the optimality conditions for the edges cannot be used to obtain an optimal image segmentation by simply solving a system of equations. However, given an (N-segment) image segmentation γ, represented either by general parametrized curves or by splines, the continuity set C_γ is easily evaluated. The optimal estimated image function over C_γ can then be found by solving the boundary value problem (2.39) (for z). This does in turn make it possible to compute the image cost density differences $\blacktriangle\varrho_1, \ldots, \blacktriangle\varrho_N$ and hence the total cost variation with respect to the image segmentation itself according to (2.42) or with respect to its defining control vertices according to (2.46) and (2.47). An appropriate adjustment of the image segmentation or the control vertices for lowering the total cost can then be calculated. If the space of image segmentations is given a norm such as (some version of) (2.1) or (2.5), this edge adjustment can furthermore be made in the "direction" of steepest descent.

The discussion above suggests the general procedure displayed in figure 4.1 for solving the global curve-represented edge detection problem posed in section 2.2. The meaning of the general term "the edges", which appears thrice in the flow chart, depends on how the actual edges are represented. If the actual edges are represented by general parametrized curves, it refers to the image segmentation itself, which for example can be represented by a table of sampled values of the edge segment parametrizations $\gamma_1, \ldots, \gamma_N$. The total cost variation with respect to "the edges" is in this case given by the functional in (2.42), which can also be represented in tabular form. If on the other hand the actual edges are represented by splines, "the edges" refer to the independent control vertices, that is the intermediate vertices v_{nm}, $m = 1 + 2o_n, \ldots, M_n - o_n$, $n = 1, \ldots, N$, and the junctions $w_1, \ldots, w_J$ associated with the image segmentation. The total cost variation with respect to "the edges" is then given by (2.46) and (2.47). Finally, the convergence test in the flow chart can refer to either the cost itself or to its variation with respect to the edges. In the latter case it is necessary to have some measure of the magnitude of the variation, which should converge to zero.

For implementation of the procedure in figure 4.1, splines offer a more tractable edge representation than do general parametrized curves. Throughout the rest of this chapter we will therefore only consider spline-represented edges. The total cost variation with respect to the edges can in this case be identified with the total cost gradient with

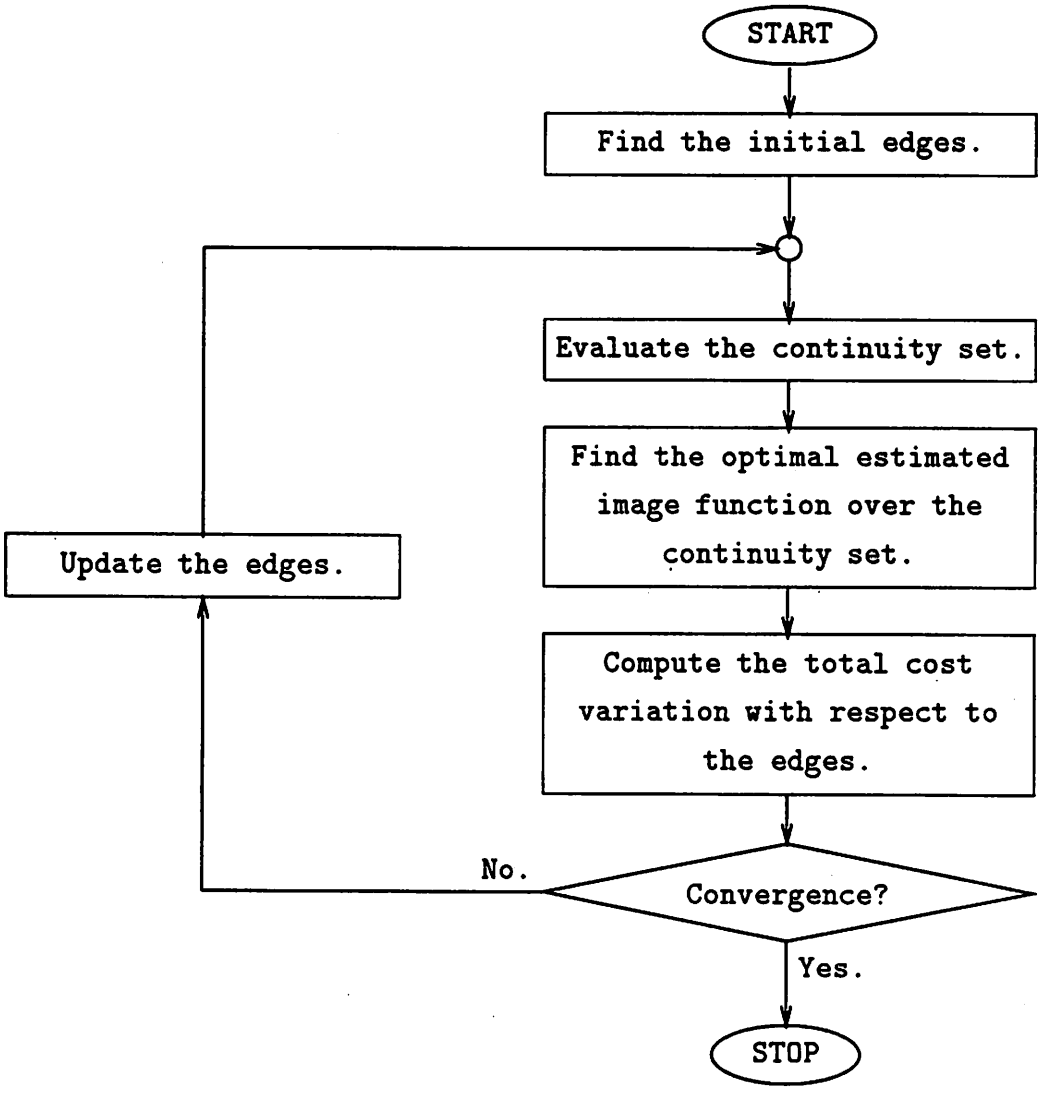


Figure 4.1: General procedure for solving the global curve-represented edge detection problem.

respect to the independent control vertices. The 2×1 block components g_{nm}, $m = 1 + 2o_n, \ldots, M_n - o_n$, $n = 1, \ldots, N$, and $g_1, \ldots g_J$, of this gradient are as we recall given by (2.47). The gradient is good for two purposes: Convergence (of the variation) can be tested by comparing the Euclidean norm of the gradient with a fixed threshold $\varepsilon_v > 0$. If the space of image segmentations is given the norm (2.5), the (negative) gradient also determines the direction of steepest descent. For the global spline-represented edge detection problem the general procedure in figure 4.1 thus naturally takes the form shown in figure 4.2. This procedure is as we see a steepest descent scheme with fixed step size $\beta > 0$. It assumes that the image segmentation configuration $\mathcal{C}$ and hence that the number of edge segments N as well as the number of junctions J are given beforehand. The configuration can either be determined in a preprocessing stage or updated in an outer loop.

4.2 Implementation

In order to verify that the algorithm in figure 4.2 operates as intended, it was implemented in software for the total cost functional

$$c_{N\zeta}(\gamma, z) = \varpi \sum_{n=1}^{N} \sum_{m=0}^{M_n - 1} \|v_{n,m+1} - v_{nm}\| + \int_{C_\gamma} [(z - \zeta)^2 + \mu \|\nabla z^T\|^2] \, dx \qquad (4.1)$$

where the constants $\varpi, \mu > 0$. This cost functional, which is one of the simplest accounted for by the paradigm presented in section 2.2, is obtained by choosing the edge cost (2.20) with $\nu = 0$ and the stabilizer (2.8) specified by $I = 1$, $\mu_0 = 0$ and $\mu_1 = \mu > 0$. With this stabilizer the optimality condition for the estimated image function reduces to (2.40). The image cost density introduced in section 2.3 is of course given by

$$\varrho = (z - \zeta)^2 + \mu \|\nabla z^T\|^2 \qquad (4.2)$$

Since the number of edge segments n is fixed throughout the procedure, the presence cost is constant. The (zero) value of the coefficient ν does hence not affect the algorithm.

In this section we describe our implementation of the algorithm in some detail. Although this particular global curve-represented edge detector is just one of many possible implementations, the description illustrates the kinds of issues that any implementation has to cope with. The description is also of interest for the understanding of our experimental results as well as for anybody who wants to identify the weak points in order to improve on the procedure.

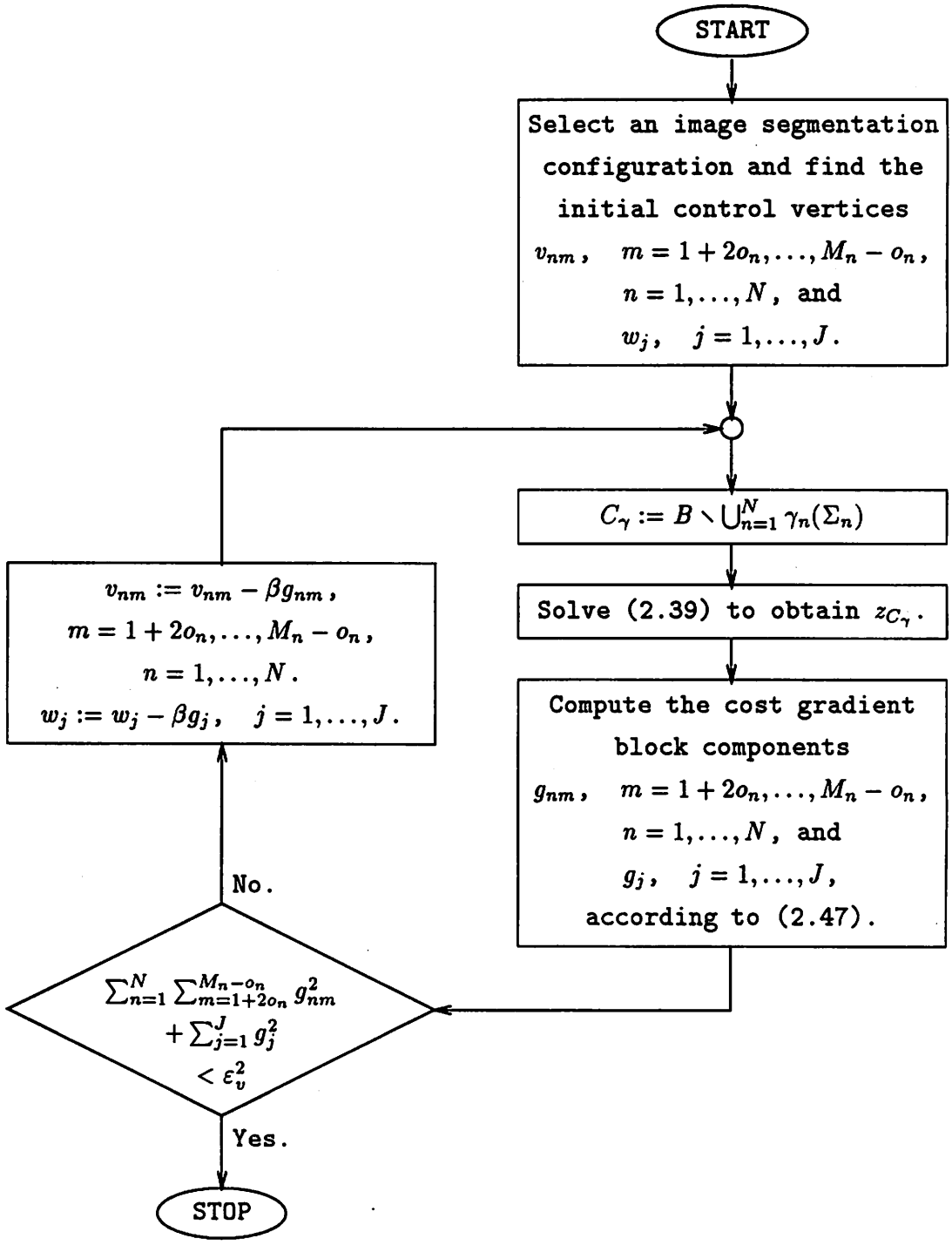


Figure 4.2: Fixed step size steepest descent scheme for solving the global spline-represented edge detection problem.

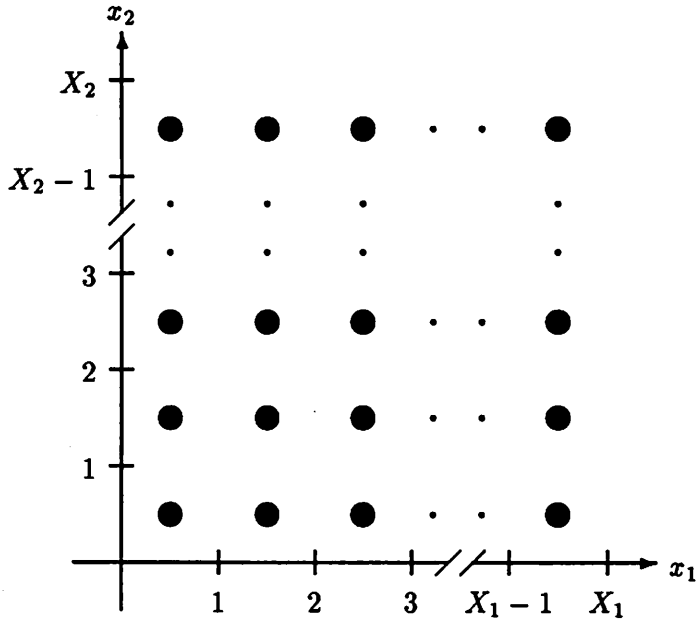


Figure 4.3: Pixel grid for global curve-represented edge detector.

The input to the global edge detector consists of an $X_1 \times X_2$ array of real numbers representing samples of an original image function $\zeta : B \doteq [0, X_1] \times [0, X_2] \rightarrow \mathbf{R}$ on a squared pixel grid $X_\zeta \doteq (\{1, \ldots X_1\} - \frac{1}{2}) \times (\{1, \ldots X_2\} - \frac{1}{2})$. This grid is shown in figure 4.3. Each grid point $x \in X_\zeta$, also referred to as a *pixel site*, can thus be thought of as the center of a square pixel of unit width. The estimated image function z is naturally also represented by an array of samples on the pixel grid X_ζ. The edges are basically represented by a list of control vertex sequences $\langle v_{nm} \rangle_{m=0}^{M_n+2}$, $n = 1, \ldots, N$, specifying N splines $\gamma_1, \ldots, \gamma_N$. This data structure, which is generated by the early processing stages of the global edge detector, will be discussed in more detail shortly.

The output data, which are available after each convergence test (in the flow chart in figure 4.2), consist of

1. the edge cost, the image cost and the total cost.

2. a list of the control vertex sequences representing the edges.

3. an octal image (eight grey levels) displaying the edges, that is the spline curves and possibly their defining control vertices.

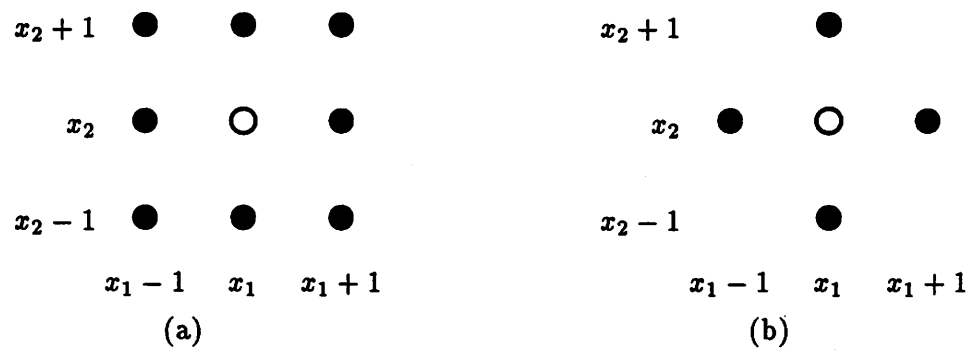


Figure 4.4: Neighboring pixel sites (filled circles) of center pixel site $[x_1 \quad x_2]^T$ (empty circle). (a) Eight-connected neighbors. (b) Four-connected neighbors.

4. a grey level image displaying the estimated image (function).

Before we continue the description of the global edge detector, we review two common notions of connectedness among rectangular pixels on a rectangular pixel grid. Two (distinct) rectangular pixels are said to be *eight-connected* if their boundaries have one side or one corner in common. In the former case they are also said to be *four-connected.* If two pixels are eight- or four-connected, we also say that their sites are eight- or four-connected respectively. Two pixel sites $x, y \in X_\zeta$ are thus eight-connected iff $\|x-y\|_\infty = 1$, and four-connected iff $\|x-y\|_1 = 1$. Figure 4.4 shows a pixel site $x = [x_1 \quad x_2]^T \in X_\zeta$ (in the center) with its eight- and four-connected neighbors.

The organization of the rest of this section follows the block structure of the global edge detector quite closely. Indeed, each of the four subsections corresponds to one of the conceptual subroutines in the flow chart in figure 4.5. The convergence test and the edge update subroutines are simple enough to be adequately described by the flow chart in figure 4.2. They will therefore not be discussed any further.

4.2.1 The Initial Edge Finder

The goal of the initial edge finder is to find a starting point for the steepest descent procedure, that is the loop that follows. Conceptually this involves the two following separate tasks:

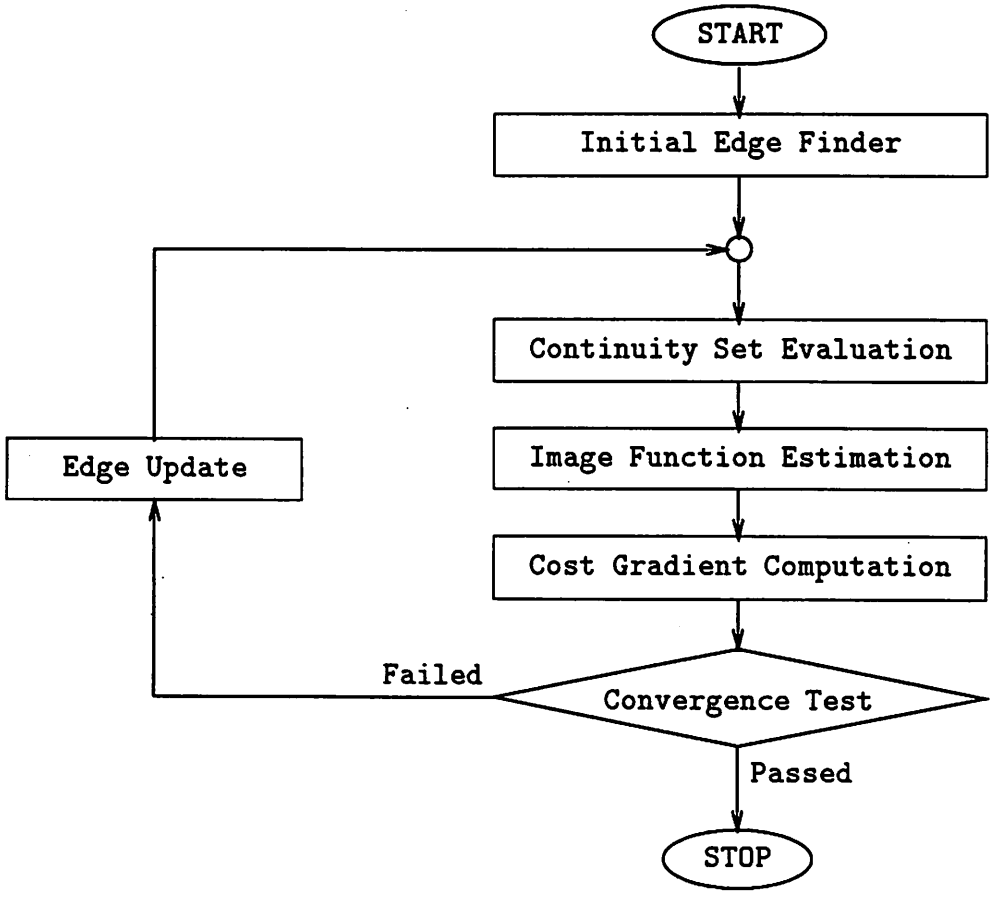


Figure 4.5: Subroutine oriented flow chart of global curve-represented edge detector.

1. Find an appropriate image segmentation configuration.

2. Select reasonable initial values for the independent control vertices, that is the intermediate vertices and the junctions.

In practice these two tasks are carried out in parallel as a data structure representing the image segmentation is grown spline by spline. While the operation by which this data structure is obtained has a significant influence on the output of the initial edge finder, it is by no means central to the ultimate design of the global edge detector. The initial edge finder is in its present status just *one possible* tool for getting the steepest descent procedure started, and can as such almost certainly be improved. A fancier version of the global edge detector might even include an additional outer loop in which the image segmentation configuration is being updated. Since the initial edge finder is also quite intricate, we will here only discuss the data structure that it generates as its output. A description of the initial edge finder operation can be found in appendix C.

The data structure generated by the initial edge finder basically consists of a list of (initial) independent control vertices and a list of splines representing the image segmentation configuration. A diagram of these lists and a typical example of their structural relationships is shown in figure 4.6. The *vertex list*, which originally contains the initial independent control vertices selected by the initial edge finder, will later repeatedly be adjusted by the edge update routine inside the loop of the steepest descent procedure. It will thus always contain the *current* values of the independent control vertices. The *spline list* on the other hand is completely determined by the image segmentation configuration, and therefore set once and for all by the initial edge finder.

Each spline in the spline list is represented by a sequence of pointers, each of which points to one of the independent control vertices in the vertex list. The pointer sequence thereby defines a control vertex sequence, which in turn specifies the spline. This unnecessarily complicated representation of the splines themselves has the advantage of simultaneously specifying the image segmentation configuration in a way that, as we later shall see, also supports the computation of the (total) cost gradient. The splines can of course be either closed or open. In the spline list in figure 4.6, for example, the first two splines are open while the last spline is closed.

In addition to the pointer sequence each spline record in the spline list also contains a reference to one of five procedures to be used by the cost gradient computation routine.

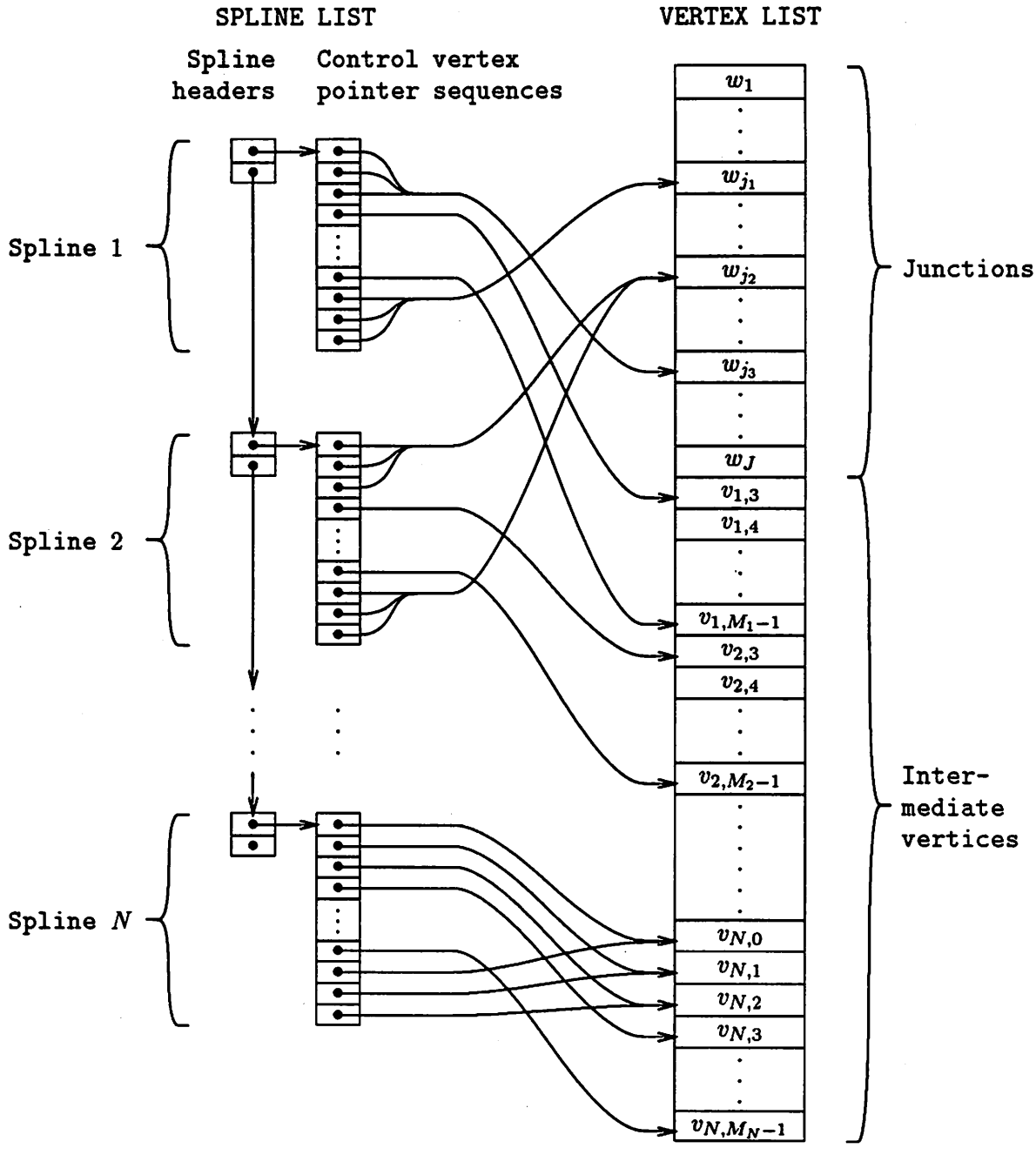


Figure 4.6: Basic output data structure of the initial edge finder. ($1 < j_1 < j_2 < j_3 < J$.)

Conceptually this reference is equivalent to three binary, so called, *type variables*, o, β and τ, which indicate whether the spline is closed or open, and if open, whether its end vertices are *free* or *constrained* to coincide with any other end vertex. The *openness variable* o, which was introduced in section 2.1, takes the value 0 if the spline is closed, and the value 1 if the spline is open. The *end condition variables* β and τ are only defined for open splines and associated with the *beginning* end vertex $v_0 = v_1 = v_2$ and the *terminating* end vertex $v_M = v_{M+1} = v_{M+2}$ respectively. An end condition variable associated with a free end vertex takes the value 0. One that is associated with a constrained end vertex takes the value 1.

The type variables can indeed be derived from the pointer sequences in the spline list, and are hence redundant as far as the representation of the image segmentation configuration is concerned. However, the end condition variables for any given spline depend on *all* the pointer sequences, so the derivation is far from immediate. Since the type variables remain constant and are frequently referred to during the course of the steepest descent procedure, they are therefore precomputed by the initial edge finder.

4.2.2 Continuity Set Evaluation

Before the boundary value problem (2.40) can be solved for the optimal estimated image function (over the current continuity set), the domain, that is the continuity set C_γ and its boundary ∂C_γ have to be found. Since (2.40) must be solved numerically, this boils down to determining which grid points (of the numerical method) that are inside C_γ, and how ∂C_γ affects the computational molecules centered at those grid points. The appropriate way of representing C_γ and ∂C_γ thus depends on how the boundary value problem (2.40) is discretized. The numerical method that we use, naturally uses the grid X_ζ on which the original image function ζ is sampled. Its basic computational molecules are of the form depicted in figure 4.7.

Representation

Since the discontinuity set D_γ is a null set (in $\mathbf{R}^2$) and hence has empty interior, the situation when the finite grid X_ζ intersects D_γ, is very rare, and can furthermore always be circumvented by an arbitrarily small perturbation of X_ζ. We shall therefore simply assume that $X_\zeta \subseteq B \setminus D_\gamma = C_\gamma$. The representation of C_γ is thereby trivially provided by the grid

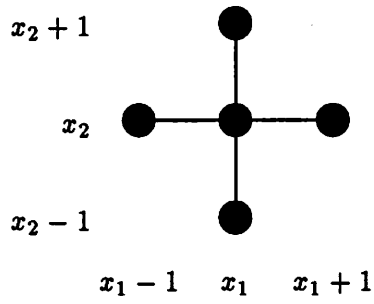


Figure 4.7: Basic computational molecule.

X_ζ alone.

The representation of ∂C_γ can be explained in terms of *(computational molecule) bonds.* In the context of the numerical method that we are using, a bond is simply a line segment between a pixel site and one of its four-connected neighbors, (or would-be neighbors if the pixel site in question is next to the boundary of the image domain.) Each pixel site $x \in X_\zeta$ is thus attached to four bonds whose centers form the set $x + \Xi$ where

$$\Xi \doteq \{\xi \in \mathbf{Z}^2 : \|\xi\|_1 = 1\} \tag{4.3}$$

Figure 4.8 shows the pixel sites in the grid X_ζ and their associated bonds. The *bond centers* are marked with dashes. The square regions enclosed by the bonds (and the dashed box joining the would-be pixel sites immediately outside the image domain) are referred to as *cells.* The cells will as the pixels be labeled by their centers, which of course are points in $\{0,\ldots,X_1\} \times \{0,\ldots,X_2\} \subseteq \mathbf{N}_0^2$. They are also subject to notions of eight- and four-connectedness similar to those concerning the pixels.

A bond that intersects ∂C_γ, is said to be *broken.* A bond that is not broken is said to be *intact.* The boundary ∂C_γ is represented by a binary "continuity control function" w, defined on the bond centers, that is $w : [\{0,\ldots,X_1\} \times (\{1,\ldots,X_2\} - \frac{1}{2})] \cup [(\{1,\ldots,X_1\} - \frac{1}{2}) \times \{0,\ldots,X_2\}] \to \{0,1\}$. On the centers of the broken bonds w takes the value 0, and on the centers of the intact bonds w takes the value 1.

Evaluation Procedure

From the intuitively obvious fact B.2.7 we know that the boundary of the continuity set is given by $\partial C_\gamma = \partial B \cup (B \cap D_\gamma)$. In order to determine the function w so that it

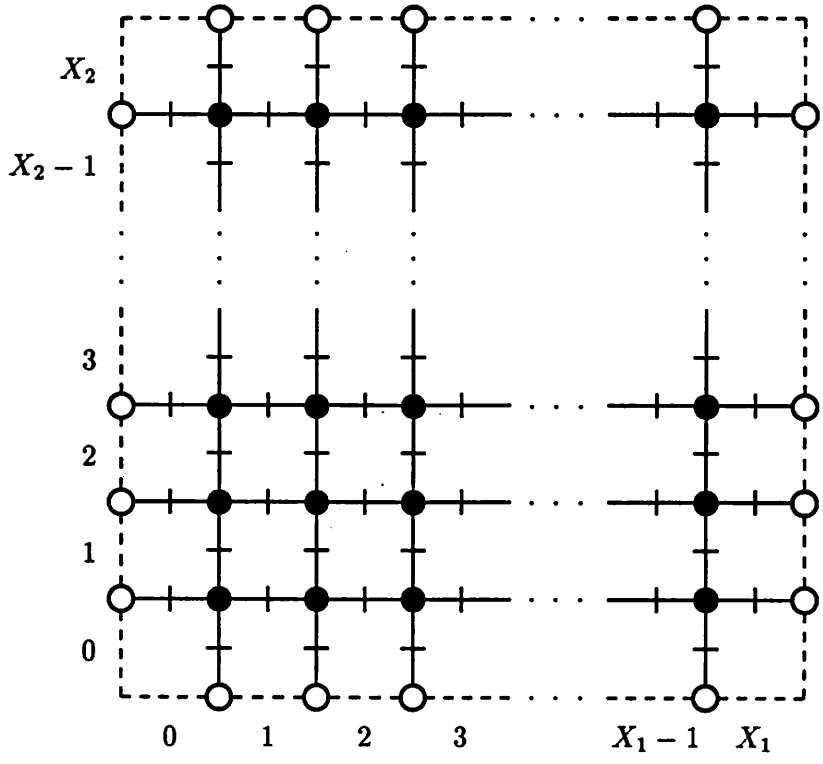


Figure 4.8: Pixel sites (filled circles), would-be pixel sites (empty circles) and bonds (line segments crossed by a dash at the center).

represents this set, the array holding the values of w is first initialized so as to represent ∂B. In other words, all the bonds intersecting ∂B are recorded broken by setting

$$w\left(0, x_2 - \frac{1}{2}\right) = w\left(X_1, x_2 - \frac{1}{2}\right) = 0 \qquad x_2 = 1, \ldots, X_2$$

$$w\left(x_1 - \frac{1}{2}, 0\right) = w\left(x_1 - \frac{1}{2}, X_2\right) = 0 \qquad x_1 = 1, \ldots, X_1$$

and all the other (interior) bonds are recorded intact by setting

$$w\left(x_1, x_2 - \frac{1}{2}\right) = 1 \qquad x_1 = 1, \ldots, X_1 - 1 \qquad x_2 = 1, \ldots, X_2$$

$$w\left(x_1 - \frac{1}{2}, x_2\right) = 1 \qquad x_1 = 1, \ldots, X_1 \qquad x_2 = 1, \ldots, X_2 - 1$$

The second and much less trivial step is to record all the bonds intersecting $B \cap D_\gamma$ as broken. This is done by sampling each function γ_{nm}, $m = 0, \ldots, M_n - 1$, $n = 1, \ldots, N$, in (2.4) at a number of points $0 = \sigma_{nm0} < \sigma_{nm1} < \cdots < \sigma_{nmL_{nm}} = 1$, chosen so that any two successive samples belong to two (distinct) four-connected cells. Denoting (the center of) the cell that contains $\gamma_{nm}(\sigma_{nml})$ by c_{nml}, $l = 0, \ldots, L_{nm}$, we thus have $\|c_{nm,l+1} - c_{nml}\|_1 = 1$, $l = 0, \ldots, L_{nm} - 1$. The bonds between (four-connected) cells containing successive samples are then recorded broken, that is we set

$$w\left(\frac{c_{nml} + c_{nm,l+1}}{2}\right) = 0 \qquad l = 0, \ldots, L_{nm} - 1 \qquad m = 0, \ldots, M_n - 1 \qquad n = 1, \ldots, N$$

The evaluation of the samples $\gamma_{nm}(\sigma_{nml}) = \sum_{r=0}^{3} v_{n,m+r} b_r(\sigma_{nml})$, $l = 0, \ldots, L_{nm}$, $m = 0, \ldots, M_n - 1$, $n = 1, \ldots, N$, normally requires access to a large number of sample values of the basis functions $b_0, \ldots, b_3$. Since the evaluation moreover takes place inside the steepest descent loop, it is therefore sped up by having the cubic polynomials $b_0, \ldots, b_3$ in (2.3) tabulated in a precomputed array.

4.2.3 Image Function Estimation

Discretization

The optimal estimated image function over a given continuity set is, as we recall from section 2.4, given by the solution to the boundary value problem

$$z - \mu \Delta z = \zeta \qquad \text{on } C_\gamma \tag{4.4a}$$

$$\frac{\partial z}{\partial e_n} = 0 \qquad \text{on } \partial C_\gamma \tag{4.4b}$$

In practice this problem must of course be solved numerically, and for this purpose both the Laplacian operator in (4.4a) and the normal differentiation operator in (4.4b) have to be approximated by finite difference operators. Since the image function estimation is by far the most computationally expensive subroutine of the global edge detector, it is important to keep the expressions of the numerical method as simple as possible. For this reason the simplest possible finite difference approximations were chosen.

For the Laplacian operator we use the five-point molecule approximation given by

$$\Delta z(x) \approx \sum_{\xi \in \Xi} z_\xi(x) - 4z(x) = \sum_{\xi \in \Xi} [z_\xi(x) - z(x)] \tag{4.5}$$

where the displacement set Ξ as before is defined by (4.3), and $z_\xi(x)$ is the actual or an extrapolated value of z at the point $x + \xi$. If the bond between the pixel sites x and $x + \xi$ is intact, we obviously choose the actual value, that is

$$z_\xi(x) \doteq z(x + \xi) \text{ if } w\left(x + \frac{\xi}{2}\right) = 1 \tag{4.6}$$

If on the contrary the bond between x and $x + \xi$ is broken, then the two pixel sites x and $x + \xi$ are considered to be separated by the boundary ∂C_γ. In this case $z_\xi(x)$ must be obtained by extrapolating z across ∂C_γ.

In the interest of simplicity we make the crude assumption that ∂C_γ intersects the bonds at right angles. At the intersection with a (broken) bond centered at $x + \frac{\xi}{2}$, ($\xi \in \Xi$,) the normal derivative of z in the ξ-direction can then be approximated according to

$$\frac{\partial z}{\partial e_n} \approx z_\xi(x) - z(x)$$

When applied to the Neumann condition (4.4b), this approximation yields the discrete boundary conditions

$$z_\xi(x) = z(x) \text{ if } w\left(x + \frac{\xi}{2}\right) = 0$$

which combined with (4.6) leads us to the definition

$$z_\xi(x) \doteq z\left(x + w\left(x + \frac{\xi}{2}\right)\xi\right) \qquad x \in X_\zeta \qquad \xi \in \Xi \tag{4.7}$$

Substituting (4.7) in (4.5) we now obtain

$$\Delta z(x) \approx \sum_{\xi \in \Xi} z\left(x + w\left(x + \frac{\xi}{2}\right)\xi\right) - 4z(x) = \sum_{\xi \in \Xi} w\left(x + \frac{\xi}{2}\right)[z(x + \xi) - z(x)]$$

The two equivalent discrete approximations of the boundary value problem (4.4) that these expressions suggest are finally given by

$$(1+4\mu)z(x) - \mu \sum_{\xi \in \Xi} z\left(x + w\left(x + \frac{\xi}{2}\right)\xi\right) = \zeta(x) \qquad x \in X_\zeta \tag{4.8}$$

and

$$[1+\omega(x)\mu]z(x) - \mu \sum_{\xi \in \Xi} w\left(x + \frac{\xi}{2}\right) z(x+\xi) = \zeta(x) \qquad x \in X_\zeta \tag{4.9}$$

where

$$\omega(x) \doteq \sum_{\xi \in \Xi} w\left(x + \frac{\xi}{2}\right)$$

The linear system (4.9) can easily be rewritten on the matrix form

$$A\vec{z} = \vec{\zeta} \tag{4.10}$$

Indeed, let

$$\vec{\zeta}_{k(x)} \doteq \zeta(x) \qquad x \in X_\zeta \tag{4.11a}$$

$$\vec{z}_{k(x)} \doteq z(x) \qquad x \in X_\zeta \tag{4.11b}$$

where

$$k(x) \doteq \left(x_1 - \frac{1}{2}\right) X_2 + x_2 + \frac{1}{2}$$

Then A is readily seen to be a symmetric matrix with diagonal elements $A_{k(x),k(x)} = 1 + \omega(x)\mu$, $x \in X_\zeta$. All the other nonzero elements are equal to $-\mu$, and the sum of the elements in each row equals 1. The symmetric matrix A is thus both real and diagonally dominant. By Geršgorin's circle theorem [62, p371] it is therefore also strictly positive definite.

Estimation Procedure

The algorithm that we use for solving the linear system (4.10), was obtained by rewriting (4.8) as

$$z(x) = \frac{1}{1+4\mu}\left[\mu \sum_{\xi \in \Xi} z\left(x + w\left(x + \frac{\xi}{2}\right)\xi\right) + \zeta(x)\right] \qquad x \in X_\zeta$$

and then interpreting this equation as a component by component update law for the vector $\vec{z}$. This update law, which is extremely simple to implement on a computer, is equivalent

to the following (locally underrelaxed) iteration scheme:

$$
\begin{aligned}
&z^{(i+1)}(x) \doteq \\
&\quad \doteq \frac{1}{1+4\mu}\left(\mu w\left(x_1-\frac{1}{2},x_2\right) z^{(i+1)}(x_1-1,x_2)+\mu w\left(x_1,x_2-\frac{1}{2}\right) z^{(i+1)}(x_1,x_2-1)\right. \\
&\qquad +\mu w\left(x_1+\frac{1}{2},x_2\right) z^{(i)}(x_1+1,x_2)+\mu w\left(x_1,x_2+\frac{1}{2}\right) z^{(i)}(x_1,x_2+1) \\
&\qquad \left.+\mu[4-\omega(x)]z^{(i)}(x)+\zeta(x)\right) \qquad (4.12)
\end{aligned}
$$

The central question at this point is of course whether the algorithm above converges to a solution of (4.10). Fortunately the answer is affirmative. In order to demonstrate this fact we make use of the following theorem from [63, p355].

Theorem 4.2.1 *Suppose $\eta \in \mathbf{R}^K$ and $F-G \in \mathbf{R}^{K\times K}$ is nonsingular. If F is nonsingular and the spectral radius of $F^{-1}G$ is strictly less than unity, then the sequence $\langle y^{(i)}\rangle_{i\in\mathbf{N}_0}$ defined by $Fy^{(i+1)} = Gy^{(i)}+\eta$ converges to $(F-G)^{-1}\eta$ for any starting vector $y^{(0)} \in \mathbf{R}^K$.*

Defining the vectors $\vec{z}^{(i)} \in \mathbf{R}^{X_1X_2}$, $i \in \mathbf{N}_0$, in analogy with (4.11b) we then have the following.

Theorem 4.2.2 *The sequence $\langle \vec{z}^{(i)}\rangle_{i\in\mathbf{N}_0}$ in $\mathbf{R}^{X_1X_2}$ defined by the algorithm (4.12) converges to the solution of (4.10) for any starting vector $\vec{z}^{(0)} \in \mathbf{R}^{X_1X_2}$.*

Proof: Collecting all the terms with iteration index $i+1$ on the left hand side and all the other terms on the right hand side, the iteration scheme (4.12) takes the form

$$(I+L)\vec{z}^{(i+1)} = (D+U)\vec{z}^{(i)} + \frac{\vec{\zeta}}{1+4\mu} \qquad (4.13)$$

where $I, L, D, U \in \mathbf{R}^{X_1X_2\times X_1X_2}$, L is strictly lower triangular, U is strictly upper triangular, I is the identity matrix, and D is diagonal with the diagonal entries

$$D_{k(x),k(x)} = \frac{4-\omega(x)}{1+4\mu}\mu \geq 0 \qquad x \in X_\zeta$$

It moreover follows that

$$I+L-D-U = \frac{A}{1+4\mu}$$

and hence by the symmetry of A that $U = -L^T$. Let $F \doteq I+L$ and $G \doteq D+U = D-L^T$. Then F and $F-G = A/(1+4\mu)$ are both nonsingular. Let λ be an eigenvalue of $F^{-1}G$. Then $\exists y \in \mathbf{C}^{X_1X_2}$ such that $F^{-1}Gy = y\lambda$ and $y^Hy = 1$, from which it follows that

$$(1+y^HLy)\lambda = y^HFy\lambda = y^HGy = y^HDy - y^HL^Ty$$

Denoting the real and imaginary parts of $y^H Ly$ by a and b respectively and letting $d \doteq y^H Dy$ we thus have that

$$(1 + a + ib)\lambda = d - a + ib \tag{4.14}$$

Since A is strictly positive definite, we note that

$$0 < y^H(I + L - D + L^T)y = 1 + 2a - d$$

whence $1 + a > d - a$. Since the positive semidefiniteness of D implies that $d \geq 0$, it is also true that $a + 1 > a - d$. Hence $1 + a > |d - a| \geq 0$. From (4.14) we then see that

$$|\lambda|^2 = \frac{(d-a)^2 + b^2}{(1+a)^2 + b^2} < 1$$

which proves that the spectral radius of $F^{-1}G$ is strictly less than 1. From (4.13) and theorem 4.2.1 it now follows that

$$\lim_{i \to \infty} \vec{z}^{(i)} = (F - G)^{-1} \frac{\vec{\zeta}}{1 + 4\mu} = A^{-1}\vec{\zeta}$$

■

Although the choice of starting vector $\vec{z}^{(0)}$ does not have any influence on whether and to which limit the iteration (4.12) converges, it does of course affect the number of iterations that are required in order to reach a certain convergence criterion. The closer $\vec{z}^{(0)}$ is to the solution $A^{-1}\vec{\zeta}$ of (4.10) the faster convergence can in general be expected. It is therefore desirable to choose some prior estimate of $A^{-1}\vec{\zeta}$ as starting vector. The first time the image function is estimated, the original image function vector $\vec{\zeta}$ is without much doubt the best such prior estimate that is available without further processing. In the first cycle of the steepest descent loop we therefore choose $\vec{z}^{(0)} \doteq \vec{\zeta}$. If the updates of the edges are sufficiently small the difference between the image functions estimated during successive steepest descent loop cycles can be expected to be fairly small as well. The estimated image function from the previous cycle is then likely to be a better prior estimate than the original image function. Beginning from the second cycle we accordingly choose $\vec{z}^{(0)}$ to be the solution of (4.10) from the previous cycle in the loop. This choice furthermore has the convenient consequence that $\vec{z}^{(0)}$ is already stored in the estimated image function array since the previous cycle. The iteration (4.12) can thus begin without prior loading of $\vec{z}^{(0)}$.

While the estimated image functions from successive steepest descent loop cycles most likely differ very little at most pixel sites, their values are almost completely unrelated

at the relatively few pixel sites that happen to be on different sides of the same edge during the two cycles, due to the edge update that takes place in between. With the choice of starting vector described above one should therefore expect the most significant updates of $z^{(i)}$, (as far as both magnitude and importance for the ultimate convergence are concerned,) to be concentrated to the relatively few pixel sites close to the edges. In order to assure adequate convergence of $\langle z^{(i)} \rangle_i$ at these pixel sites, the convergence criterion (of the iteration (4.12)) must be sensitive to discrepancies over small subregions of the pixel grid X_ζ. The termination condition for (4.12) is therefore of the form

$$\|\bar{z}^{(i+1)} - \bar{z}^{(i)}\|_\infty < \varepsilon_z$$

where $\varepsilon_z > 0$ is a constant threshold.

4.2.4 Cost Gradient Computation

Cost Gradient Components

After the image function has been estimated, as described above, the global edge detector computes (a slightly modified version of) the gradient of the total cost (4.1) with respect to the independent control vertices in the vertex list. The main purpose for doing this is to determine an appropriate edge update for reducing the total cost. If the (total) cost gradient components are computed according to (4.10), and the total cost functional satisfies the simplifying assumptions underlying the identity (2.26), then the sequence of such updates generated throughout the steepest descent procedure will finally yield an image segmentation at which the cost gradient vanishes. That is, the control vertex (optimality) conditions in section 2.4 will be attained. This is basically what we want. However, the control vertex conditions can in similarity with the other edge conditions in the same section be expected to interfere with the possibilities of detecting edge segments with free endpoints. From the perspective of the steepest descent update law the problem is that the update term associated with the (arc or polygon) length cost makes the edge segments "pull" on their endpoints in the tangential direction. At endpoints that are constrained to form a junction with some other endpoint(s), there are sufficiently many other "update forces" present to balance this "pull". At free endpoints, however, the only other possible update force is (more or less) orthogonal to the "pull" direction. In this case equilibrium can therefore in general not be achieved.

The difficulty considered above could of course be eliminated by simply excluding edge segments with free endpoints from our model. Rather than taking such a drastic measure, however, we have chosen to get around the problem in a heuristic manner by artificially inhibiting the "pull" on the free endpoints. For the total cost (4.1) this is achieved by modifying the formula (2.47b) for the cost gradient components associated with the junctions to

$$
\begin{aligned}
g_j \doteq & \\
\doteq & \sum_{n\in N_{j0}} \left[\frac{v_{n2} - v_{n3}}{\|v_{n2} - v_{n3}\|} \varpi \beta_n \right. \\
& \left. + R_\times \sum_{m=0}^{2} \sum_{r=0}^{m} \sum_{s=0}^{3} v_{n,m-r+s} \int_0^1 \blacktriangle \varrho_n(m - r + \sigma) b_r(\sigma) \dot{b}_s(\sigma)\, d\sigma \right] \\
& + \sum_{n\in N_{j1}} \left[\frac{v_{n,M_n} - v_{n,M_n-1}}{\|v_{n,M_n} - v_{n,M_n-1}\|} \varpi \tau_n \right. \\
& \left. + R_\times \sum_{m=M_n}^{M_n+2} \sum_{r=m-M_n+1}^{3} \sum_{s=0}^{3} v_{n,m-r+s} \int_0^1 \blacktriangle \varrho_n(m - r + \sigma) b_r(\sigma) \dot{b}_s(\sigma)\, d\sigma \right] \\
& j = 1, \ldots, J
\end{aligned}
\tag{4.15}
$$

where β_n and τ_n are the end condition variables of the nth spline. The terms associated with the polygon length cost are thereby removed from the updates of the free end vertices. The intermediate vertices are of course not directly affected by the free endpoint difficulties. The cost gradient components associated with these control vertices are therefore left as given by (2.47a), that is

$$
\begin{aligned}
g_{nm} \doteq & \\
\doteq & \left(\frac{v_{nm} - v_{n,m-1}}{\|v_{nm} - v_{n,m-1}\|} + \frac{v_{nm} - v_{n,m+1}}{\|v_{nm} - v_{n,m+1}\|} \right) \varpi \\
& + R_\times \sum_{r=0}^{3} \sum_{s=0}^{3} v_{n,(m-r) \bmod M_n + s} \int_0^1 \blacktriangle \varrho_n((m - r) \bmod M_n + \sigma) b_r(\sigma) \dot{b}_s(\sigma)\, d\sigma \\
& m = 1 + 2o_n, \ldots, M_n - o_n, \qquad n = 1, \ldots, N
\end{aligned}
\tag{4.16}
$$

Image Cost Density Difference Approximation

In order to compute the modified cost gradient components as given by (4.15) and (4.16), the image cost density difference functions $\blacktriangle\varrho_1, \ldots, \blacktriangle\varrho_N$ have to be expressed in

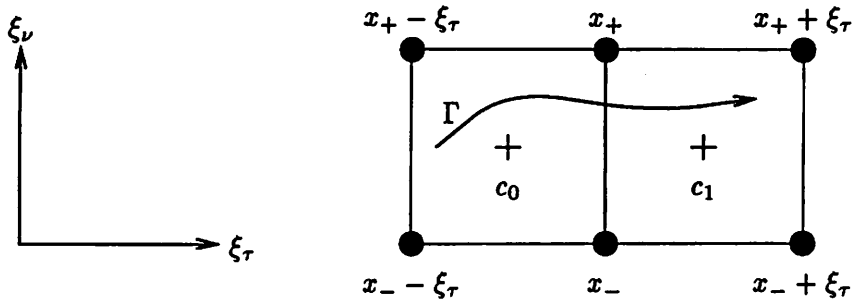


Figure 4.9: Notions for approximation of image cost density on each side of short directed curve Γ.

terms of some known quantities. From section 2.3 we know that

$$\blacktriangle\varrho_n(\sigma) = \varrho_{+,n}(\sigma) - \varrho_{-,n}(\sigma) \qquad \sigma \in \Sigma_n \qquad n = 1, \ldots, N$$

where $\varrho_{+,n}$ and $\varrho_{-,n}$ are the image cost densities evaluated on the left and right sides respectively of the (directed) edge segment $\gamma_n(\Sigma_n)$ at the point $\gamma_n(\sigma)$. The image cost density (4.2) is, however, given in terms in terms of the image functions ζ and z of which only samples on the pixel grid X_ζ are available. The functions $\varrho_{\pm,n}, \blacktriangle\varrho_n : \Sigma_n \to \mathbf{R}$, $n = 1, \ldots, N$, must thus all be approximated in terms of these samples.

Consider first a general short directed curve $\Gamma \subseteq \mathbf{R}^2$ that goes from one cell (centered at) c_0 to one of its four-connected neighbors (centered at) c_1 without entering any other cell in between. Define two orthogonal unit vectors in Ξ by

$$\begin{aligned} \xi_\tau &\doteq c_1 - c_0 \\ \xi_\nu &\doteq R_\times^T \xi_\tau \end{aligned}$$

where $R_\times$ as before is the 90° clockwise rotation matrix defined in (2.11). The six pixel sites at the corners of the two cells are then given by $x_\pm + p\xi_\tau$, $p = -1, 0, 1$, where

$$x_\pm \doteq \frac{c_0 + c_1}{2} \pm \frac{\xi_\nu}{2}$$

and oriented relative to Γ as shown in figure 4.9. The image cost density on the left and right sides of Γ can therefore be approximated by the constants ϱ_+ and ϱ_- respectively given by

$$\varrho_\pm \doteq [z(x_\pm) - \zeta(x_\pm)]^2 + \mu \left(\left[\frac{z_{\xi_\tau}(x_\pm) - z_{-\xi_\tau}(x_\pm)}{2} \right]^2 + [z_{\mp\xi_\nu}(x_\pm) - z(x_\pm)]^2 \right)$$

where z_ξ, $\xi \in \Xi$, is defined by (4.7). If Γ is part of an edge segment—the case of interest, the bond centered at $(c_0+c_1)/2$, which joins $x_\pm$ with $x_\pm \mp \xi_\nu$, is necessarily broken. Thence $z_{\mp\xi_\nu}(x_\pm) = z(x_\pm)$ and we have

$$\varrho_\pm = [z(x_\pm) - \zeta(x_\pm)]^2 + \frac{\mu}{4}[z_{\xi_\tau}(x_\pm) - z_{-\xi_\tau}(x_\pm)]^2$$

The discussion above strongly suggests that the image cost density differences $\blacktriangle\varrho_1, \ldots, \blacktriangle\varrho_N$ be approximated by piecewise constant functions. With the sample point sequences and cell sequences from the continuity set evaluation conveniently at hand we therefore choose the approximation

$$\blacktriangle\varrho_n(m+\sigma) \approx \blacktriangle\varrho_{nml} \doteq \varrho_{+nml} - \varrho_{-nml} \qquad \sigma \in]\sigma_l, \sigma_{l+1}[,$$
$$l = 0, \ldots, L_{nm} - 1, \qquad m = 0, \ldots, M_n - 1, \qquad n = 1, \ldots, N \tag{4.17}$$

where the constants $\varrho_{\pm nml}$ are defined by

$$\varrho_{\pm nml} \doteq [z(x_{\pm nml}) - \zeta(x_{\pm nml})]^2 + \frac{\mu}{4}[z_{\xi_{nml}}(x_{\pm nml}) - z_{-\xi_{nml}}(x_{\pm nml})]^2$$

$$x_{\pm nml} \doteq \frac{c_{nml} + c_{nm,l+1}}{2} \pm \frac{R_\times^T \xi_{nml}}{2}$$

$$\xi_{nml} \doteq c_{nm,l+1} - c_{nml}$$

Computational Procedure

With the image segmentation configuration represented as described earlier in this section, the terms of the sums in (4.15) and (4.16) are most efficiently accumulated spline by spline. We therefore return to (4.1) and (4.2) and rewrite the total cost variation (2.46) with respect to the control vertices as

$$\delta_v c_{N\zeta}(\gamma, z) = \sum_{n=1}^{N} \delta_n$$

where

$$\delta_n \doteq \varpi \delta_v \Pi(v_{n0}, \ldots, v_{n,M_n+2}) - \int_{\gamma_n(\Sigma_n)} \blacktriangle\varrho_n \delta\gamma_{n\nu}\, d\ell$$

obviously is the contribution from the nth spline. Expanding δ_n according to (2.28), (2.29) and (the second last line of) (2.30) we find that

$$
\begin{aligned}
\delta_n &= \\
&= \varpi \left[o_n \left(\frac{v_{n2} - v_{n3}}{\|v_{n2} - v_{n3}\|} \right)^T \delta v_{n2} \right. \\
&\quad + \sum_{m=1+2o_n}^{M_n - o_n} \left(\frac{v_{nm} - v_{n,m-1}}{\|v_{nm} - v_{n,m-1}\|} + \frac{v_{nm} - v_{n,m+1}}{\|v_{nm} - v_{n,m+1}\|} \right)^T \delta v_{nm} \\
&\quad \left. + o_n \left(\frac{v_{n,M_n} - v_{n,M_n-1}}{\|v_{n,M_n} - v_{n,M_n-1}\|} \right)^T \delta v_{n,M_n} \right] \\
&\quad - \sum_{m=0}^{M_n-1} \sum_{r=0}^{3} \sum_{s=0}^{3} \int_0^1 \blacktriangle \varrho_n (m+\sigma) b_r(\sigma) \dot{b}_s(\sigma)\, d\sigma\, v_{n,m+s}^T R_\times \delta v_{n,m+r}
\end{aligned}
$$

The modified contributions from the nth spline obtained after deleting the terms, if any, that are responsible for the "pull" on the free endpoints, and incorporating the approximation (4.17) is thus given by the linear combination

$$
\begin{aligned}
&\sum_{p=1}^{P_n} \alpha_{np}^T \delta v_{n,m_{np}} = \\
&\quad = \varpi \left[\beta_n o_n \left(\frac{v_{n2} - v_{n3}}{\|v_{n2} - v_{n3}\|} \right)^T \delta v_{n2} \right. \\
&\qquad + \sum_{m=1+2o_n}^{M_n - o_n} \left(\frac{v_{nm} - v_{n,m-1}}{\|v_{nm} - v_{n,m-1}\|} + \frac{v_{nm} - v_{n,m+1}}{\|v_{nm} - v_{n,m+1}\|} \right)^T \delta v_{nm} \\
&\qquad \left. + \tau_n o_n \left(\frac{v_{n,M_n} - v_{n,M_n-1}}{\|v_{n,M_n} - v_{n,M_n-1}\|} \right)^T \delta v_{n,M_n} \right] \\
&\qquad - \sum_{m=0}^{M_n-1} \sum_{r=0}^{3} \sum_{s=0}^{3} \sum_{l=0}^{L_{nm}-1} \blacktriangle \varrho_{nml} \int_{\sigma_{nml}}^{\sigma_{nm,l+1}} b_r(\sigma) \dot{b}_s(\sigma)\, d\sigma\, v_{n,m+s}^T R_\times \delta v_{n,m+r}
\end{aligned}
$$

where $\alpha_{n1}, \ldots, \alpha_{n,P_n} \in \mathbb{R}^2$ and $m_{n1}, \ldots, m_{n,P_n} \in \{0, \ldots, M_n + 2\}$.

Because of the interconnection constraints and the spline end conditions, the control vertex variations $\delta v_{n,m_{np}}$, $p = 1, \ldots, P_n$, $n = n, \ldots, N$, are of course not independent. The block components of the cost gradient can therefore not just be identified with the vectors α_{np}, $p = 1, \ldots, P_n$, $n = n, \ldots, N$. Instead each block component consists of the sum of all those vectors that multiply any of its associated control vertices. For the purpose of computing these sums each 2×1 block component of the cost gradient of equivalently each control vertex in the vertex list has a designated accumulator, which the cost gradient computation routine initializes to zero. Each such accumulator is actually contained in the record of its associated control vertex in the vertex list. The accumulator designated to

the block component associated with any particular control vertex of any given spline is thereby addressable from the spline list by the same mechanism as are the coordinates of that control vertex. For each $p = 1, \ldots, P_n$, $n = 1, \ldots, N$, the vector α_{np} is computed and added to the contents of the accumulator designated to the block component associated with $v_{n,m_{np}}$. When this procedure is completed, all the block components, as given by (4.15) and (4.16), are contained in their designated accumulators.

The computation of the vectors α_{np}, $p = 1, \ldots, P_n$, $n = n, \ldots, N$, normally necessitates access to the value of the integral $\int_{\sigma_0}^{\sigma_1} b_r(\sigma)\dot{b}(\sigma)\, d\sigma$ for all the 16 combinations of $r, s \in \{0, \ldots, 3\}$ and for a large number of values of σ_0 and σ_1. For higher speed the functions

$$a_{rs}(\sigma) \doteq \int_0^{\sigma} b_r(\varsigma)\dot{b}(\varsigma)\, d\varsigma \qquad r, s \in \{0, \ldots, 3\}$$

are therefore tabulated (alongside of the basis functions $b_0, \ldots, b_3$,) thereby allowing the integral above to be computed by means of the single subtraction $a_{rs}(\sigma_1) - a_{rs}(\sigma_0)$. Although defined as an indefinite integral, each such function a_{rs} is merely a known sixth order polynomial, and thus very simple to pre-evaluate.

4.3 Experimental Results

The global curve-represented edge detector described in section 4.2 was, as mentioned earlier, implemented for studying the performance of the steepest descent procedure. Our experiments with it had two major purposes. Most importantly we wanted to demonstrate that the steepest descent procedure does indeed adjust the edges so as to reduce the total cost, and that the edge adjustment moreover represents an improvement to human evaluation. Secondly we wanted to show how the estimated image, the detected edges and the convergence rate are affected by the edge cost and stabilizing cost coefficients λ and μ as well as by the number of control vertices used in the image segmentation configuration. In this section we present some of the results of these experiments.

4.3.1 Edge Adjustment

A First Example

For the first experiment to be presented we used the original image (of a personal computer) shown in figure 4.10. The initial edges obtained with the initial edge finder are

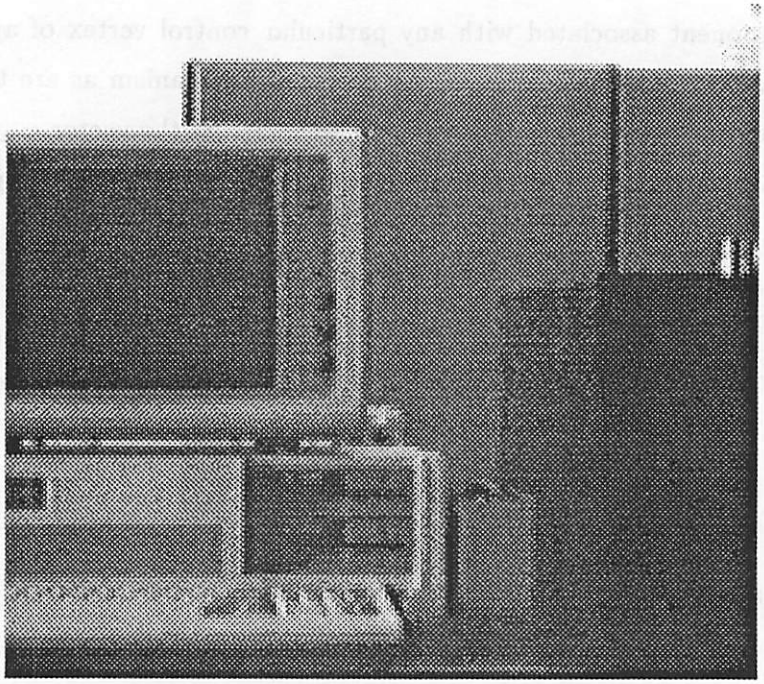


Figure 4.10: Original image used in first example.

shown in figure 4.11. As we see, the initial edges are located way off the corresponding contours of high gradient magnitude of the original image function. After 20 iterations—cycles of the steepest descent loop—the edges had been adjusted as shown (in black) in figure 4.12 (a). For comparison the initial edges are also superimposed (in grey). The edges obtained after 60 and 100 iterations are shown (in black) in figure 4.12 (b) and (c) respectively with the edges obtained 40 iterations earlier superimposed (in grey). As we see, most of the edge adjustment took place during the 20 first iterations. Indeed, the edges from the 60th iteration hardly show at all in figure 4.12 (c), indicating that the adjustment from the 60th to the 100th iteration is practically negligible.

In figure 4.13 the edges obtained after 100 iterations are superimposed on the original image. The match between these edges and the contours of high gradient magnitude of the original image function is, as we see, quite satisfactory.

The total edge adjustment during the 100 first iterations represents a significant improvement to a human observer. This is easily seen from figure 4.14, which shows the initial edges (in grey) superimposed on those obtained after 100 iterations (in black). The improvement can also be appreciated from figure 4.15, which shows the estimated image

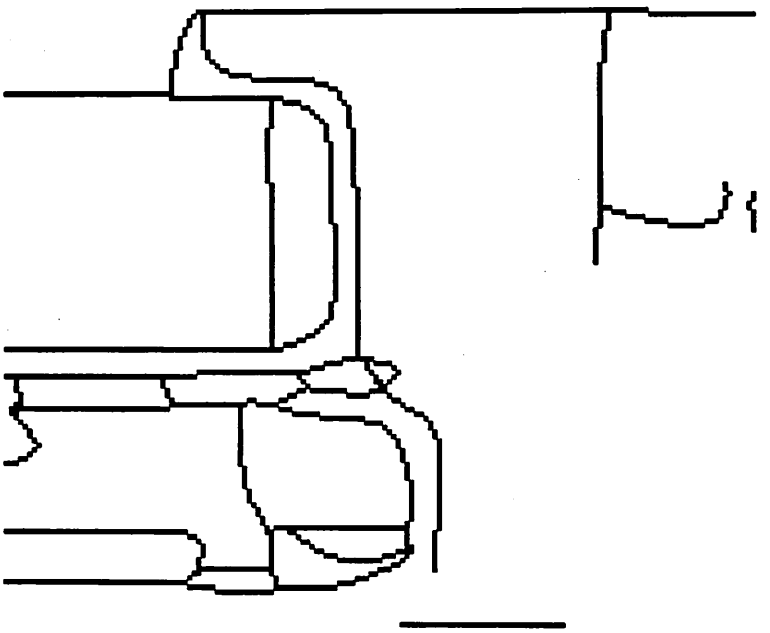


Figure 4.11: Initial edges obtained from original image in figure 4.10.

after 0 and 100 iterations.

Going back to figure 4.10 and 4.11 we note that some of the visible edges in the original image were not detected by the initial edge finder. The present implementation of the global edge detector does neither add to nor delete any of the initial edges, and is therefore essentially unable to change this condition, (except by stretching or shrinking the initial edges). If the missing edges were to be detected one would have had to adjust the parameters of the preliminary edge detector (described in section C.1) accordingly.

A Second Example

The original image in the foregoing example was quite simple. For our next example we used the much more detailed original image (of a painting) shown in figure 4.16. Figure 4.17 shows the initial edges obtained with the initial edge finder. The edges resulting after 180 iterations are shown alone in figure 4.18 and superimposed on the original image in figure 4.19. Once again the edges were adjusted so as to match the contours of high gradient magnitude of the original image function. The most apparent improvements are those involving the shapes of the little statue, the decanter and the top of the glass. From

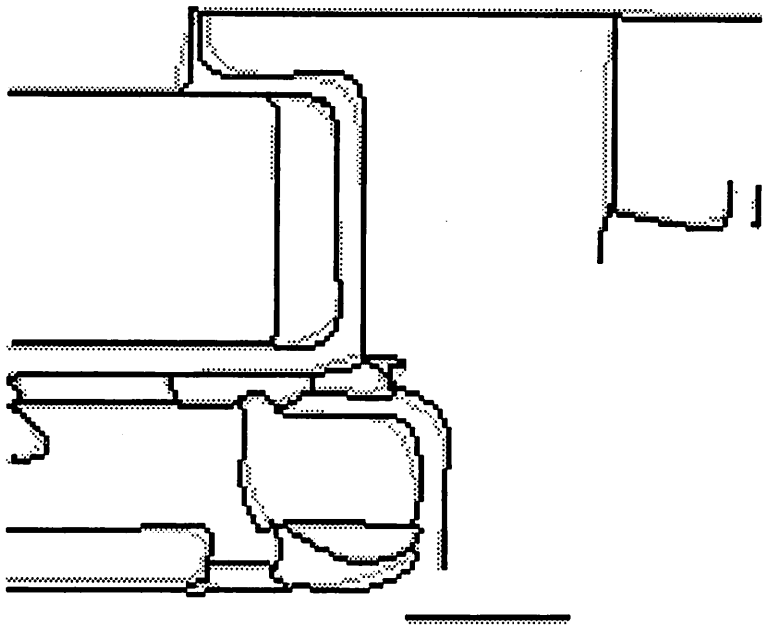


Fig. 4.12: (a)

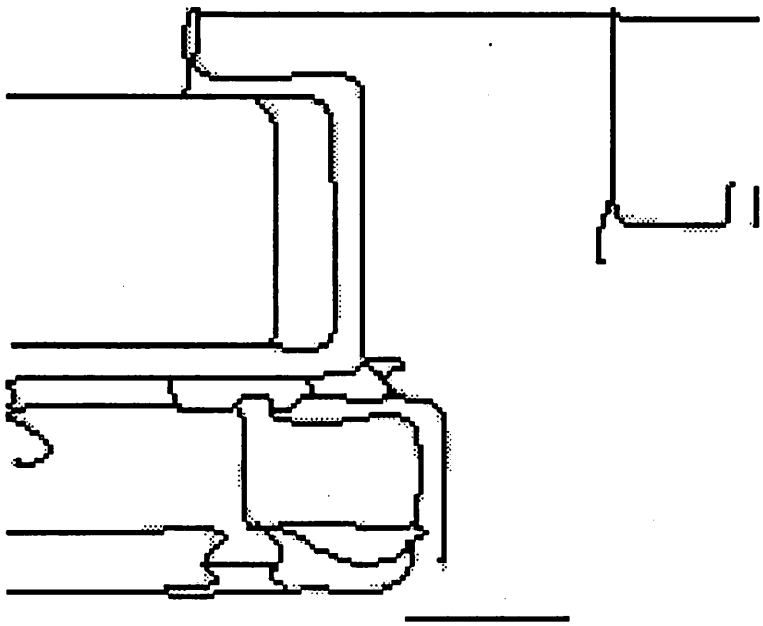


Fig. 4.12: (b)

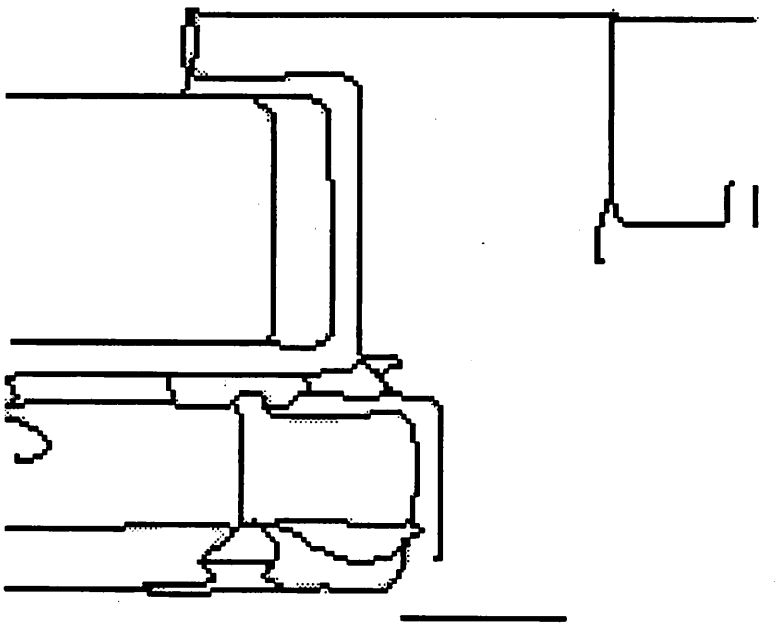


Fig. 4.12: (c)

Figure 4.12: Adjusted edges obtained from original image in figure 4.10 after: (a) 0 (grey) and 20 (black) iterations. (b) 20 (grey) and 60 (black) iterations. (c) 60 (grey) and 100 (black) iterations.

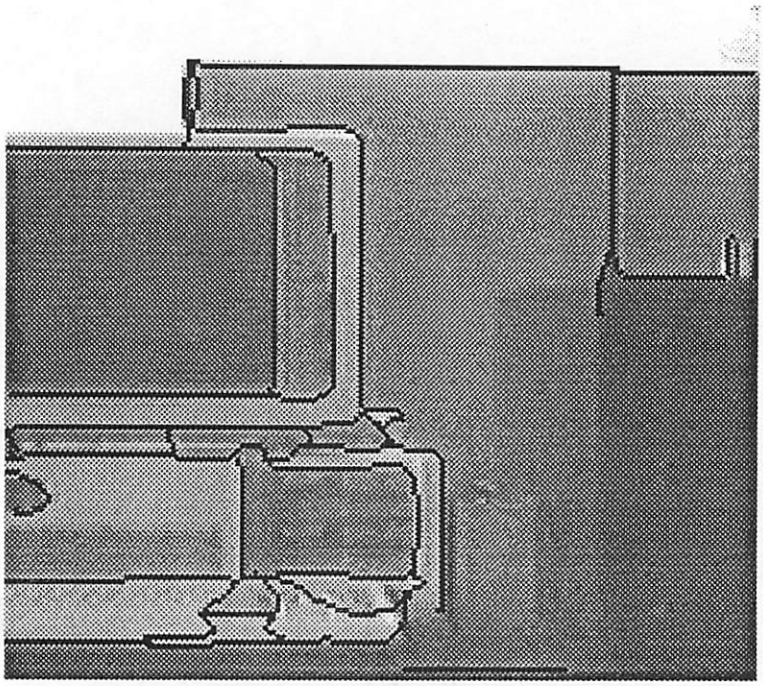


Figure 4.13: Adjusted edges after 100 iterations superimposed on original image.

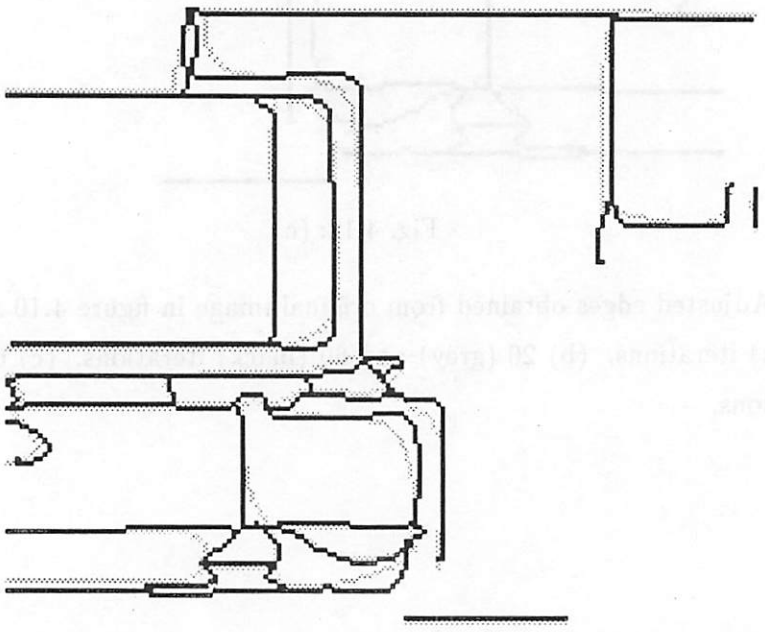
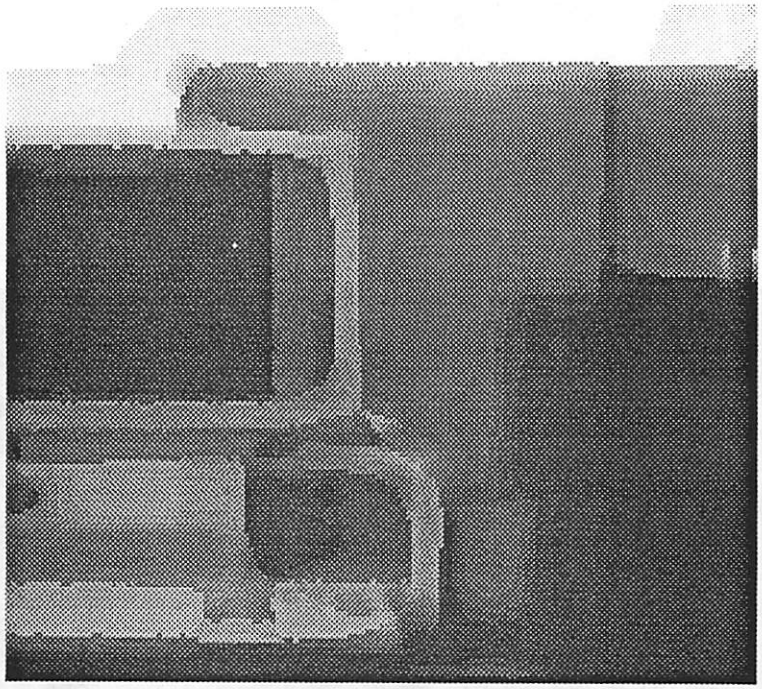
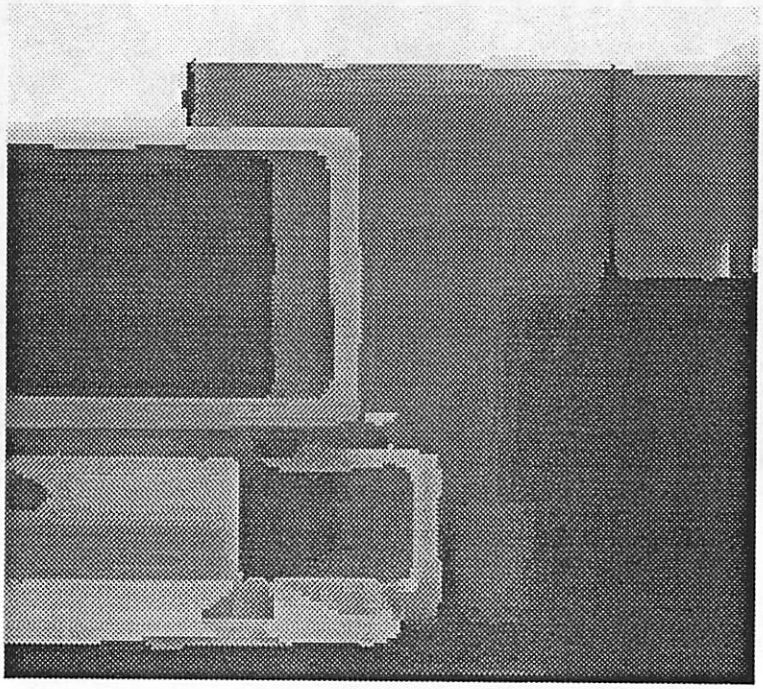


Figure 4.14: Adjusted edges after 100 iterations (black) and initial edges (grey) obtained from original image in figure 4.10.



(a)



(b)

Figure 4.15: Estimate of original image in figure 4.10 after: (a) 0 iterations. (b) 100 iterations.

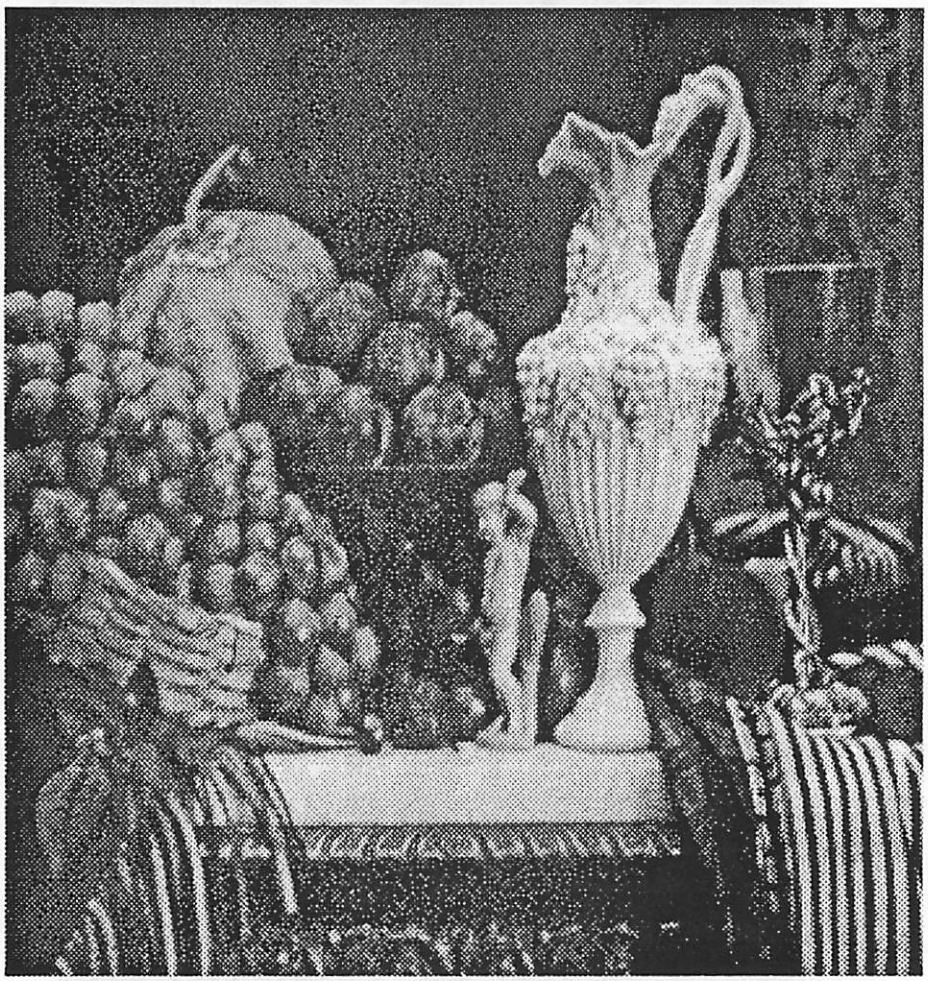


Figure 4.16: Original image used in second example.

Figure 4.17: Initial edges obtained from original image in figure 4.16.

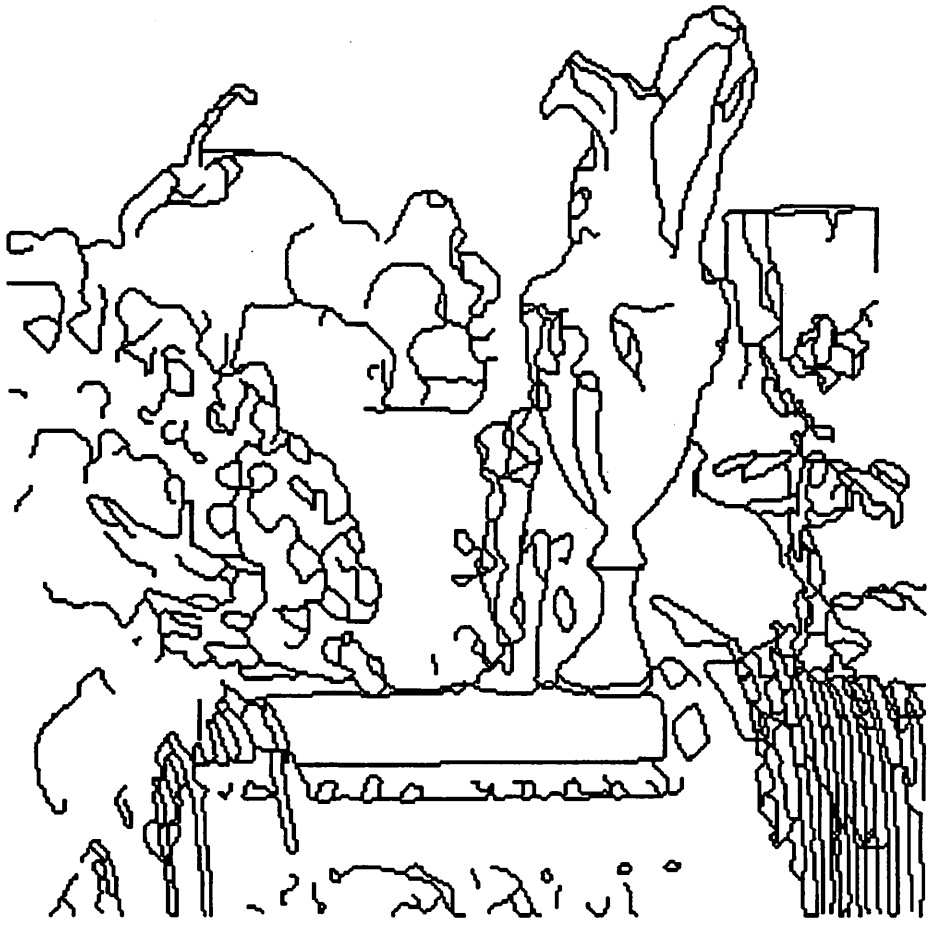


Figure 4.18: Adjusted edges obtained from original image in figure 4.16 after 180 iterations.

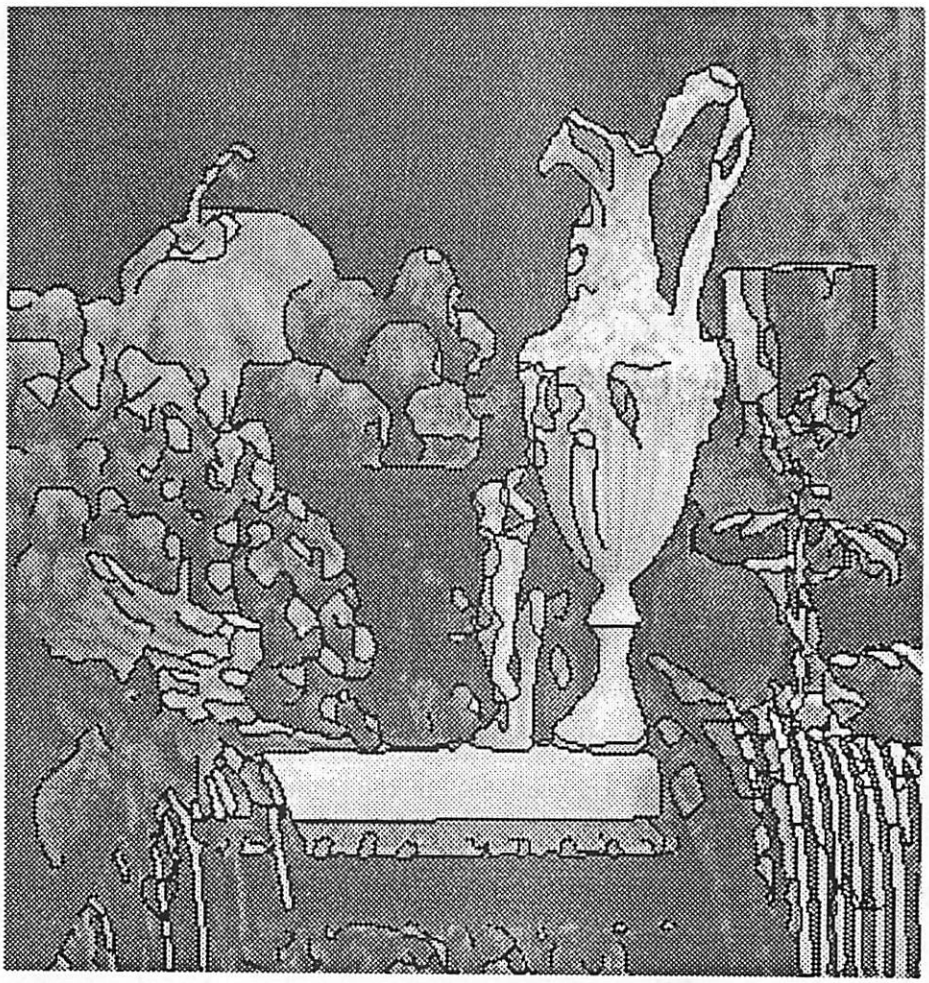


Figure 4.19: Adjusted edges after 180 iterations superimposed on original image.

the estimated images shown in figure 4.20 it is also clear that the edges associated with the stripes on the cloth in the lower right corner agree much better with the contours of high gradient magnitude of the original image function, after than before the edge adjustment.

Free Endpoints

In both the examples discussed above the initial edges possess (at least) a few free endpoints. Comparing these edges with those obtained after the edge adjustment we find neither essential growth nor essential shrinkage of the edges (in the tangential directions) at those free endpoints. Edge adjustment in the direction normal to the edge does, however, take place as much at the free endpoints as anywhere else. This phenomenon is probably easiest to notice on the edges in figure 4.17 and figure 4.18 that correspond to the right hand side of the glass in the right part of the image in figure 4.16. Altogether the observations, which are in full agreement with our prior expectations, indicate that our heuristic method for dealing with the free endpoint problem works as intended.

Cost Reduction

The total cost reduction from the edge adjustment generated by the steepest descent procedure amounted to 30% in both the above examples. The cost reduction resulting from the entire edge detection process, that is the reduction from the minimum image cost (of the estimated image function) in the complete absence of edges to the total cost after the steepest descent procedure adjustment of the initial edges, was of course greater. In the first example above this reduction was 55%. In the second example it was 49%.

In all our experiments the total cost decreased steadily during the beginning of the edge adjustment, that is the early cycles of the steepest descent loop. As expected, the steady decline in the cost then slowly diminished, whereupon the cost started to fluctuate up and down. In some of the experiments a pronounced trend of slow cost reduction was sustained many iterations after this fluctuation began. This pattern seems to be caused by slow convergence of some of the edges after the majority of the edges have already converged.

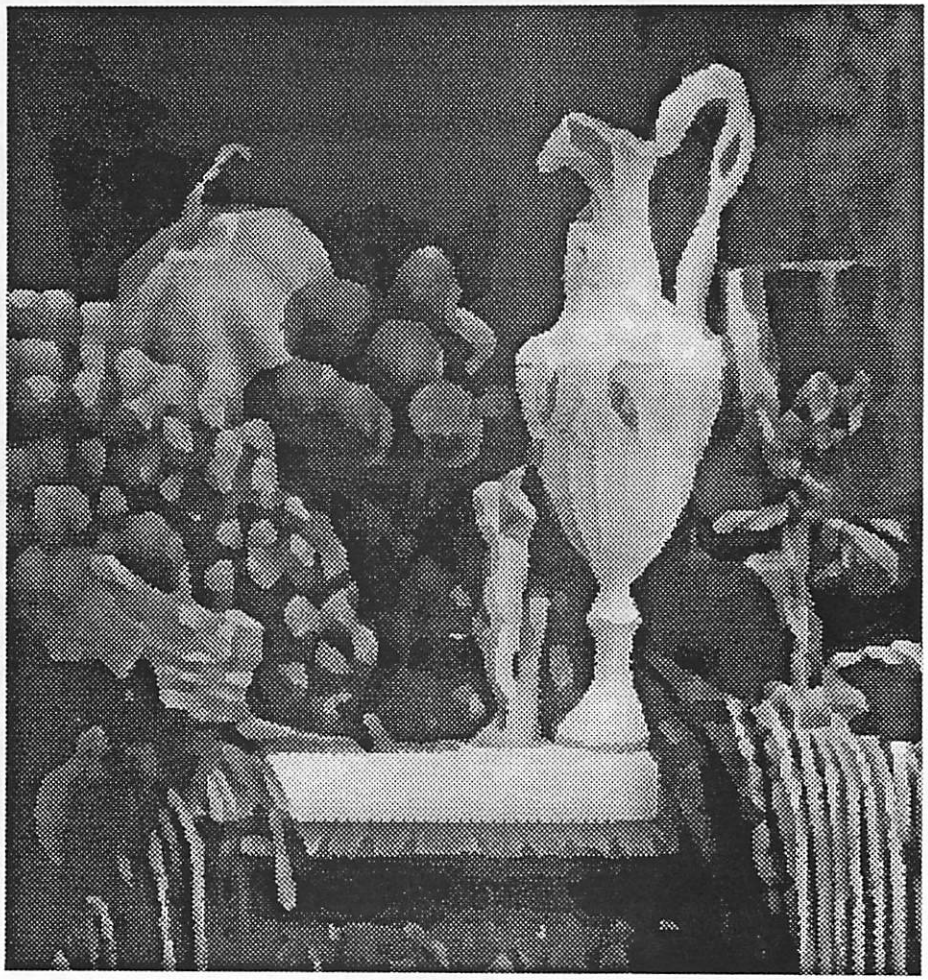


Fig. 4.20: (a)

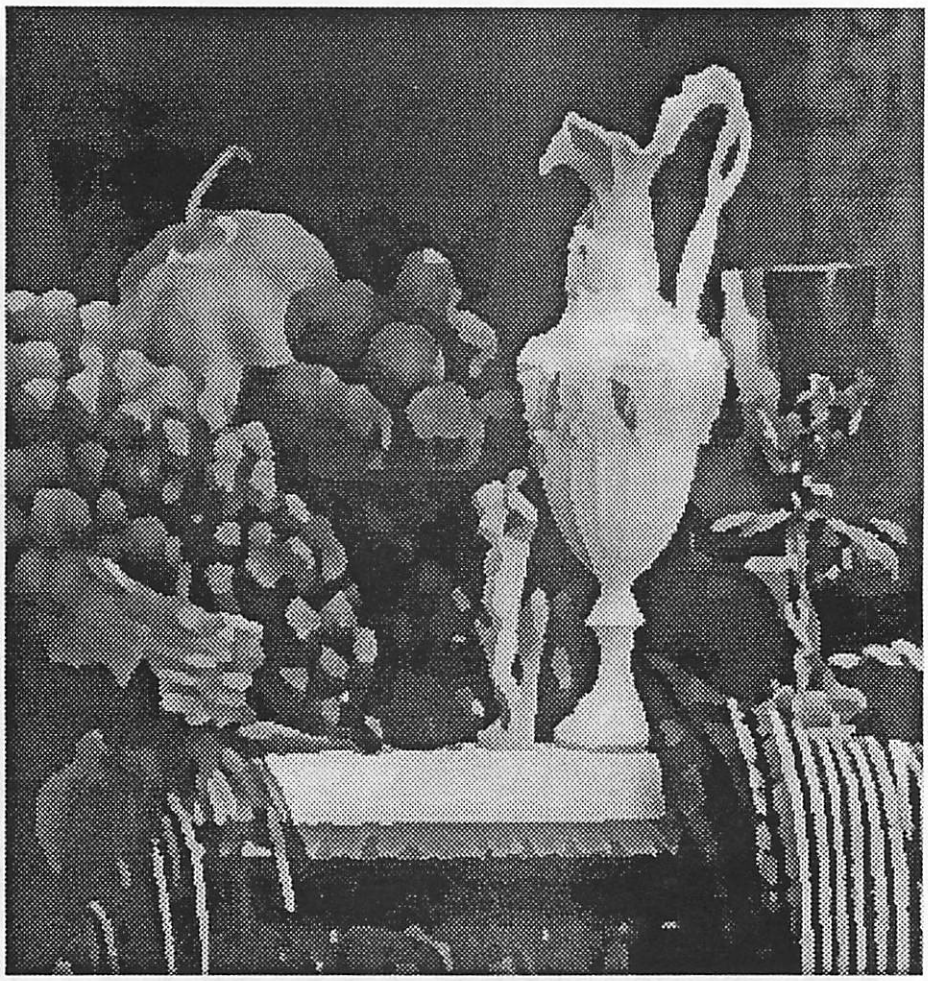


Fig. 4.20: (b)

Figure 4.20: Estimate of original image in figure 4.16 after: (a) 0 iterations. (b) 180 iterations.

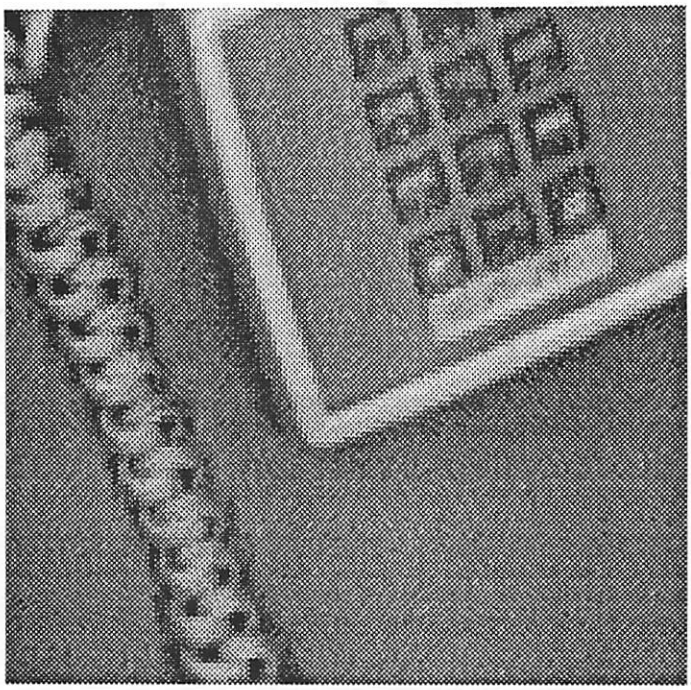


Figure 4.21: Original image used in edge cost coefficient experiment.

4.3.2 Parameter Dependence

The Edge Cost Coefficient

For our experiments regarding the *edge cost coefficient* λ we used the original image (of a telephone) shown in figure 4.21. This image was a relatively hard one for the initial edge finder to process. As a result the initial edges, which are shown in figure 4.22, are not as good as one would have hoped. However, the example illustrates quite well how the edge adjustment is affected by the choice of the parameter λ.

We first processed the initial edges with a quite low value of λ. As a consequence of the relatively mild penalty for the length of (the control polygons defining) the edges the edge adjustment proceeded for well more than 250 iterations without reaching convergence. The edges obtained after 100 and 250 iterations are shown in figure 4.23 (a) and (b) respectively. As we see, the edges were adjusted considerably so as to make up for the inadequacies of the initial edge finder output. Most notable are the added wiggles along the receiver cord boundary and the addition of the edges outlining the boarder and the side of the phone body. We also note some spurious edges resulting from occurrences of one spline

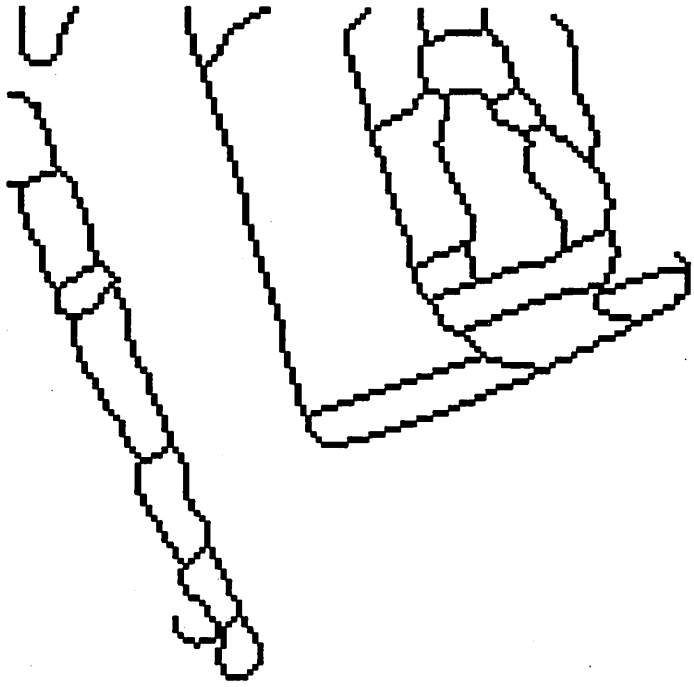
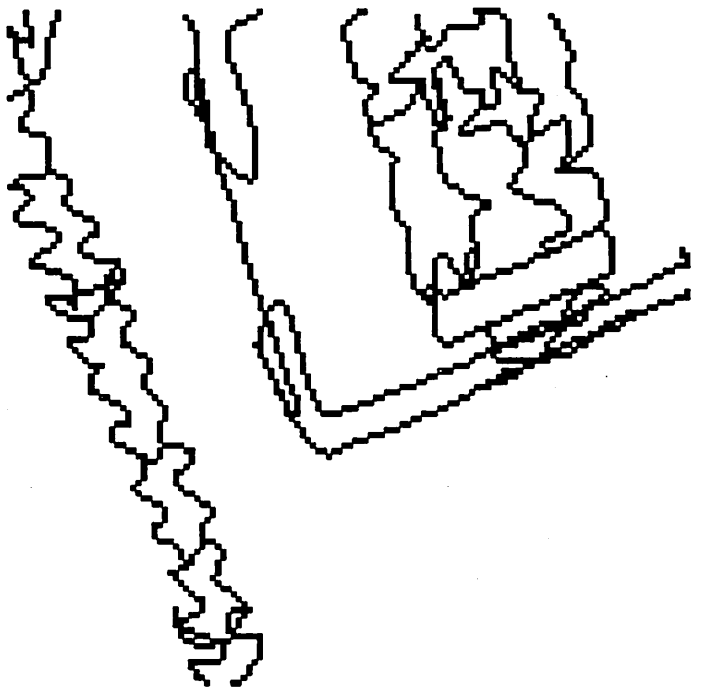


Figure 4.22: Initial edges obtained from original image in figure 4.21.

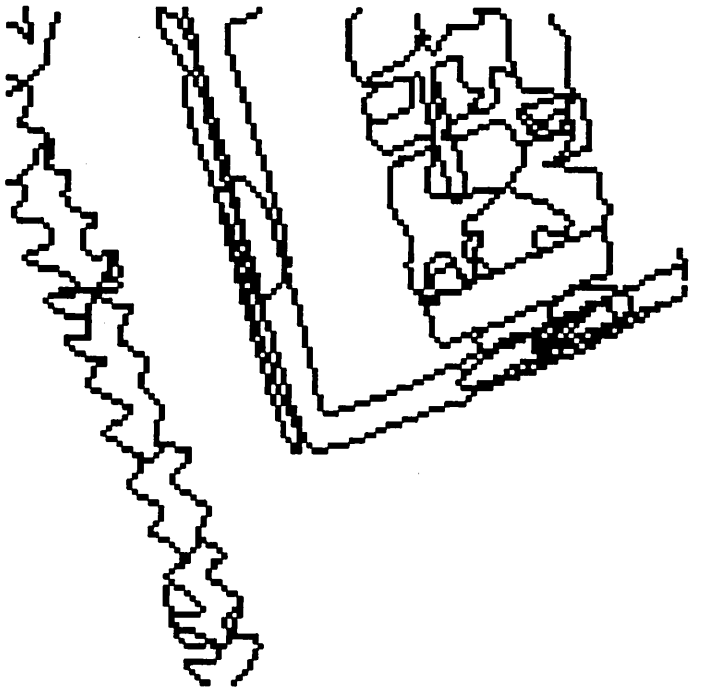
curve stretching so as to represent more than one distinct segments of the contours of high gradient magnitude of the original image function.

For our second processing of the initial edges in figure 4.22 we increased the value of λ by a factor 50. In this case the edges that converged after about 50 iterations. The resulting edges are shown in figure 4.24. As expected the high edge cost made the edges less inclined to stretch. Consequently the adjusted edges are less wiggly and there are no spurious edges present besides the few produced by the initial edge finder. The shape of the edges outlining the receiver cord and the key pad are moreover less accurate, and the boarder on the left rim of the phone body does not appear.

In summary a higher value of λ seems to lead to faster convergence and to prevent the appearance of spurious edges. The price paid for these advantages is a less accurate representation of wiggly edge shapes and a weaker tendency for the edge adjustment to make up for those edges that the initial edge finder did not detect.



(a)



(b)

Figure 4.23: Adjusted edges obtained with $\lambda = 65$ from original image in figure 4.21 after: (a) 100 iterations. (b) 250 iterations.

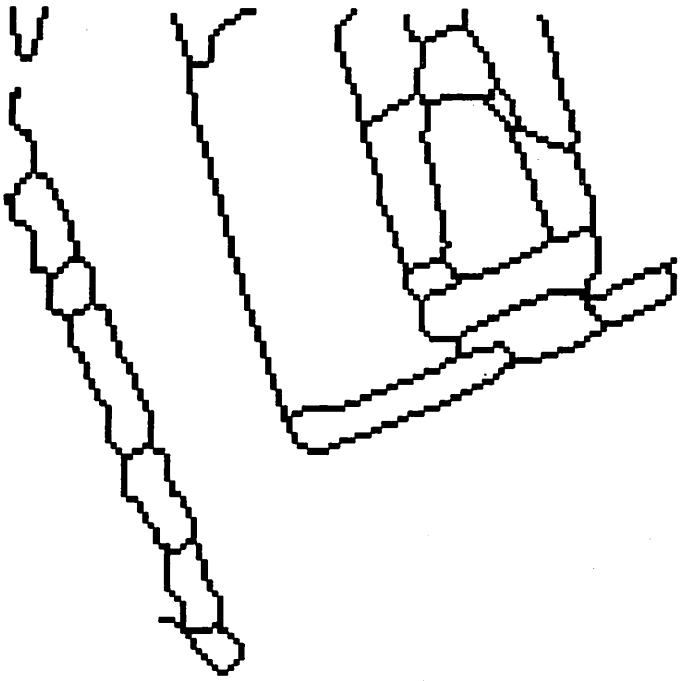


Figure 4.24: Adjusted edges obtained from original image in figure 4.21 with $\lambda = 3250$ after 50 iterations.

The Stabilizing Cost Coefficient

Our next example illustrates the influence of the *stabilizing* (or *nonsmoothness*) *cost coefficient* μ. The original image that was used for this experiment is shown in figure 4.25. (It shows a detail of the surface of a chip.) The edges produced by the initial edge finder are shown in figure 4.26.

We first processed these edges with the extremely low stabilizing cost coefficient $\mu = 0.5$. The edges had then converged already after 50 iterations. The resulting edges are shown alone in figure 4.27 (a) and superimposed on the original image in figure 4.27 (b). The adjusted edges are, as we see, fairly accurate. However, the edge outlining each of the dark circular regions inside the brighter strips is broken, or erroneously brought into contact with the edge corresponding to the boundary of the surrounding strip.

The estimated images obtained after 0 and 50 iterations are shown in figure 4.28 (a) and (b) respectively. Because of the low value of μ they are both very similar to the original image and hence to each other, despite the fact that the edges obtained after 50 iterations exhibit significant differences from the initial edges.

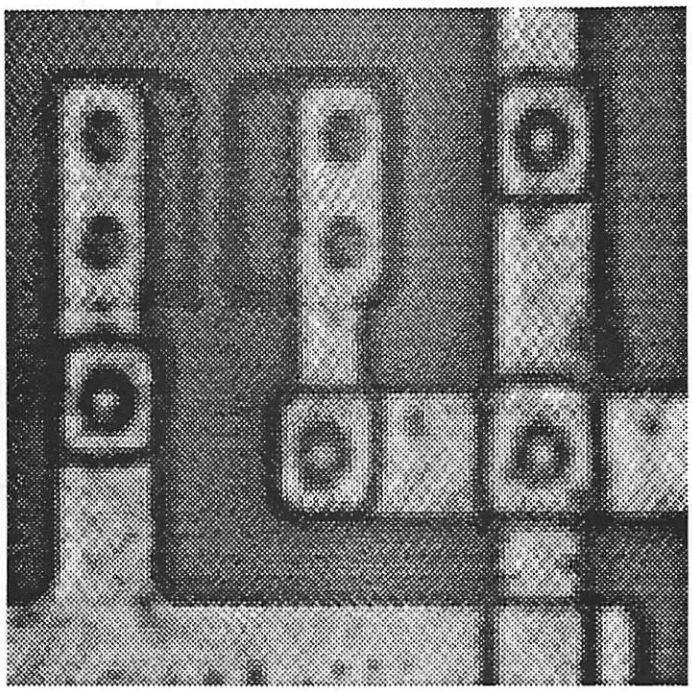


Figure 4.25: Original image used in stabilizing cost coefficient experiment.

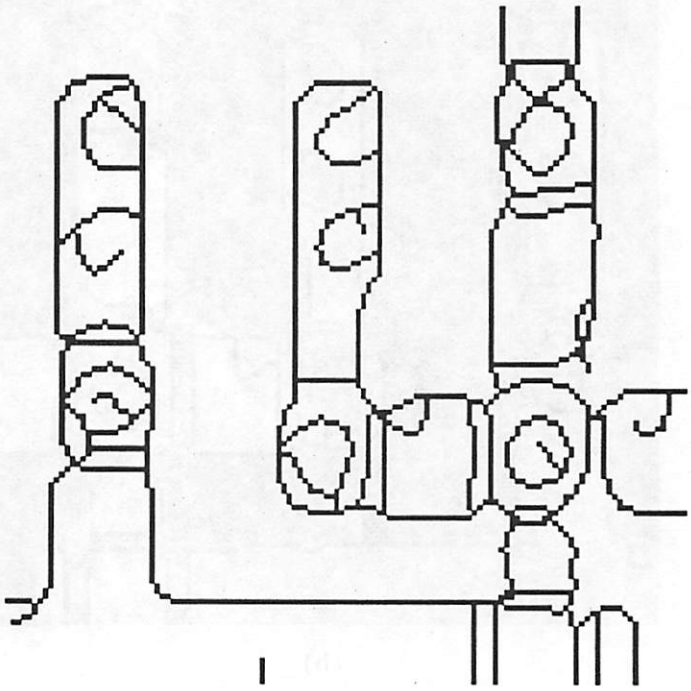
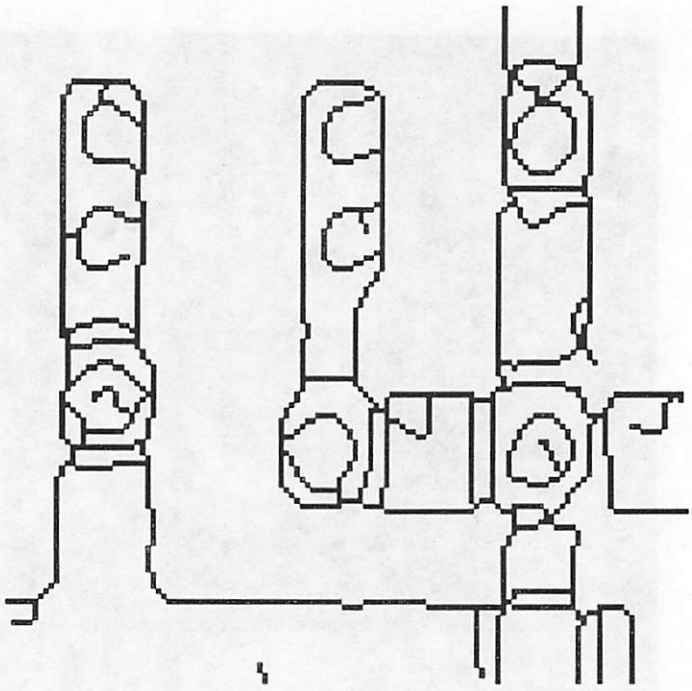
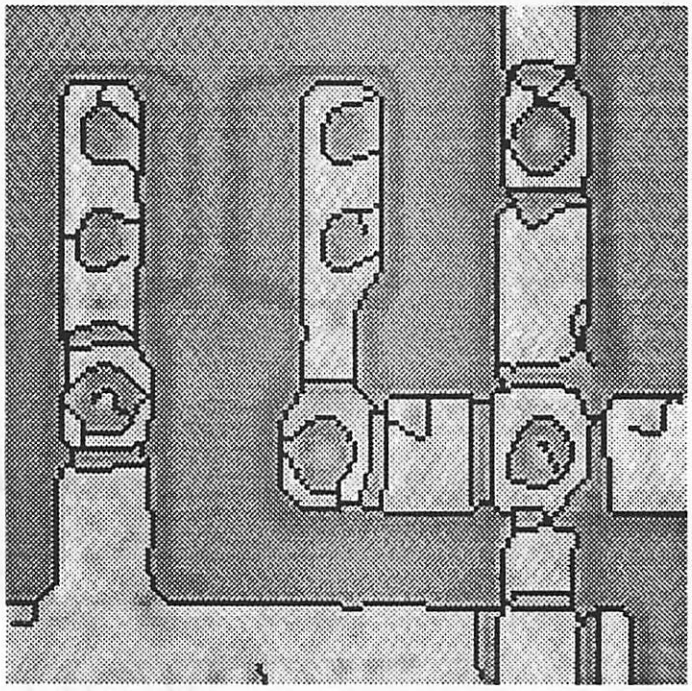


Figure 4.26: Initial edges obtained from original image in figure 4.25.

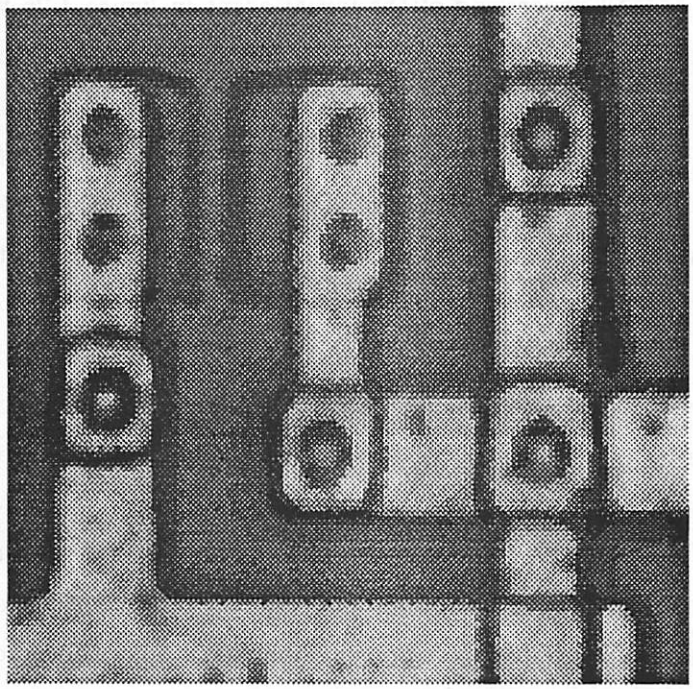


(a)

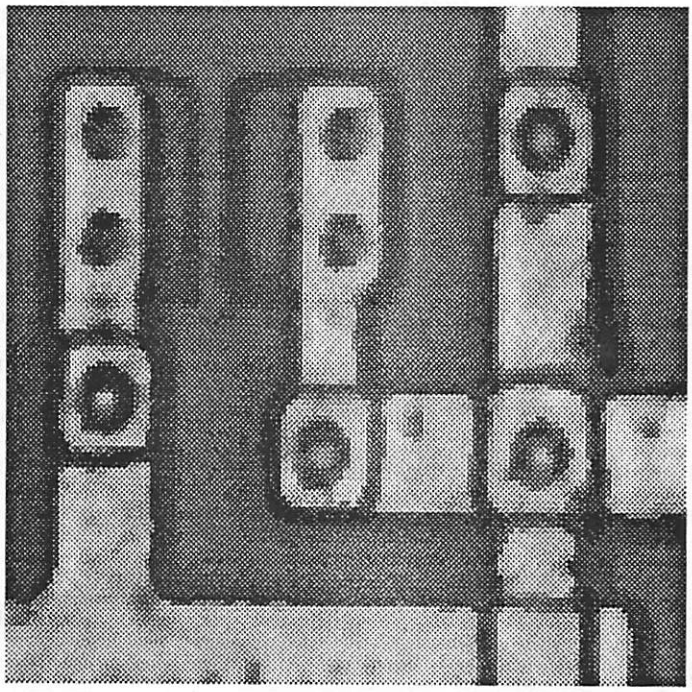


(b)

Figure 4.27: Adjusted edges obtained with $\mu = 0.5$ from original image in figure 4.25 after 50 iterations.



(a)



(b)

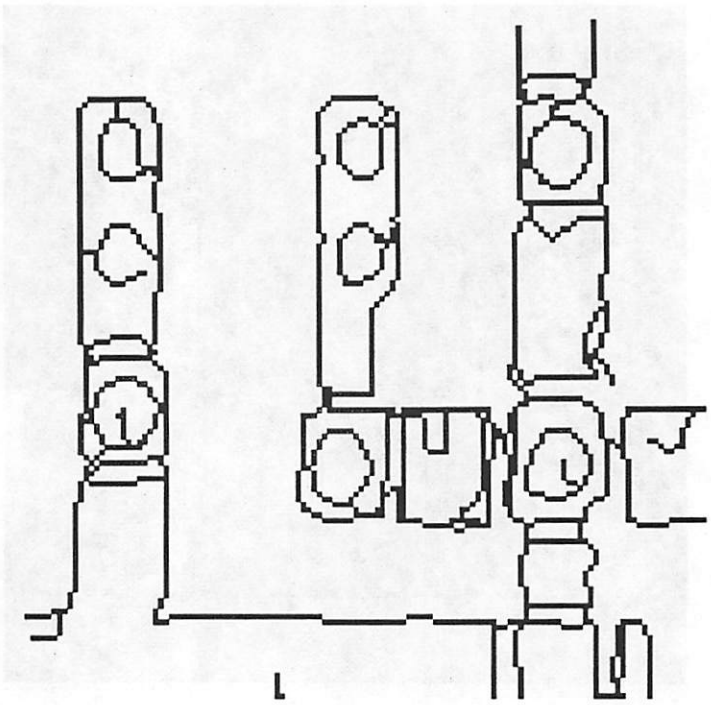
Figure 4.28: Estimate obtained with $\mu = 0.5$ of original image in figure 4.25 after: (a) 0 iterations. (b) 50 iterations.

For the second processing of the initial edges in figure 4.26 we used the 20 times greater stabilizing cost coefficient $\mu = 10$. In this case it took about 350 iterations for the edges to converge. The resulting edges are shown alone and superimposed on the original image in figure 4.29 (a) and (b) respectively. Comparing these pictures with those in figure 4.27 we note that the increase of the parameter μ improved the accuracy of the edges. In particular, all but one of the dark circular regions in the interior of the brighter strips were now correctly detected as such.

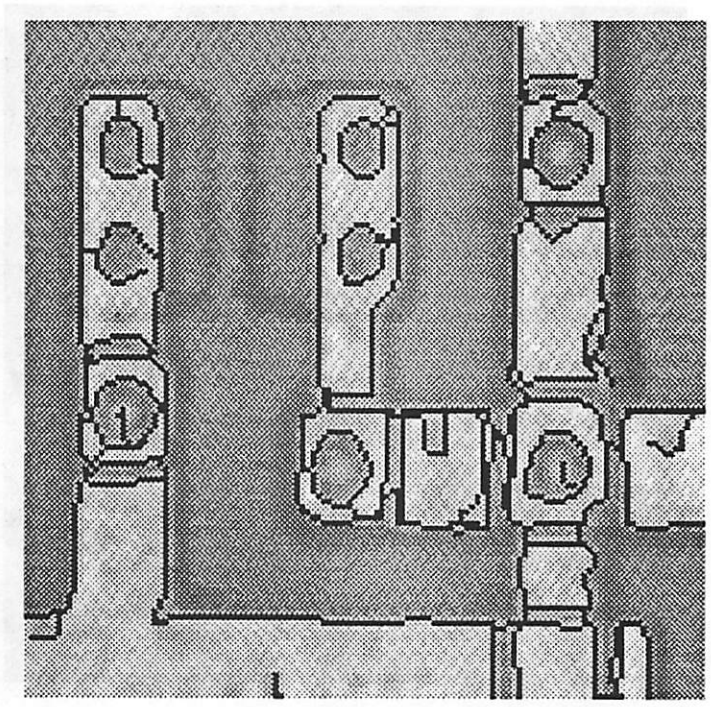
Figure 4.30 (a) and (b) show the estimated images obtained after 0 and 350 iterations respectively. With the higher value of μ the image estimation results in severe smoothing of the original image. Prior to the edge adjustment a substantial part of this smoothing takes place across the contours of high gradient magnitude of the original image function. As a result the estimated image in figure 4.30 (a) is relatively blurry. After the edge adjustment, on the other hand, the edges are sufficiently well lined up with the high gradient magnitude contours to prevent almost all smoothing across these contours. Thence, as figure 4.30 (b) well illustrates, the estimated image function is close to being piecewise constant.

In summary a high value of the parameter μ seems to yield more accurate edges and estimated image functions that are closer to being piecewise constant than does a low value of μ. The price one pays for the higher accuracy of the edges is a longer time to convergence. There are two contributing factors to the slower convergence. First of all, with a significantly higher value of μ the smoothing resulting from the image function estimation is, as we just have seen, much more severe. The numerical solution of the system (4.10) therefore requires many more iterations of (4.12) (inside the image function estimation routine), whence each iteration of the steepest descent procedure takes a substantially longer time. Secondly, for reasons soon to be discussed a high value of μ in general stimulates more edge adjustment than a low value of μ. Convergence of the edges can therefore in general be expected to require more iterations (of the steepest descent procedure).

Since multiplication of the total cost with a strictly positive constant has no influence on the steepest descent procedure and thus neither on the edge adjustment, an increase of the stabilizing cost coefficient μ is equivalent to a decreased emphasis on the edge and deviation costs. It should therefore not be surprising that an increase of μ in similarity with a decrease of λ, promotes the tendency of the edges to adjust to the contours of high gradient magnitude of the original image function. However, since an increase of μ unlike

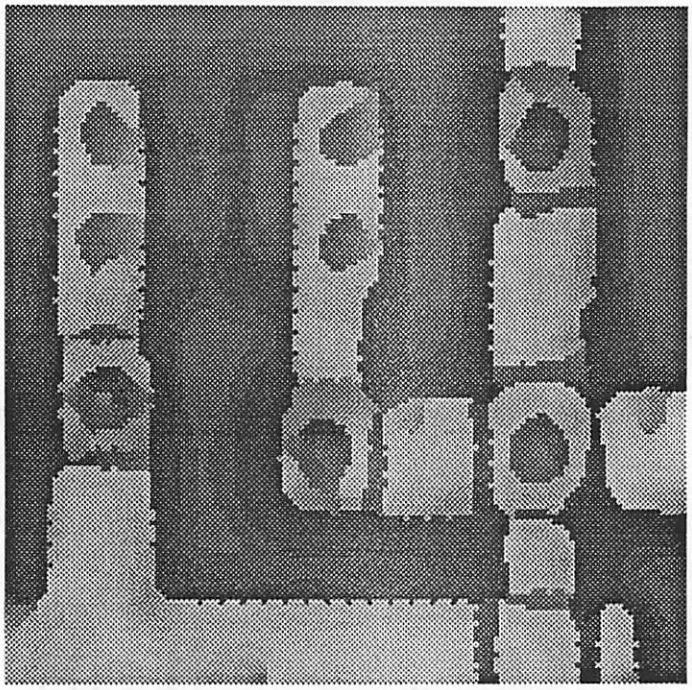


(a)

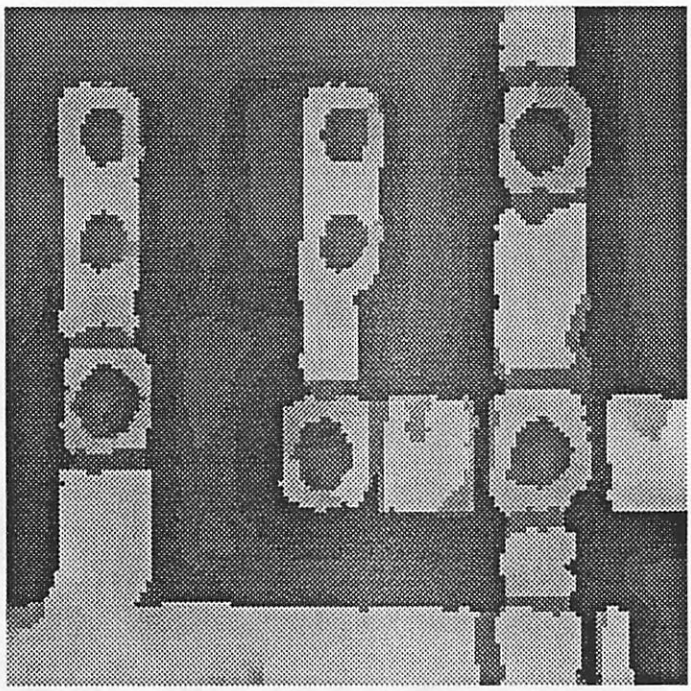


(b)

Figure 4.29: Adjusted edges obtained with $\mu = 10$ from original image in figure 4.25 after 350 iterations.



(a)



(b)

Figure 4.30: Estimate obtained with $\mu = 10$ of original image in figure 4.25 after: (a) 0 iterations. (b) 350 iterations.

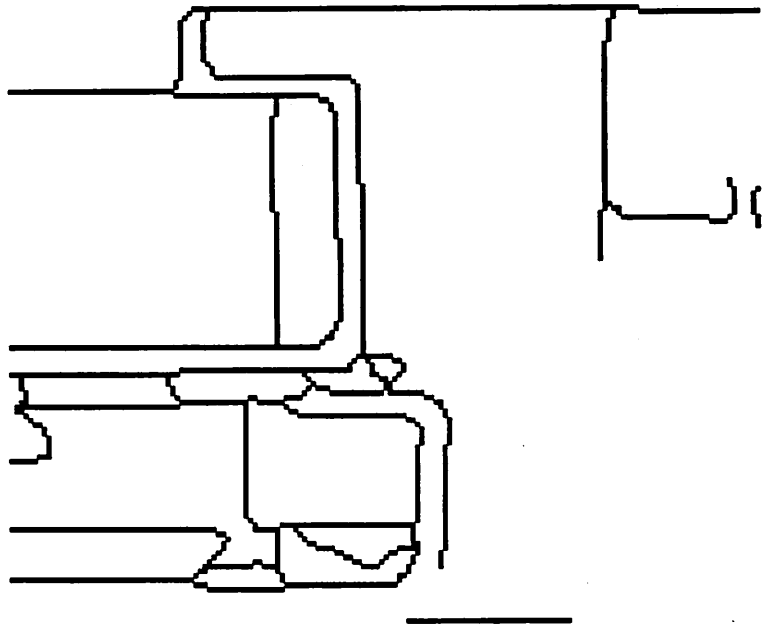


Figure 4.31: Initial edges obtained with $i_m = 4$ from original image in figure 4.10.

a decrease of λ besides de-emphasizing the edge cost also de-emphasizes the deviation cost, the edges are less likely to get attracted to small isolated such contours if μ is increased than if λ is decreased an "equivalent" amount. In other words, an increase of μ selectively promotes adjustment of the edges towards the "more significant" high gradient magnitude contours.

Control Vertex Density

In order to illustrate how the number of control vertices in the image segmentation configuration affects the detected edges we reprocessed the original image in figure 4.10 with a higher control vertex density than before. For our first example we used the maximum sampling interval $i_m = 16$. In other words, there was roughly one control vertex per sequence of 16 (consecutive) preliminary edge pixels. (See section C.3 for a precise definition of i_m.) This time around we chose $i_m = 4$, resulting in roughly four times as many control vertices for each edge segment as before. In this case the initial edge finder produced the edges shown in figure 4.31. These initial edges are, as we see, less rounded and therefore more accurate than those shown in figure 4.11 from our first example. The edges obtained after 80

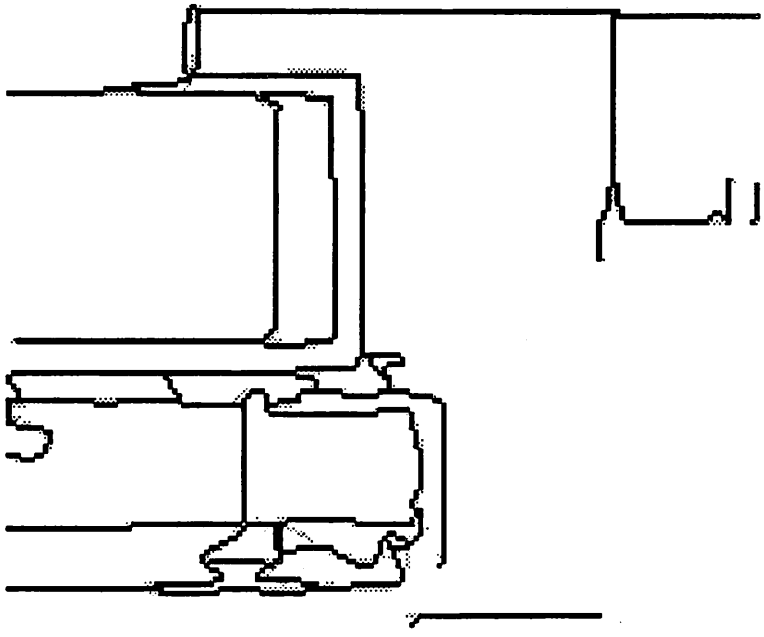


Figure 4.32: Adjusted edges obtained from original image in figure 4.10 with $i_m = 4$ after 80 iterations (black) and with $i_m = 16$ after 100 iterations (grey).

iterations are shown (in black) in figure 4.32. For comparison the adjusted edges obtained in our first example after 100 iterations are superimposed (in grey). In spite of the smaller number of iterations, (80 vs. 100,) the adjusted edges are also less rounded and more accurate than those obtained in our first example with the lower control vertex density.

In general a higher control vertex density—lower value of i_m—seems to yield more accurate edges as well as faster and more well-behaved convergence. The price one pays for these advantages is a less compact parametrization of the edges and a stronger tendency for the edges to take on quite irregular shapes and thereby also to get attracted by less significant contours of high gradient magnitude of the original image function.

Chapter 5

Biased Anisotropic Diffusion

In this chapter we present and analyze a second global edge detection approach based on variational regularization. While the new paradigm was originally intended as a general improvement on variational edge detection, the resulting algorithm can also be viewed as a new anisotropic diffusion method. We thereby unify these two, from the original outlook, quite different methods. This puts anisotropic diffusion, as a method in early vision, on more solid grounds; it is just as well-founded as the well-accepted standard regularization techniques. The algorithm to be presented moreover has a number of attractive properties, which makes it very competitive with other existing global edge detection methods.

5.1 Introduction

Of all the global edge detection approaches that we have encountered so far, all but one—the anisotropic diffusion method—are based on some kind of regularization. Regularization can, as we saw in chapter 1, be achieved in different ways. In probabilistic regularization [25, 26] the problem is reformulated as Bayesian estimation. In variational regularization [20, 21, 40, 41, 16, 22, 23, 24, 42], (of which the approach presented in chapter 2–4 obviously is an example,) it is posed as a cost (or energy) functional minimization problem, leading to a variational principle. In spite of the different outlooks of these approaches they essentially end up with the same mathematical and computational problem; given an *original image function* $\zeta : B \to \mathbf{R}$, defined on some open bounded connected *image domain* $B \subseteq \mathbf{R}^2$, minimize a cost functional $C_\zeta(w, z)$, where w is some function representing the edges, and $z : B \to \mathbf{R}$ is the *estimated* (or reconstructed) *image function.* In

each case the total cost can furthermore be divided into three components according to

$$C_\zeta(w,z) = \mathcal{E}(w) + \mathcal{D}(z,\zeta) + \mathcal{S}(w,z)$$

where the *edge cost* $\mathcal{E}$ measures the extent of the edges, the *deviation cost* $\mathcal{D}$ measures the discrepancy between the estimated and the original image functions, and the *stabilizing cost* measures the nonsmoothness or the a priori "unlikeliness" of the estimated image function.

The "edge function" w can be defined in a variety of ways. It might for example be an image segmentation or a vector of control vertices, such as those considered in chapter 2–4. In this chapter, however, it is, as most frequently in the literature, simply a function of the form $w : B \to \mathbf{R}$, which to each point in the image domain assigns a "measure of continuity" or a "discontinuity type".

Given a specific edge function w it is generally the case, that there exists a unique optimal estimated image function $\check{z}_w$, which can be found by solving a linear partial differential equation. While most of the regularization approaches do take advantage of this condition, none of them is capable of solving for the optimal edges in a similar way. The optimality conditions for the edges do either not exist, or else they consist of unsolvable equations. For the minimization of $C_\zeta(w,z)$ with respect to w all of the regularization approaches referred to above therefore resort to some kind of stochastic or deterministic search method such as the "Metropolis algorithm" or "steepest descent". Because of the tremendous size of the solution space any such search method is by itself quite expensive. In addition the general nonconvexity of the cost function causes any converging search algorithm to get stuck at local minima. The common response to this unfortunate situation has been to solve whole sequences of minimization problems, as a mechanism for "gravitating" towards a good local (hopefully a global) minimum. The GNC-algorithm introduced in [23, 24] and simulated annealing [25] are both examples thereof. As a consequence every global edge detection method up to date except the anisotropic diffusion methods involves some form of repeated iterative minimization process, and because of the high computational cost that this implies, the optimality of the solution is often compromised.

For anisotropic diffusion—the only global edge detection method that does not require the repeated iterations associated with the regularization based methods—the concerns are naturally of a different character. This method, as we recall from section 1.5, does not seek an optimal solution of any kind. Instead it operates by repeatedly filtering the image function with a smoothing kernel of small support, thereby producing a sequence

of *diffused image functions* of successively lower resolution. At some stage in the iterated filtering process remarkably impressive edges can be extracted by postprocessing the diffused image function with a rudimentary local edge detector. In the limit, however, all edges disappear, and the diffused image function converges to a constant. Needless to say, any solution of interest therefore has to be selected from the sequence of diffused image functions way before convergence.

The selection itself has so far been a matter of manual inspection. If automation is necessary, one can of course, in the absence of more sophisticated rules, simply prespecify a number of filter iterations. A more serious problem due to the necessity to select a solution prior to convergence, may arise in an analog circuit implementation where the diffusion process must be latched or halted in order to retrieve the diffused image function of interest.

In this chapter we show how the variational regularization approach by Terzopoulos [21, 40] can be modified so that the calculus of variations yields useful optimality conditions, not only for the estimated image function, but for the edges as well. The result is a global edge detection algorithm, which does not suffer from the high computational costs of most, if not all, of the other regularization-based such methods.

As it turns out, the new algorithm can also be viewed as a (new) *biased anisotropic diffusion* method. (The term "biased" will be explained in section 5.4.) This unification of the apparently quite different regularization and diffusion approaches is in itself very interesting. It also shows that it is completely fair to think of anisotropic diffusion as a global edge detection method. (The doubts, which were legitimately raised in section 1.5, are indeed reduced to whether a given anisotropic diffusion algorithm is truly global, or just approximates a global method.) Finally the unification brings the anisotropic diffusion approach an appealing sense of optimality. Anisotropic diffusion is thus a method for solving a well-defined mathematical problem, not just an image processing technique, by which one image can be transformed into another more pleasing looking one. With this face-lift of the foundations of the anisotropic diffusion method its extraordinary performance is no longer so surprising.

Even more exciting than the unification just discussed, is the fact that the new algorithm shares the better properties of both the regularization based methods and the anisotropic diffusion method. Indeed:

1. It only requires the solution of a *single* boundary value problem on the *entire* image domain—almost always a very simple region.

2. It converges to a solution of interest.

The first of these properties implies a number of advantages over other existing regularization methods. In particular:

(i) No explicit search method is necessary.

(ii) No sequence of minimization problems has to be solved.

The computational cost is therefore relatively low. The second property represents a couple of advantages over the existing diffusion methods:

(i) It removes the problem of manual selection of, which one in the sequence of diffused image functions, to be postprocessed with the local edge detector.

(ii) It is superior for circuit implementations.

The rest of this chapter is organized as follows: In the next section we review Terzopoulos' edge representation in terms of continuity control functions. In section 5.3 we propose our modification of his paradigm, and derive the resulting conditions for optimality. In section 5.4 we compare our variational edge detection method with the anisotropic diffusion algorithm introduced by Perona and Malik. In section 5.5 and 5.6 we study some properties of the biased anisotropic diffusion. In section 5.7 we discuss discretizations of the variational edge detection problem, and propose numerical and analog circuit solutions. Section 5.8 is devoted to convergence, uniqueness and stability analysis of the discretized problem and the proposed algorithm. Finally section 5.9 covers our experimental results.

5.2 Terzopoulos' Edge Representation

From our brief review of stabilization in chapter 2 we recall that the "classical" stabilizers that first appeared in early vision problems did not allow estimation or reconstruction of image functions with discontinuities. In order to improve on this framework, Terzopoulos [21, 40] introduced a more general class of stabilizing functionals referred to as

controlled-continuity stabilizers. These are of the form

$$\mathcal{S}(w,z) \doteq \int_{\mathbf{R}^K} \sum_{i=1}^{I} w_i \sum_{k_1=1}^{K} \cdots \sum_{k_i=1}^{K} \left(\frac{\partial^i z}{\partial x_{k_1} \cdots \partial x_{k_i}} \right)^2 dx$$

where $w \doteq [w_1 \cdots w_I]^T$, and the weighting functions $w_1, \ldots, w_I : \mathbf{R}^K \to [0,1]$, referred to as *continuity control functions* are in general discontinuous. They are in particular able to make jumps to zero, and edges, where the partial derivatives of z of order $\geq j$ are allowed to be discontinuous, are represented by the sets

$$\bigcap_{i=j+1}^{I} w_i^{-1}(\{0\}) \qquad j = 0, \ldots, I-1 \tag{5.1}$$

For the edge cost Terzopoulos proposes the functional

$$\mathcal{E}(w) \doteq \int_{\mathbf{R}^K} \sum_{i=1}^{I} \lambda_i (1 - w_i)\, dx$$

where the constants $\lambda_1, \ldots, \lambda_I \in \overline{\mathbf{R}_+}$ satisfy $\sum_{i=1}^{I} \lambda_i > 0$. Unfortunately this paradigm fails to support a genuine variational technique for minimizing the total cost with respect to the continuity control function vector w. In fact it does so for a couple of reasons.

First of all, calculus of variations with respect to w requires that the space W of admissible continuity control functions is embedded in some topological vector space. Any continuity control function, which can be separated from the set of all strictly positive continuity control functions by this topology, that is any continuity control function, which represents an essential set of edges according to (5.1), will necessarily belong to the boundary of W. Hence the continuity control function vectors of particular interest, that is those representing edges, can be optimal, without being critical, that is, without resulting in a zero variation of the total cost with respect to w.

Secondly, if the variation of the total cost with respect to w is set to zero, one obtains the ridiculous condition

$$\sum_{k_1=1}^{K} \cdots \sum_{k_i=1}^{K} \left(\frac{\partial^i z}{\partial x_{k_1} \cdots \partial x_{k_i}} \right)^2 \equiv \lambda_i \qquad i = 1, \ldots, I$$

under which the total cost is completely independent of w. Thus the optimal continuity control function vector can not be found by means of calculus of variations, even if it does not represent an essential set of edges. Terzopoulos resolves this problem by first discretizing the entire space of continuity control functions; w is defined on a finite subset D—a dual

pixel grid—and only allowed to take the values 0 or 1. The edge cost is modified accordingly to

$$\mathcal{E}(w) \doteq \sum_{x \in D} \sum_{i=1}^{I} \lambda_i [1 - w_i(x)]$$

For a solution he then applies a descent method in the continuity control function vector space W^I. Prior to each update of w, the optimal estimated image function $\check{z}_w$ for the present w is computed, by solving the Euler equation—a partial differential equation in $\check{z}_w$—associated with the variational principle $\delta_z C_\zeta(w, z) = 0$. This method is expensive, and since the update $\blacktriangle w$ is based on the cost difference $C_\zeta(w + \blacktriangle w, \check{z}_w) - C_\zeta(w, \check{z}_w)$, as opposed to $C_\zeta(w + \blacktriangle w, \check{z}_{w+\blacktriangle w}) - C_\zeta(w, \check{z}_w)$—computation of $\check{z}_{w+\blacktriangle w}$ for all possible updates $\blacktriangle w$ would be far too expensive—convergence to a global minimum cannot be guaranteed.

5.3 Genuinely Variational Edge Detection

For our problem of detecting discontinuities of a bivariate image function, the appropriate deviation and stabilizing costs in the paradigm above are given by

$$\mathcal{D}(z, \zeta) \doteq \int_B (z - \zeta)^2 \, dx$$

and

$$\mathcal{S}(w, z) \doteq \int_B w \|\nabla z^T\|^2 \, dx$$

As in the earlier chapters we will assume that the image domain $B \subseteq \mathbf{R}^2$ is open bounded and connected. In order to remedy the difficulties with Terzopoulos' method, we propose the use of a smooth continuity control function $w : B \to \overline{\mathbf{R}_+}$. If w was prespecified, this would amount to the simplest straight forward generalization of Tikhonov stabilization to bivariate regularization. However, as Terzopoulos we will consider w to be a variable, and optimaize the total cost with respect to both w and z. To avoid the problem with optimal continuity control functions, which are noncritical, and thus impossible to find by means of variational calculus, we will arrange the edge cost, so that for each estimated image function z, the total cost $C_\zeta(w, z)$ attains its minimum for exactly one optimal continuity control function $\check{w}_z$, whose range is confined to lie in $]0, 1]$. This idea is similar to the use of barrier functions in finite dimensional optimaization [64]. The uniqueness of $\check{w}_z$ for a given z, also allows us to solve for $\check{w}_z$ in terms of z in a way similar to Blake and Zisserman's elimination

of their "line process" [23, 24]. The edge costs, we propose for this purpose, are of the form

$$\mathcal{E}(w) \doteq \int_B \lambda f \circ w \, dx$$

where the *edge cost coefficient* $\lambda > 0$ is constant, and the *edge cost density function* $f : \mathbf{R}_+ \to \mathbf{R}$ is twice differentiable. Our total cost functional is thus given by

$$C_\zeta(w, z) \doteq \int_B \left[\lambda f \circ w + (z - \zeta)^2 + w\|\nabla z^T\|^2\right] dx \tag{5.2}$$

It would be appropriate to multiply the stabilizing cost $\mathcal{S}(w, z)$ by the square of a (constant) scale-space parameter $\mu > 0$. However a true magnification of the scale of resolution of the edge detector should be equivalent to a shrinkage of the width and height of the image domain (along with the induced space scaling of the functions defined thereon) by the same factor. For any consistent discretization of the problem the effective scale-space parameter will therefore be inversely proportional to, and might as well be absorbed in, the pixel width h.

Setting the first variation of $C_\zeta(w, z)$ to zero yields the Euler equations

$$z(x) - \zeta(x) - \nabla \cdot (w \nabla z)(x) = 0 \qquad \forall x \in B \tag{5.3a}$$

$$\lambda f'(w(x)) + \|\nabla z(x)^T\|^2 = 0 \qquad \forall x \in B \tag{5.3b}$$

$$w(x)\frac{\partial z}{\partial e_n}(x) = 0 \qquad \forall x \in \partial B \tag{5.3c}$$

where $\nabla\cdot$ denotes the divergence operator, and $\partial/\partial e_n$ denotes the directional derivative in the direction of the outward normal. The second variation of C_ζ with respect to w is also easily found to be

$$\delta^2_{ww} C_\zeta(w, z) = \int_B \frac{\lambda}{2}(f'' \circ w)(\delta w)^2 \, dx \tag{5.4}$$

Together with the desired existence of a unique optimal continuity control function $\bar{w}_z$ for each possible estimated image function z these equations put some restrictions on the edge cost density f. In fact from (5.3b) it follows, that $f'|]0, 1] \to \overline{\mathbf{R}_-}$ must be bijective, and that $f'(]1, \infty[) \subseteq \mathbf{R}_+$. Likewise from (5.4) we see, that f'' must be strictly positive on $]0, 1[$, and that $f''(1) \geq 0$. the simplest functions, which satisfy these requirements are given by

$$f(\omega) \doteq \omega - \ln \omega \qquad \Rightarrow \qquad f'(\omega) = 1 - \frac{1}{\omega} \tag{5.5}$$

and

$$f(\omega) \doteq \omega \ln \omega - \omega \qquad \Rightarrow \qquad f'(\omega) = \ln \omega \tag{5.6}$$

but there are of course many other possibilities, for example:

$$f(\omega) \doteq \omega + \frac{1}{(p-1)\omega^{p-1}} \qquad \Rightarrow \qquad f'(\omega) = 1 - \frac{1}{\omega^p} \qquad p \in \mathbf{R}_+ \setminus \{1\} \tag{5.7}$$

However, some choices of p might be better than others. In section 5.6 we will present an argument supporting the further restriction, that $p < 2$. Another example, which might be of special interest to circuit designers, consists of the somewhat involved edge cost density

$$f(\omega) \doteq \begin{cases} -\omega(\ln\omega)^2 + 2\omega\ln\omega - 2\omega & \text{if } \omega \in]0,1] \\ \omega(\ln\omega)^2 - 2\omega\ln\omega + 2\omega - 4 & \text{if } \omega \in]1,\infty[\end{cases} \tag{5.8}$$

with derivative

$$f'(\omega) = \begin{cases} -(\ln\omega)^2 & \text{if } \omega \in]0,1] \\ (\ln\omega)^2 & \text{if } \omega \in]1,\infty[\end{cases} \tag{5.9}$$

Given that f satisfies these conditions, $f'|]0,1]$ is invertible, and since w is strictly positive, we end up with the equations

$$z(x) = \zeta(x) + \nabla \cdot (w\nabla z)(x) \qquad \forall x \in B \tag{5.10a}$$

$$w(x) = g(\|\nabla z(x)^T\|) \qquad \forall x \in B \tag{5.10b}$$

$$\frac{\partial z}{\partial e_n}(x) = 0 \qquad \forall x \in \partial B \tag{5.10c}$$

where the function $g : \overline{\mathbf{R}_+} \to]0,1]$, (for reasons soon to make sense,) referred to as the *diffusivity anomaly*, is defined by

$$g(\gamma) \doteq (f'|]0,1])^{-1}\left(-\frac{\gamma^2}{\lambda}\right) \qquad \gamma \geq 0 \tag{5.11}$$

The properties of the edge cost density f clearly imply, that g is a strictly positive strictly decreasing differentiable bijection. In particular $g(0) = 1$, and $\lim_{\gamma\to\infty} g(\gamma) = 0$. For the edge cost densities in (5.5) and (5.6) the diffusivity anomaly depends explicitly on the square of its argument, and takes the forms

$$g(\gamma) \doteq \frac{1}{1 + \frac{\gamma^2}{\lambda}} \qquad \gamma \geq 0 \tag{5.12}$$

and

$$g(\gamma) \doteq e^{-\frac{\gamma^2}{\lambda}} \qquad \gamma \geq 0 \tag{5.13}$$

respectively. In contrast the edge cost density in (5.9) yields

$$g(\gamma) \doteq e^{-\frac{\gamma}{\sqrt{\lambda}}} \qquad \gamma \geq 0 \tag{5.14}$$

Since our method necessarily yields continuity control functions, for which

$$w^{-1}(\{0\}) = \varnothing$$

Terzopoulos' edge representation is inadequate. The simplest and most reasonable modification is to consider the edges to consist of the set $w^{-1}(]0,\theta])$, where θ is a fixed threshold. Since the diffusivity anomaly g is strictly decreasing, we have

$$w^{-1}(]0,\theta]) = \|\nabla z^T\|^{-1}([g^{-1}(\theta),\infty[)$$

whence the edges are obtained by thresholding the magnitude of the gradient of the estimated image function.

Other possibilities are of course possible. One could for example attempt to detect various desired edge patterns by filtering w. One could also try to make the threshold adaptive, and/or let it depend on the position x in some clever way. With attempts along these lines however, one will most likely tend to stray away from the original optimality principle, and End up in the kind of hacker's nest that the introduction of such a principle was initially meant to avoid.

5.4 Biased Anisotropic Diffusion

Perona and Malik [27, 29] have introduced anisotropic diffusion as a method of suppressing finer details, without weakening or dislocating the larger scale edges. The initial value problem governing their method is given by

$$\frac{\partial z}{\partial t}(x,t) = \nabla \cdot (w\nabla z)(x,t) \qquad \forall x \in B \qquad \forall t > 0 \tag{5.15a}$$

$$w(x,t) = g(\|\nabla z(x,t)^T\|) \qquad \forall x \in B \qquad \forall t > 0 \tag{5.15b}$$

$$\frac{\partial z}{\partial e_n}(x,t) = 0 \qquad \forall x \in \partial B \qquad \forall t > 0 \tag{5.15c}$$

$$z(x,0) = \zeta(x) \qquad \forall x \in B \tag{5.15d}$$

where the diffused image function z and the *diffusivity* w are functions of both position $x \in B$ and time $t \geq 0$, $\nabla\cdot$ and ∇ denote the divergence and the gradient respectively with respect to x, and the diffusivity anomaly $g : \overline{\mathbf{R}_+} \to \overline{\mathbf{R}_+}$ is a decreasing function.

As the name "anisotropic diffusion" suggests, these equations have appealing physical interpretations. The function z can for example be thought of as representing the

temperature T in a thin slab S of a material, whose initial temperature is given by ζ, and whose space- and time-varying thermal diffusivity (or thermal conductivity, if time is scaled appropriately,) is given by w. This analogy is depicted in figure 5.1.

The Euler equations we derived in the previous section are very similar to the initial value problem (5.15). In fact, a solution of (5.10) is given by the steady state of the initial value problem

$$\frac{\partial z}{\partial t}(x,t) = \zeta(x,t) - z(x,t) + \nabla \cdot (w \nabla z)(x,t) \qquad \forall x \in B \qquad \forall t > 0 \qquad (5.16a)$$

$$w(x,t) = g(\|\nabla z(x,t)^T\|) \qquad \forall x \in B \qquad \forall t > 0 \qquad (5.16b)$$

$$\frac{\partial z}{\partial e_n}(x,t) = 0 \qquad \forall x \in \partial B \qquad \forall t > 0 \qquad (5.16c)$$

$$z(x,0) = \chi(x) \qquad \forall x \in B \qquad (5.16d)$$

which is obtained from (5.15) by replacing the anisotropic diffusion equation (5.15a) by the closely related *"biased" anisotropic diffusion* equation (5.16a). Since our interest is in the steady state solution, the initial condition (5.15d) can also be replaced by an arbitrary initial condition (5.16d). The continuity control function w thus plays the role of the diffusivity, and will be referred to as such, whenever the context so suggests.

The bias term $\zeta - z$ in (5.16a) intuitively has the effect of locally moderating the diffusion as the diffused image function z diffuses further away from the original image function ζ. It is therefore reasonable to believe, that a steady state solution does exist. The following physical interpretation of this initial value problem further substantiates this belief: Let S be a thin slab of some material resting on top of another slab S_0 of some (other) material as in figure 5.2. Suppose that the space- and time-varying thermal conductivity of S is given by αw, where the constant $\alpha > 0$ is the coefficient of heat transfer between S and S_0. If the initial temperature at each point $x \in B$ of S is given by $\chi(x)$, and the temperature distribution of S_0 is held fixed at ζ, then z represents the space- and time-varying temperature of S. Besides supporting useful intuition about our variational edge detection method, the analogy above suggests a physical mechanism, which could serve as a model for the design of analog circuits, which realize the solutions of the boundary value problem (5.10).

The possibility of suppressing finer details, while the more significant edges remain intact, or are even strengthened, is a consequence of the anisotropy, which in both the diffusions described above in turn is caused by the nonconstancy of the diffusivity anomaly g.

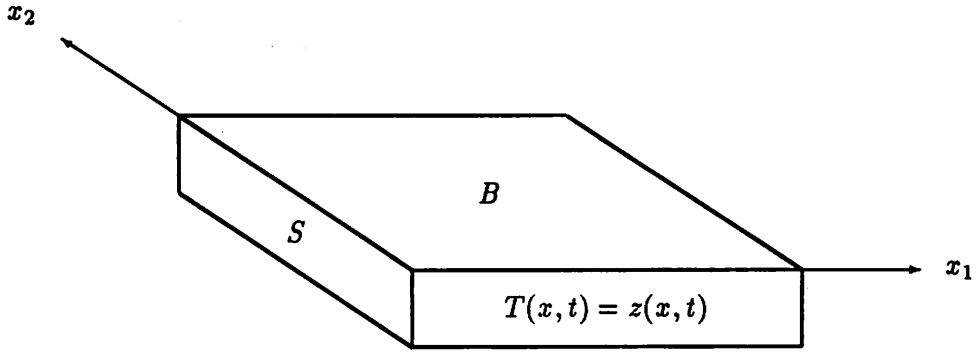


Figure 5.1: Physical model of unbiased anisotropic diffusion.

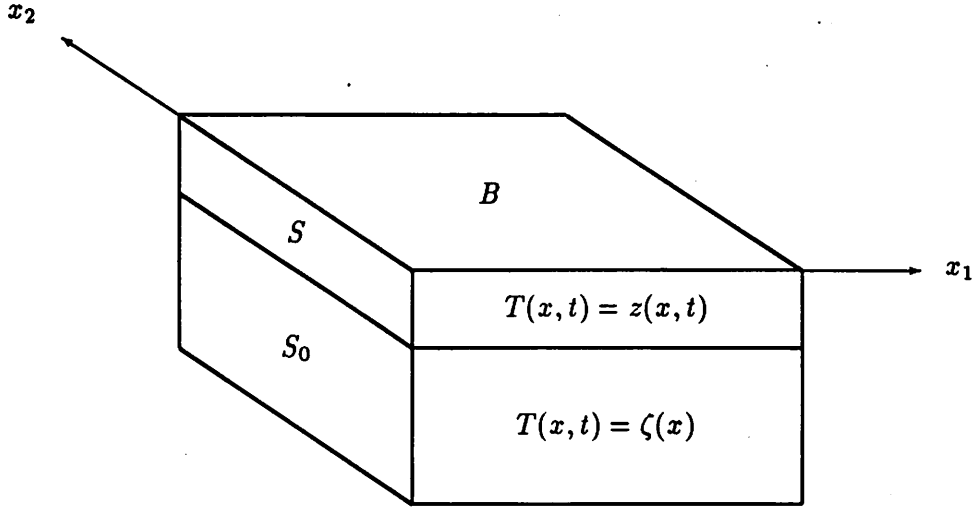


Figure 5.2: Physical model of biased anisotropic diffusion.

If g is constant, the *unbiased* diffusion (5.15a) reduces to Gaussian blurring, while the steady state of the biased diffusion (5.16a) in a sense corresponds to filtering with a doubly cascaded first order Butterworth filter. For our variational method governed by the boundary value problem (5.10), the choice of g was based on optimality considerations. Perona and Malik select their function g, by demanding, that the resulting unbiased anisotropic diffusion enhances the already pronounced edges, while the less significant edges are weakened. Based on an analysis including only blurred linear step edges—an unnecessary restriction, as we shortly shall see—they vouch for diffusivity anomalies of the form

$$g(\gamma) \doteq \frac{c}{1+\left(\frac{\gamma^2}{\lambda}\right)^a} \tag{5.17}$$

where $c, \lambda > 0$ and $a > 1/2$ are constants. It is easy to check, that, if these functions were substituted in the Euler equation (5.10b), the corresponding edge cost densities would satisfy the requirements of our variational method. (To be precise, the constant c would actually have to be equal to unity. This is however an artifact, which would not have surfaced, had we incorporated the scale-space parameter μ, and vanishes regardless in the discretization process.) Incidentally, for their experimental results, Perona and Malik use exactly the functions, we proposed in (5.12) and (5.13), of which only the former belongs to the class specified by (5.17).

Finally we note, that the heuristically motivated method that Perona and Malik used for extracting a set of edges from the diffused image function, is practically identical to the method implied by our edge representation in terms of the continuity control function. While they threshold the absolute difference between four-connected neighbor pixel values, our edge representation leads, as we saw in the previous section, to thresholding of the magnitude of the gradient.

5.5 The Extremum Principle

The extremum principle is a common tool for proving uniqueness and stability with respect to boundary data for linear elliptic and linear parabolic problems [65]. For quasi-linear equations, such as the Euler equation (5.10a) and the biased anisotropic diffusion equation (5.16a), it is not quite as conclusive. Nevertheless it provides bounds on the solution and useful insight for convergence analysis of the numerical methods employed to find such a solution. We will present an extremum principle for the biased anisotropic

diffusion problem (5.16) as well as for the boundary value problem (5.10). In both cases we will assume that the diffusivity anomaly $g : \overline{\mathbf{R}_+} \to \overline{\mathbf{R}_+}$ is continuously differentiable.

Theorem 5.5.1 *Let $z : \overline{B} \times \overline{\mathbf{R}_+} \to \mathbf{R}$ be a solution of the biased anisotropic diffusion problem (5.16), where it is assumed, that $\zeta : B \to \mathbf{R}$ is uniformly continuous. Assume further, that z and its first and second order partial derivatives with respect to x are continuous (on $\overline{B} \times \overline{\mathbf{R}_+}$). Then the following claims are true:*

(i) *If $\pm y_\tau : \overline{B} \to \mathbf{R} : x \mapsto \pm z(x,\tau)$ has a local maximum at $\xi \in \overline{B}$ for some fixed $\tau > 0$, then*

$$\pm \frac{\partial z}{\partial t}(\xi,\tau) \leq \pm\zeta(\xi) \mp z(\xi,\tau)$$

(ii) *If $\pm z$ has a local maximum at $(\xi,\tau) \in \overline{B} \times \mathbf{R}_+$, then*

$$\pm z(\xi,\tau) \leq \pm\zeta(\xi)$$

(iii) $$\inf_{\xi\in B}[\zeta(\xi) \wedge \chi(\xi)] \leq z(x,t) \leq \sup_{\xi\in B}[\zeta(\xi) \vee \chi(\xi)] \qquad \forall x \in \overline{B} \qquad \forall t \geq 0$$

Proof: From (5.16a) and the continuity assumptions regarding g, ζ and z one can show, that $\partial z/\partial t$ is uniformly continuous on $B \times T$ for every bounded interval $T \subseteq \mathbf{R}_+$, and therefore has a unique continuous extension to $\overline{B} \times \mathbf{R}_+$. By the bounded convergence theorem of integration it also follows, that this extension equals $\partial z/\partial t$ on the boundary $\partial B \times \mathbf{R}_+$. Hence (5.16a) is satisfied on all of $\overline{B} \times \mathbf{R}_+$ (with the appropriate one-sided derivatives on $\partial B \times \mathbf{R}_+$). Suppose that $\pm y_\tau$ has a local maximum at $\xi \in \overline{B}$ for some $\tau > 0$. Then by Taylor's formula (and the Neumann condition (5.16c), if $\xi \in \partial B$) we have, that $\nabla y_\tau(\xi) = 0$, and $\pm\Delta y_\tau(\xi) \leq 0$. Thus

$$\pm\nabla \cdot (w\nabla z)(\xi,\tau) = \pm\nabla w(\xi,\tau)\nabla y_\tau(\xi)^T \pm w(\xi,\tau)\Delta y_\tau(\xi) \leq 0$$

whence (i) follows. Suppose next, that $\pm z$ has a local maximum at $(\xi,\tau) \in \overline{B} \times \mathbf{R}_+$. Then $\pm y_\tau$ has a local maximum at ξ, and $\partial z/\partial t(\xi,\tau) = 0$. Hence (ii) follows from (i). Finally consider the compact set $\overline{B} \times [0, T_1]$, on which the continuous functions $\pm z$ attain their maximal values, say at $(\xi_\pm, \tau_\pm)$. If $\tau_\pm = T_1$, then $\pm\partial z/\partial t(\xi_\pm,\tau_\pm) \geq 0$, and $\pm y_{\tau_\pm}$ has a local maximum at $\xi_\pm$. Hence (i) implies, that $\pm z(\xi_\pm,\tau_\pm) \leq \pm\zeta(\xi_\pm)$. If $\tau_\pm \in]0, T_1[$, the same conclusion follows immediately from (ii). Since $T_1 > 0$ was arbitrarily chosen, this shows, that

$$\sup_{(x,t)\in\overline{B}\times\mathbf{R}_+} \pm z(x,t) \leq \sup_{x\in B} \pm\zeta(x)$$

from which (iii) follows. ■

For the boundary value problem (5.10) governing our variational edge detection method a proof similar to that above yields the following extremum principle:

Theorem 5.5.2 *Let $z : \overline{B} \to \mathbf{R}$ be a solution of the boundary value problem (5.10). Assume further, that z and its first and second order partial derivatives are continuous (on $\overline{B}$). Then*

$$\inf_{\xi \in B} \zeta(\xi) \leq z(x) \leq \sup_{\xi \in B} \zeta(\xi) \qquad \forall x \in \overline{B}$$

We remark, that in both the theorems above, the assumption, that the derivatives of z are continuous up to *and including* the boundary ∂B ($\times\overline{\mathbf{R}_+}$), (or equivalently uniformly continuous on every bounded subset of the interior of the domain of z,) could have been traded for a weaker bound on z, which in addition to the values of z on B ($\times\{0\}$) also includes those on ∂B ($\times\overline{\mathbf{R}_+}$). However, for the discretized problem, that we eventually will have to solve, the subtle difference between plain vs. uniform continuity of z and its derivatives is of no consequence. The "stronger-assumption-conclusion" versions of the extremum principles presented above are therefore more useful in this context.

According to the two theorems above the solutions of the biased anisotropic diffusion problem are well-behaved, in that they do not stray too far away from the original image function ζ, unless forced to by the initial condition, and even if so, they eventually approach the range of ζ as $t \to \infty$. In plain language condition (i) of the first theorem says, that the diffused image process, at each of its momentary critical points (with respect to x) is headed towards the original image function. Condition (ii) of the same theorem says, that all the noninitial local extrema of the diffused image process are within the range of the original image function, and condition (iii) gives explicit bounds on the entire collection of diffused image functions in terms of the initial and original image functions. The second theorem bounds the steady state diffused image function in terms of the original data alone. In other words, our variational edge detection method produces an estimated image function, whose range is contained inside that of the original image function.

5.6 Edge Enhancement

It was mentioned earlier, that the biased anisotropic diffusion (5.16), in similarity with its unbiased counterpart (5.15), has the important property of suppressing finer details, while strengthening the more significant edges. Indeed, the edges are roughly either sharpened or blurred depending on their present strength, viz. the magnitude of the gradient of the diffused image function z.

In order to see this, we define the edges to consist of the points in the image domain B, at which the magnitude of the gradient of the diffused image function has a strict local maximum along the direction perpendicular to the edge, that is the direction of the gradient. For simpler notation we let $\sigma \doteq \|\nabla z^T\|$. We also define e_ν and e_τ to be the unit vectors in the directions of $[\partial z/\partial x_1 \quad \partial z/\partial x_2]$ and $[\partial z/\partial x_2 \quad -\partial z/\partial x_1]$ respectively, that is e_ν is normal, and e_τ is tangential to the edge. Since $\sigma > 0$ on the edges, e_ν and e_τ are well-defined on the points of interest. The edge points can now be characterized by:

$$\frac{\partial \sigma}{\partial e_\nu} = 0 \tag{5.18a}$$

$$\frac{\partial^2 \sigma}{\partial e_\nu^2} < 0 \tag{5.18b}$$

For a typical edge of interest it is reasonable to assume that its strength σ exhibits a fairly pronounced peak along its *perpendicular* direction, resulting in a large value of $|\partial^2\sigma/\partial e_\nu^2|$. On the other hand σ can be expected to vary quite moderately—at most with a fairly constant derivative (shading component)—*along* the edge, with a small value of $|\partial^2\sigma/\partial e_\tau^2|$ as a consequence. We will therefore at little loss allow ourselves to restrict attention to edge points, at which

$$\frac{\partial^2 \sigma}{\partial e_\nu^2} \approx \Delta\sigma < 0 \tag{5.19}$$

Our discussion includes in particular all symmetrically blurred (smooth) step edges. For points on such edges the approximation (5.19) is indeed exact, even if the size of the step varies linearly with arc length along the edge.

We begin by noting, that

$$\nabla\sigma\nabla z^T = \frac{\partial \sigma}{\partial e_\nu}\sigma$$

and

$$\frac{\partial z}{\partial e_\nu} = \sigma \tag{5.20}$$

Assuming that all functions involved are sufficiently smooth, and that the diffusivity is of the usual form $w = g \circ \sigma$, from (5.18a) and (5.19) we then have.

$$\begin{aligned}
&\frac{\partial}{\partial e_\nu}\nabla \cdot (w\nabla z) = \\
&\quad = \frac{\partial}{\partial e_\nu}(\nabla w \nabla z^T + w\Delta z) \\
&\quad = \frac{\partial}{\partial e_\nu}[(g' \circ \sigma)\nabla\sigma\nabla z^T] + \frac{\partial w}{\partial e_\nu}\Delta z + w\frac{\partial}{\partial e_\nu}\Delta z \\
&\quad = \frac{\partial}{\partial e_\nu}\left[(g' \circ \sigma)\frac{\partial \sigma}{\partial e_\nu}\sigma\right] + (g' \circ \sigma)\frac{\partial \sigma}{\partial e_\nu}\Delta z + w\Delta\frac{\partial z}{\partial e_\nu} \\
&\quad = (g' \circ \sigma)\frac{\partial^2 \sigma}{\partial e_\nu^2}\sigma + (g \circ \sigma)\Delta\sigma \\
&\quad \approx [(g' \circ \sigma)\sigma + g \circ \sigma]\Delta\sigma
\end{aligned}$$

From (5.20) it also follows that

$$\frac{\partial \sigma}{\partial t} = \frac{\partial}{\partial e_\nu}\frac{\partial z}{\partial t}$$

Hence on the edges, the biased anisotropic diffusion (5.16) causes the edge strength to vary with time according to

$$\frac{\partial \sigma}{\partial t} \approx \frac{\partial \zeta}{\partial e_\nu} - \frac{\partial z}{\partial e_\nu} + (\varphi' \circ \sigma)\Delta\sigma$$

where

$$\varphi(\gamma) \doteq g(\gamma)\gamma \qquad \gamma \geq 0 \tag{5.21}$$

Rewriting this equation as

$$\frac{\partial}{\partial t}(\sigma - \sigma_\zeta) \approx -(\sigma - \sigma_\zeta) + (\varphi' \circ \sigma)\Delta\sigma \tag{5.22}$$

where $\sigma_\zeta \doteq \partial\zeta/\partial e_\nu$, it is clear, that the bias term $-(\sigma - \sigma_\zeta)$ merely has a moderating effect on the enhancement/blurring of the edge, while the decision between enhancement vs. blurring depends on the sign of the "driving" term $(\varphi' \circ \sigma)\Delta\sigma$ associated with the unbiased anisotropic diffusion.

For the desired performance of weakening the weak edges, while strengthening the strong ones in a consistent manner, since $\Delta\sigma < 0$, it is therefore necessary, that there exists an *edge enhancement threshold* $\gamma_0 \in \mathbf{R}_+$, such that

$$\varphi'^{-1}(\mathbf{R}_-) =]\gamma_0, \infty[\tag{5.23a}$$

$$\varphi'^{-1}(\{0\}) = \{\gamma_0\} \tag{5.23b}$$

$$\varphi'^{-1}(\mathbf{R}_+) = [0, \gamma_0[\tag{5.23c}$$

Furthermore, if so, the threshold γ_0 clearly controls the sensitivity of the edge detector, and one would hence expect it to be closely related to the intuitively similarly acting edge cost coefficient λ. Indeed from (5.11) and (5.21) it immediately follows, that $\varphi'(\gamma)$ is a function of γ^2/λ. Since $\overline{\mathbf{R}_+} \to \overline{\mathbf{R}_+} : \gamma \mapsto \gamma^2/\lambda$ is strictly monotone, γ_0 must therefore be proportional to $\sqrt{\lambda}$. It is easy to verify, that the diffusivity anomalies given in (5.12), (5.13) and (5.14) satisfy (5.23) with $\gamma_0 = \sqrt{\lambda}$, $\gamma_0 = \sqrt{\lambda/2}$ and $\gamma_0 = \sqrt{\lambda}$ respectively. For the diffusivity anomalies corresponding to the edge cost densities in (5.7) on the other hand, an edge enhancement threshold $\gamma_0 = \sqrt{p\lambda/(2-p)}$ satisfying (5.23) will exist if and only if $p \in]0,1[\cup]1,2[$.

Although the discussion above generates some useful insight, and offers guidelines for sensible choices of the diffusivity anomaly g, it is not completely satisfactory, in that it does not account for the change in location and orientation of the edges in the image domain during the diffusion process. In fact, by evaluating the second partial derivative $\partial^2\sigma/\partial t\partial e_\nu$, one can show, that only edges with certain symmetry properties, for example symmetrically blurred linear step edges, will remain fixed in position, during the diffusion. If one neglect this weakness—a forgotten subject in previous papers—one could be misled to believe, that the enhancement/blurring decisions about the edges are completely determined by the local properties of the original image function ζ at the edge points. If this was true, one could just as well detect the edges, by checking these properties, amounting to nothing more, than thresholding the directional derivative of ζ in the direction of its gradient at points, where this derivative has local maxima—a previously explored paradigm in local edge detection [13, 66].

If the diffused image function converges to a *unique* steady state solution, that is a solution of the boundary value problem (5.10), the edge enhancement/blurring is of course in the limit independent of the initial condition (5.16d). Indeed from (5.22) we immediately obtain the steady state edge enhancement

$$\sigma - \sigma_\zeta = (\varphi' \circ \sigma)\Delta\sigma \tag{5.24}$$

This equation is clearly valid independently of how the edges move during the diffusion process. On the other hand σ_ζ is not representative of the original edge strength $\|\nabla\zeta^T\|$, if ∇z and $\nabla\zeta$ differ too much in orientation.

Since the range of the steady state solution, by the extremum principle, is confined to lie within the range of the original image function, an amply enhanced edge strength σ

can only be maintained along a very short distance across the edge. Such edges are therefore sharpened.

For the numerical solution of the boundary value problem (5.10) on a regular computer there are, as we shall shortly discuss, good reasons for updating the estimated image function according to a rule, different from a straight forward discretization of the biased anisotropic diffusion equation. However, the final edge enhancement (5.24) is independent of the path to the solution, so the discussion above is still valid.

Besides being of vital importance for the edge enhancement mechanism, the existence of the edge enhancement threshold γ_0 also provides a natural choice for the threshold to be used in the postprocessing, whereby the edges are finally extracted from the estimated image function. It is intuitively clear, that, for our edge representation to be consistent with the edge enhancement mechanism, the edge representation threshold in section 5.3 should be given by $\theta \doteq g(\gamma_0)$. The edge set $w^{-1}(]0,\theta])$ will then consist of the points in the image domain, where the magnitude of the gradient of the estimated image function exceeds γ_0, that is exactly those points, where the edge strength has been enhanced. On the other hand, and this is in a sense the essential benefit with our regularization approach, the bistability of the edge enhancement mechanism will deplete the set of points, at which the gradient magnitude of the estimated image function takes values close to γ_0. The edge set $w^{-1}(]0,\theta])$ will therefore be almost indifferent to changes in θ, as long as θ belongs to some substantial neighborhood of $g(\gamma_0)$. These circumstances are clearly ideal for thresholding, and consequently our edge representation is practically consistent with the edge enhancement mechanism for a whole interval of edge representation thresholds, corresponding to a relatively wide range of gradient magnitudes.

5.7 Discretization

For a numerical solution of the variational edge detection problem in section 5.3 the boundary value problem (5.10) has to be discretized. The original image function ζ is most likely already given only on a squared pixel grid. Assuming that this is the case, the simplest way of discretizing the image functions z and ζ for the numerical problem, is obviously to use the same grid. For the evaluation of the expression $\nabla \cdot (w \nabla z)$ there are on the contrary a number of more or less equally sensible choices. One can for example expand

$\nabla \cdot (w\nabla z)$ in terms of z and its derivatives according to

$$\nabla \cdot (w\nabla z) = (g' \circ \sigma)\frac{1}{\sigma}\nabla z\, \mathbf{H}z\, \nabla z^T + (g \circ \sigma)\Delta z$$

where $\sigma \doteq \|\nabla z^T\|$, and $\mathbf{H}$ denotes the Hessian operator. With the numerous discrete approximations of ∇z and $\mathbf{H}z$ at hand, this leaves a multitude of open possibilities. Alternatively one can treat $w\nabla z$ as a single function, readily evaluated at appropriate points in terms of some discrete approximation of ∇z, and then take some discrete approximation of its divergence. We have settled for the latter approach, which has the special quality of highlighting the diffusion mechanism. This in turn naturally leads to expressions, which are particularly convenient for both hardware and software implementation as well as for theoretical analysis of the resulting algorithms.

To be more specific, let us consider an original image function ζ, given on a grid $\{jh : j \in J\}$, where $h > 0$ is the pixel width, and $J \doteq \{1, \ldots, J_1\} \times \{1, \ldots, J_2\}$ for some $J_1, J_2 \in \mathbb{N}$. The corresponding image domain is thus given by $B \doteq]\frac{1}{2}, J_1 + \frac{1}{2}[\times]\frac{1}{2}, J_2 + \frac{1}{2}[$. We then define the discretized shifted functions

$$\begin{array}{lll} \zeta_0(j) \doteq \zeta(jh) & j \in J & \\ z_q(j) \doteq z(\arg\min_{l \in J} \|l - j - q\| \cdot h) & j \in J & q \in S \\ w_q(j) \doteq g(\sigma_q(j)) & j \in J & q \in S \end{array}$$

where $S \doteq \{-1, 0, 1\}^2$, and $\sigma_q(j)$ is some discrete approximation of $\|\nabla z((j + \frac{q}{2})h)^T\|$. It is reasonable to demand, that $\sigma_q(j)$ be specified in terms of z at the smallest possible symmetric set of neighboring grid points of $(j + \frac{q}{2})h$. This requirement leads to the discrete approximations:

$$\sigma_q^2 \doteq \frac{(z_{q_1,0} - z_{0,1} + z_{q_1,-1} - z_{0,0})^2 + (z_{q_1,0} - z_{0,-1} + z_{q_1,1} - z_{0,0})^2}{8h^2} \qquad q_1 = \pm 1 \quad (5.25a)$$

$$\sigma_q^2 \doteq \frac{(z_{0,q_2} - z_{1,0} + z_{-1,q_2} - z_{0,0})^2 + (z_{0,q_2} - z_{-1,0} + z_{1,q_2} - z_{0,0})^2}{8h^2} \qquad q_2 = \pm 1 \quad (5.25b)$$

$$\sigma_q^2 \doteq \frac{(z_{q_1,q_2} - z_{0,0})^2 + (z_{q_1,0} - z_{0,q_2})^2}{2h^2} \qquad q_1, q_2 \in \{-1, 1\} \quad (5.25c)$$

where we have dropped the dependence of $j \in J$ for shorter notation, and written z_{q_1,q_2} for z_q. The two discrete approximations of (5.10), which immediately come to mind, can after some manipulation (from a variety of starting points) be written as

$$\zeta_0 - z_0 + \frac{1}{\rho^2 h^2} \sum_{q \in S_\rho} w_q(z_q - z_0) = 0 \qquad \rho^2 = 1, 2 \quad (5.26)$$

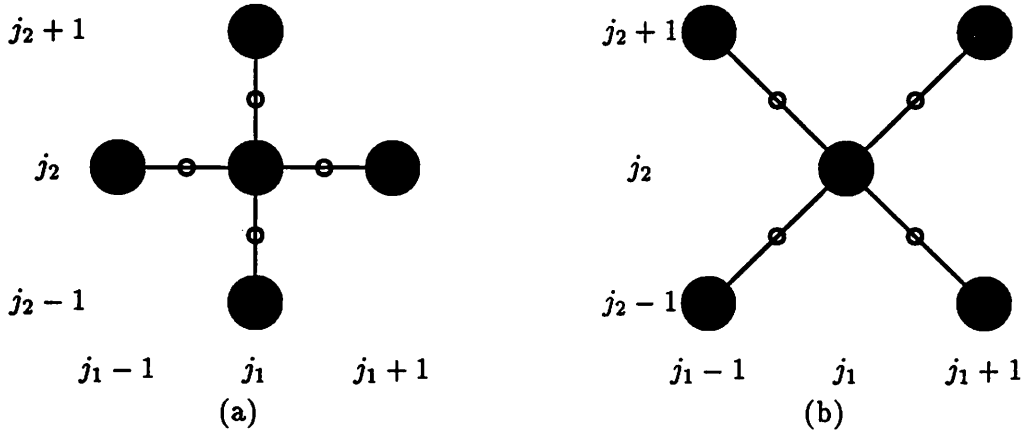


Figure 5.3: Discrete approximation molecule structures. (a) "Cartesian"; $\rho^2 = 1$. (b) "Diagonal"; $\rho^2 = 2$.

where $S_\rho \doteq \{q \in S : \|q\| = \rho\}$, $\rho^2 = 1, 2$. Note, that the Neumann condition (5.10c) is conveniently taken care of by the "arg min"-adjustment in (5.25b), which systematically replaces any otherwise required value of z at a grid point outside B, by the value of z at the closest grid point inside B, thereby ensuring that

$$\left.\begin{aligned} z_{-1,q_2}(1,j_2) - z_{0,q_2}(1,j_2) &= 0 \\ z_{1,q_2}(J_1,j_2) - z_{0,q_2}(J_1,j_2) &= 0 \end{aligned}\right\} \quad j_2 = 1,\ldots,J_2 \qquad q_2 = -1,0,1$$

$$\left.\begin{aligned} z_{q_1,-1}(j_1,1) - z_{q_1,0}(j_1,1) &= 0 \\ z_{q_1,1}(j_1,J_2) - z_{q_1,0}(j_1,J_2) &= 0 \end{aligned}\right\} \quad j_1 = 1,\ldots,J_1 \qquad q_1 = -1,0,1$$

The computational molecules associated with the two approximations, $\rho^2 = 1, 2$, in (5.26) have the structures depicted in figure 5.3, where the filled circles (atoms) mark the sites associated with the evaluation of ζ and z, and the empty circles (bonds centers) mark the sites associated with the evaluation of w. In each case the sum involved contains four terms.

The "Cartesian" approximation has a few apparent advantages. First of all the simple structure of its molecules makes it ideal for hardware implementations—an issue we will return to shortly. Secondly it provides tight coupling between all pairs of eight-connected pixel neighbors. In contrast, as one can see from figure 5.3, the "diagonal" approximation results in two interleaved but separated computational lattices. An algo-

rithm based on this approximation therefore models two more or less separate diffusion processes, which are coupled only through the shared diffusivity function, that is the coefficient function(s) of the quasi-linear equation (5.26). For original image functions with mildly well-behaved statistics however, the smoothing effect of the diffusion will, as one would guess, and as our experimental results also indicate, cure this problem. A third minor advantage of the Cartesian approximation is, that its associated truncation error is only $\sqrt{2}/4$ times that of the diagonal approximation.

The diagonal approximation also has a couple of advantages: As figure 5.3 reveals, it requires only half as many evaluations of the continuity control function, as does the Cartesian approximation. In addition these evaluations are simpler, as they are governed by (5.25c) as opposed to (5.25a) and (5.25b) in the Cartesian case. The diagonal approximation thus leads to faster and simpler software implementations. Our experiments further show, that it, despite its drawbacks, yields excellent results.

There are several possible ways of solving the discretized equation (5.26) numerically. One method, which is obvious in the light of the discussion in the previous sections, is to propagate the corresponding discretized biased anisotropic diffusion equation

$$\begin{aligned} z_0^{(0)} &\doteq \chi_0 \\ z_0^{(i+1)} &\doteq z_0^{(i)} + k\left[\zeta_0 - z_0^{(i)} + \frac{1}{\rho^2 h^2}\sum_{q\in S_\rho} w_q^{(i)}(z_q^{(i)} - z_0^{(i)})\right] \qquad \rho^2 = 1,2 \end{aligned}$$

where the *initial image function* $\chi_0 : J \to \mathbf{R}$ is arbitrary, most naturally chosen equal to ζ_0, $k > 0$ is the time step size, and $i \in \mathbf{N}_0$ is an iteration index, representing the time variable t according to: $t = ik$. However this algorithm is numerically stable only for sufficiently small values of the step size k, and safe play will necessarily bring down the convergence rate. Since we are not interested in the diffusion per se, but merely its steady state solution, this problem can be avoided, by choosing some robuster iteration method. Such methods are easily generated by treating the quasi-linear equation (5.26) as a linear elliptic equation, and applying any of the commonly used Jacobi, Gauss-Seidel or successive over-relaxation methods. The Jacobi method for example yields the iteration scheme:

$$z_0^{(0)} \doteq \chi_0 \tag{5.27a}$$

$$z_0^{(i+1)} \doteq \frac{1}{\rho^2 h^2 + \overline{w}_\rho^{(i)}}\left[\rho^2 h^2 \zeta_0 + \sum_{q\in S_\rho} w_q^{(i)} z_q^{(i)}\right] \tag{5.27b}$$

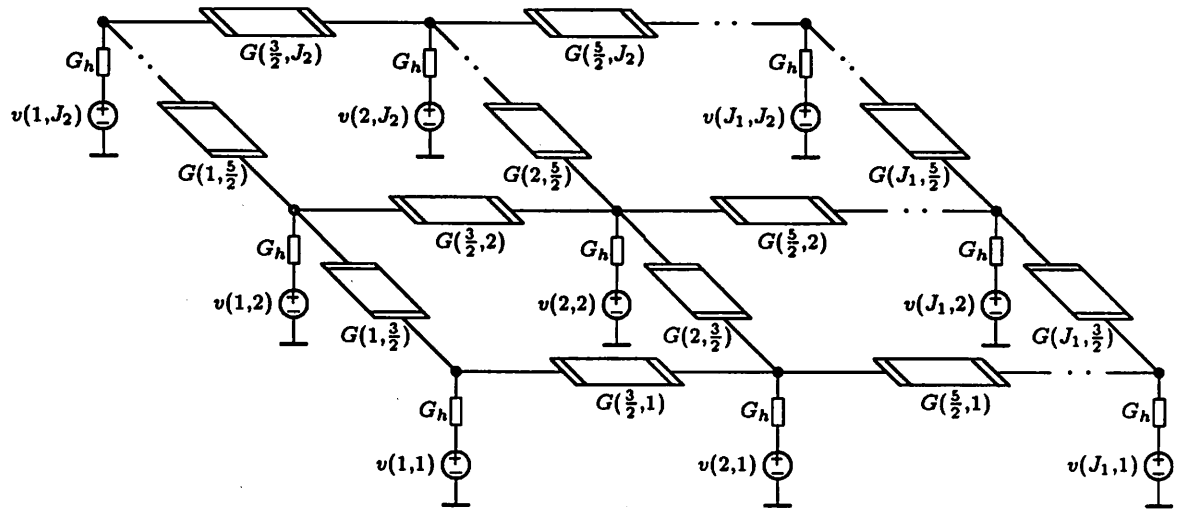


Figure 5.4: Analog circuit for solving the variational edge detection problem (5.10).

$$\overline{w}_\rho^{(i)} \doteq \sum_{q \in S_\rho} w_q^{(i)} \tag{5.27c}$$

Aside from numerical methods equation (5.26) can be solved by means of analog circuits. Indeed, by inspection of the Cartesian approximation ($\rho^2 = 1$), one immediately sees, that its solution z_0 is realized by the (steady state) potentials of the upper layer of nodes in the resistive network in figure 5.4, where the voltages $v(j)$, $j \in J$, represent the original image function, the controlled conductances $G(j + \frac{q}{2})$, $j \in J \cap (J - q)$, $q \in S_1$, model the continuity control function, and the fixed conductance G_h determines the scale of resolution. More precisely, if

$$\begin{aligned} v(j) &= v_1 \zeta_0(j) && \forall j \in J \\ G(j + \frac{q}{2}) &= G_1 w_q(j) && \forall j \in J \cap (J - q), \qquad \forall q \in S_1 \\ G_h &= G_1 h^2 \end{aligned}$$

where v_1 and G_1 are strictly positive constants, then the potentials in the upper node layer take the values $v_1 z_0(j)$, $j \in J$. The Neumann condition is trivially implemented by leaving out the "loose" connections of the nodes next to the boundary.

The fact that our variational edge detection problem (5.10) is a time independent boundary value problem, as opposed to an initial value problem, makes this circuit realiza-

tion much more tractable, than those proposed for the solution of the unbiased anisotropic diffusion problem (5.17) [28].

1. No precision capacitors are necessary to model the time dependency of the diffusion. This will save a lot of silicon area, and make the performance less sensitive to imperfections in the manufacturing process, and completely insensitive to stray capacitances.

2. No simultaneous loading or latching of the voltages, representing the diffused image function is required. The image function I/O-processes can be asynchronous. Consequently precision timing is not an issue.

3. The settling time is determined only by the stray capacitances and therefore practically nil. The particularly simple representation of the original image function ζ_0 therefore allows the circuit to process sequences of images at a frequency, which is limited only by the I/O-capacity.

The most serious difficulty with implementing the circuit in figure 5.4 is probably the realization of the controlled conductances most of which depend on the potentials at six! of the surrounding nodes. This problem can be greatly simplified, by replacing (5.25a) and (5.25b) by a cruder approximation of $\|\nabla z((j+\frac{q}{2})h)^T\|$. If $z_{0,\pm 1}$, $z_{q_1,\pm 1}$, $z_{\pm 1,0}$ and $z_{\pm 1,q_2}$ are approximated by $z_{0,0}$, $z_{q_1,0}$, $z_{0,0}$ and z_{0,q_2} respectively, one obtains the much simpler expression

$$\sigma_q \doteq \frac{|z_q - z_0|}{h} \qquad q \in S_1 \tag{5.28}$$

The conductance $G(j+\frac{q}{2})$ then only depends on the absolute value of the voltage across it, and can thus be realized by a regular voltage-controlled nonlinear resistor with i-v characteristic

$$i = \Phi(v) \doteq G_1 v_1 h \ \mathrm{sgn}(v) \varphi\left(\frac{|v|}{v_1 h}\right)$$

where as before $\varphi(\gamma) \doteq g(\gamma)\gamma$. The validity of the approximation (5.28), which by the way also simplifies the diffusivity anomaly (5.14), has been tested experimentally by Perona and Malik [29], who introduced it for solving the unbiased anisotropic diffusion problem numerically, and obtained very satisfactory results.

5.8 Convergence

In this section we will study some rudimentary convergence properties of the Jacobi-like iteration method (5.27). For certain parameter values we manage to show, that this iteration converges to a limit point, which satisfies (5.26), depends continuously on the original image function ζ_0, and is independent of the initial image function χ_0. Besides convergence of the iteration we thus obtain both uniqueness and a sense of stability with respect to the initial data. This sounds too good to be true, and as a matter of fact the assertions are valid only for parameter values, far from those of major interest for edge detection purposes. Albeit this serious weakness our analytical results give some indication, that solutions exist, and that these solutions are reasonably well-behaved—a hypothesis further supported quite strongly by our experiments. They might also serve as a starting point for future theoretical development. One could possibly obtain better results, if one applied some more sophisticated iteration method. However this would most likely drastically compromise the simplicity of the algorithm. We have therefore confined our analysis to the Jacobi-like method, which after all yields remarkably satisfactory experimental results.

We begin our discussion with a couple of observations closely related to the extremum principles from section 5.5.

Proposition 5.8.1 *Let z_0 be a solution of the discretized boundary value problem (5.26). Then*

$$\bigwedge_{l\in J} \zeta_0(l) \le z_0(j) \le \bigvee_{l\in J} \zeta_0(l) \qquad \forall j \in J$$

Proof: Let $j_\pm \doteq \arg\max_{l\in J} \pm z_0(l)$. Then $\pm[z_q(j_\pm) - z_0(j_\pm)] \le 0$, $\forall q \in S$. Hence by (5.26)

$$\bigvee_{l\in J} \pm z_0(l) = \pm z_0(j_\pm) \le \pm\zeta_0(j_\pm) \le \bigvee_{l\in J} \pm\zeta_0(l)$$

■

Proposition 5.8.2 *Let $z_0^{(i)}$, $i \in \mathbb{N}_0$ be defined by the iteration scheme (5.27). Then*

$$\bigwedge_{l\in J} [\zeta_0(l) \wedge \chi_0(l)] \le z_0^{(i)}(j) \le \bigvee_{l\in J} [\zeta_0(l) \vee \chi_0(l)] \qquad \forall j \in J \qquad \forall i \in \mathbb{N}_0$$

Proof: Let $i \in \mathbb{N}_0$, and $j \in J$. From (5.27a) and (5.27b) we see that $z_0^{(i)}(j)$ is a convex combination of $\zeta_0(j)$ and $z_q^{(i-1)}(j)$, $q \in S_\rho$, and thus in the convex hull of $\{\zeta_0(l),\ z_0^{(i-1)}(l)$:

$l \in J\}$. Since this is true $\forall j \in J$, we have

$$\bigwedge_{l\in J}[\zeta_0(l) \wedge z_0^{(i-1)}(l)] \leq \bigwedge_{l\in J} z_0^{(i)}(l) \leq \bigvee_{l\in J} z_0^{(i)}(l) \leq \bigvee l \in J[\zeta_0(l) \vee z_0^{(i-1)}(l)]$$

The proposition then follows by induction. ■

Using the bounds provided by the proposition above we can show the following two convergence results:

Lemma 5.8.3 *Let $z_0^{(i)}$, $i \in \mathsf{N}_0$, be defined by the iteration scheme (5.27) in terms of an initial image function χ_0 and an original image function ζ_0. Let $y_0^{(i)}$, $i \in \mathsf{N}_0$, be defined in a completely analogous manner, but with χ_0 and ζ_0 replaced by ψ_0 and η_0 respectively. Assume that the dependency on the edge cost coefficient λ is reflected by the diffusivity anomaly g given by (5.12). If λ is sufficiently large, then*

$$\limsup_{i\to\infty} \|y_0^{(i)} - z_0^{(i)}\|_\infty \leq c\|\eta_0 - \zeta_0\|_\infty \qquad \textit{exponentially}$$

for some known finite constant c.

Proof: To be specific we will only prove the lemma for the diagonal approximation ($\rho^2 = 2$). The proof can however easily be reconstructed to cover the Cartesian case as well. For $i \in \mathsf{N}_0$, let $y_q^{(i)}$, $\tau_q^{(i)}$, $v_q^{(i)}$, $q \in S_2$ and $\overline{v}^{(i)}$ denote the functions associated with $y_0^{(i)}$ corresponding to $z_q^{(i)}$, $\sigma_q^{(i)}$, $w_q^{(i)}$, $q \in S_2$ and $\overline{w}^{(i)} \doteq \overline{w}_2^{(i)}$ respectively. For simpler notation also define the following bounds:

$$\begin{aligned}
D^{(i)} &\doteq \|y_0^{(i)} - z_0^{(i)}\|_\infty \qquad i \in \mathsf{N}_0 \\
E &\doteq \|\eta_0 - \zeta_0\|_\infty \\
R_z &\doteq \bigvee_{j\in J}[\zeta_0(j) \vee \chi_0(j)] - \bigwedge_{j\in J}[\zeta_0(j) \wedge \chi_0(j)] \\
R_y &\doteq \bigvee_{j\in J}[\eta_0(j) \vee \psi_0(j)] - \bigwedge_{j\in J}[\eta_0(j) \wedge \psi_0(j)] \\
R &\doteq R_z + (E + D^{(0)}) \\
M &\doteq \|\zeta_0\|_\infty \vee \|\chi_0\|_\infty
\end{aligned}$$

From the definitions of the shifted functions z_q, y_q, $q \in S$, and proposition 5.8.2 we further

note that

$$\begin{aligned} D^{(i)} &\geq \|y_q^{(i)} - z_q^{(i)}\|_\infty && \forall q \in S \quad \forall i \in \mathbb{N}_0 \\ R_z &\geq \bigvee_{j\in J} z_q^{(i)}(j) - \bigwedge_{j\in J} z_q^{(i)}(j) && \forall q \in S \quad \forall i \in \mathbb{N}_0 \\ R_y &\geq \bigvee_{j\in J} y_q^{(i)}(j) - \bigwedge_{j\in J} y_q^{(i)}(j) && \forall q \in S \quad \forall i \in \mathbb{N}_0 \\ R &\geq \frac{R_y + R_z}{2} \\ M &\geq \|z_q^{(i)}\|_\infty && \forall q \in S \quad \forall i \in \mathbb{N}_0 \end{aligned}$$

Dropping the dependence on $j \in J$ and $i \in \mathbb{N}_0$ for shorter notation, from (5.25c) we then have

$$\begin{aligned} &|\tau_q^2 - \sigma_q^2| = \\ &\quad = \frac{1}{2h^2}|(y_{q_1,q_2} - y_{0,0})^2 + (y_{q_1,0} - y_{0,q_2})^2 - (z_{q_1,q_2} - z_{0,0})^2 - (z_{q_1,0} - z_{0,q_2})^2| \\ &\quad = \frac{1}{2h^2}|(y_{q_1,q_2} - y_{0,0} + z_{q_1,q_2} - z_{0,0})(y_{q_1,q_2} - y_{0,0} - z_{q_1,q_2} + z_{0,0}) \\ &\qquad + (y_{q_1,0} - y_{0,q_2} + z_{q_1,0} - z_{0,q_2})(y_{q_1,0} - y_{0,q_2} - z_{q_1,0} + z_{0,q_2})| \\ &\quad \leq \frac{2(R_y + R_z)2D}{2h^2} \\ &\quad \leq \frac{4RD}{h^2} \qquad \forall j \in J \qquad \forall i \in \mathbb{N}_0 \end{aligned}$$

Thus from (5.12) we see that

$$\begin{aligned} &|v_q - w_q| = \\ &\quad = |g \circ \tau_q - g \circ \sigma_q| \\ &\quad = \frac{v_q w_q |\sigma_q^2 - \tau_q^2|}{\lambda} \\ &\quad \leq \frac{4 v_q w_q RD}{\lambda h^2} \qquad \forall j \in J \qquad \forall q \in S_2 \qquad \forall i \in \mathbb{N}_0 \end{aligned}$$

which in turn implies that

$$|\overline{v} - \overline{w}| \leq \sum_{q\in S_2} |v_q - w_q| \leq \sum_{q\in S_2} \frac{4 v_q w_q RD}{\lambda h^2} \leq \frac{4\overline{v}RD}{\lambda h^2} \qquad \forall j \in J \qquad \forall i \in \mathbb{N}_0$$

Hence

$$\begin{aligned} &|y_0 - z_0| = \\ &\quad = \left| \frac{1}{2h^2 + \overline{v}} \left(2h^2 \eta_0 + \sum_{q\in S_2} v_q y_q \right) - \frac{1}{2h^2 + \overline{w}} \left(2h^2 \zeta_0 + \sum_{q\in S_2} w_q z_q \right) \right| \end{aligned}$$

$$\leq \left|\frac{1}{2h^2+\overline{v}} - \frac{1}{2h^2+\overline{w}}\right| \left|2h^2\zeta_0 + \sum_{q\in S_2} w_q z_q\right|$$
$$+ \frac{1}{2h^2+\overline{v}} \left|2h^2(\eta_0-\zeta_0) + \sum_{q\in S_2}(v_q y_q - w_q z_q)\right|$$
$$\leq \frac{|\overline{w}-\overline{v}|}{2h^2+\overline{v}} M + \frac{1}{2h^2+\overline{v}}(2h^2 + E + |\overline{v}-\overline{w}|M + \overline{v}D)$$
$$\leq \frac{\overline{v}}{2h^2+\overline{v}}\left(\frac{8RM}{\lambda h^2}+1\right) D + E \qquad \forall j \in J \qquad \forall i \in \mathbb{N}_0$$

Since $\overline{v}^{(i)}(j) \leq 4$, $\forall j \in J$, $\forall i \in \mathbb{N}_0$, we therefore conclude that

$$D^{(i+1)} \leq \frac{1}{1+\frac{h^2}{2}}\left(\frac{8RM}{\lambda h^2}+1\right) D^{(i)} + E \qquad \forall i \in \mathbb{N}_0 \tag{5.29}$$

If $\lambda > 16RM/h^4$, the assertion of the lemma then follows. ■

Theorem 5.8.4 *For sufficiently large values of the edge cost coefficient λ the discretized variational edge detection problem (5.26) has a unique solution, which is L_∞-norm-stable with respect to the initial data, and to which the iteration method (5.27) converges independently of its initial state χ_0.*

Proof: Assume that λ is large enough for lemma 5.8.3 to be conclusive. Let $z_0^{(i)}$, $y_0^{(i)}$, $i \in \mathbb{N}_0$, be given as in lemma 5.8.3 with $\psi_0 = z_0^{(1)}$ and $\eta_0 = \zeta_0$. Then $y_0^{(i)} = z_0^{(i+1)}$, $\forall i \in \mathbb{N}_0$, and $E = 0$. By lemma 5.8.3, a simple Cauchy sequence argument (in $L_\infty(J)$) and the observation, that the left hand side of (5.26) is a continuous function of z_0 with respect to the L_∞-topology, (z_q is a continuous function of z_0, $\forall q \in S$,) it then follows, that $z_0^{(i)}$ converges to a *solution* of (5.26) (as $i \to \infty$) independently of its initial value χ_0. Next let χ_0 and ψ_0 be two possibly different solutions of (5.26), and let $z_0^{(i)}$, $y_0^{(i)}$, $i \in \mathbb{N}_0$, be given as in lemma 5.8.3 with $\eta_0 = \zeta_0$. Then $z_0^{(i)} = \chi_0$, $y_0^{(i)} = \psi_0$, $\forall i \in \mathbb{N}_0$, and $E = 0$. Thus by lemma 5.8.3

$$\|\psi_0 - \chi_0\|_\infty = \lim_{i\to\infty} \|y_0^{(i)} - z_0^{(i)}\|_\infty = 0$$

which shows, that the solution of (5.26) is unique. Finally let χ_0 and ψ_0 be the solutions of (5.26) given, that the corresponding original image functions are ζ_0 and η_0 respectively, and let $z_0^{(i)}$, $y_0^{(i)}$, $i \in \mathbb{N}_0$, be given as before. Again $z_0^{(i)} = \chi_0$, $y_0^{(i)} = \psi_0$, $\forall i \in \mathbb{N}_0$. From proposition 5.8.1 and lemma 5.8.3 it thus follows that

$$\|\psi_0 - \chi_0\|_\infty = \lim_{i\to\infty} \|y_0^{(i)} - z_0^{(i)}\|_\infty \leq \|\eta_0 - \zeta_0\|_\infty$$

which proves the stability of the solution. ■

Unfortunately the theorem above gives a too pessimistic view of, what our experiments undoubtedly confirm, is really going on; it is only conclusive for values of the edge cost coefficient λ, far greater, than those, for which the algorithm is useful for edge detection. There are two reasons for this shortcoming.

First of all at most locations $j \in J$ the constant R in the proof of lemma 5.8.3 is an overly conservative bound for the *local* differences of $z_0^{(i)}$, it is meant to estimate. If the iteration scheme was linear, this problem could easily be remedied, by replacing the L_∞-norm in the convergence analysis by a Sobolev norm, which incorporates the evolution of these local differences as well as that of $z_0^{(i)}$ itself. However, as we discussed in section 5.6, the nonlinearity, inherited from the boundary value problem (5.10), is by our choice such, that the local differences are strengthened, wherever initially sufficiently pronounced. Local differences of magnitude of the order R are therefore eventually to be expected. The intuitive reason for the success of the scheme lies in the earlier demonstrated fact, that the strengthened edges are simultaneously sharpened, so that the set of slow convergence shrinks during the iteration—a mechanism that is not captured by the L_∞-style of the proof above. Since the nonlinearity prohibits Fourier techniques, this problem might be hard to fix.

Secondly the theorem suggests, that λ be chosen proportional to h^4. In contrast, (as one would also guess from, the way λ enters the defining expressions of the diffusivity anomaly,) our experiments indicate, that λ be chosen proportional to h^2, as if the unity term inside the parenthesis in (5.29) was missing. The intuitive reason for this discrepancy has to do with another case of competing processes. A closer examination of the proof above shows, that the source of this term is the unit bound on the continuity control function $w_q^{(i)}$, inherited from the properties of the diffusivity anomaly g. Since $w_q^{(i)}$ actually takes values close to unity at the abundant locations of almost vanishing image function gradient, this bound is tight. However rewriting (5.27b) as

$$z_0^{(i+1)} - z_0^{(i)} = \frac{1}{1+\overline{w_2}^{(i)}}\left[2h^2(\zeta_0 - z_0^{(i)}) + \sum_{q\in S_2} w_q^{(i)}(z_q^{(i)} - z_0^{(i)})\right]$$

we see, that at such locations

$$z_0^{(i+1)} - z_0^{(i)} \approx \frac{2h^2}{5}(\zeta_0 - z_0^{(i)})$$

Thus it seems, like the large values of $w_q^{(i)}$ destroy the exponential convergence rate locally, and only at those locations $j \in J$, where $z_0^{(i)}(j)$ has already practically converged.

5.9 Experimental Results

In this section we present some experimental results regarding our variational edge detection method, governed by the cost functional (5.2). In all the experiments the edge cost density was given by $f(\omega) \doteq \omega - \ln\omega$, corresponding to the diffusivity anomaly $g(\gamma) \doteq 1/(1+\gamma^2/\lambda)$. The images involved were obtained by solving the diagonal ($\rho^2 = 2$) discrete approximation (5.26) of the boundary value problem (5.10). For computational simplicity we used a Gauss-Seidel-like iteration method, rather than the Jacobi-like scheme (5.27).

As mentioned earlier, the iteration method converges to the solution of interest. In general, as one should expect, the convergence is faster, if the initial image function χ_0 is set equal to the original image function ζ_0. The sequence of images in figure 5.5 illustrates this condition. It shows, that reasonably good results are obtained well before 50 iterations, and that convergence in the "sense of insignificant perceptible changes" is reached after about 100 iterations. These observations, are as far as we can tell from our experiments, valid, whenever $\chi_0 = \zeta_0$. In particular, they seem to hold independently of the choice of the edge cost coefficient λ and the pixel width h, at least in the range of interest for edge detection.

The variational edge detection method itself as well as the iteration method, we employed to solve it, appear to be remarkably robust with respect to changes in the initial image function. Indeed if $\chi_0 \neq \zeta_0$, the iteration method still converges, if yet at a slower rate. To demonstrate this behavior, we tried the algorithm on the same original image function, as in figure 5.5, but with the particularly unfavorable initial image function $\chi_0 = 0$. Some samples from the resulting sequence of images are shown in figure 5.6. The fact that the limit image functions in the figures 5.5 and 5.6 are perceptually so close, also indicates, that the solutions, even though multiple, in large exhibit the desirable type of behavior, that mathematically stringent uniqueness would warrant. As one should expect, the significant differences seem to be limited to affect small blobs of high contrast relative to the local background. It is interesting to note, that the little dark blobs in the center of figure 5.5 (f), which are missing in figure 5.6 (f), represent pixel values, which are closer to the zero initial

Fig. 5.5: (a)

Fig. 5.5: (b)

Fig. 5.5: (c)

Fig. 5.5: (d)

Fig. 5.5: (e)

Fig. 5.5: (f)

Figure 5.5: Estimated image after i iterations, when $\chi_0 = \zeta_0$. (a) $i = 0$ (original image). (b) $i = 25$. (c) $i = 50$. (d) $i = 100$. (e) $i = 200$. (f) $i = 800$.

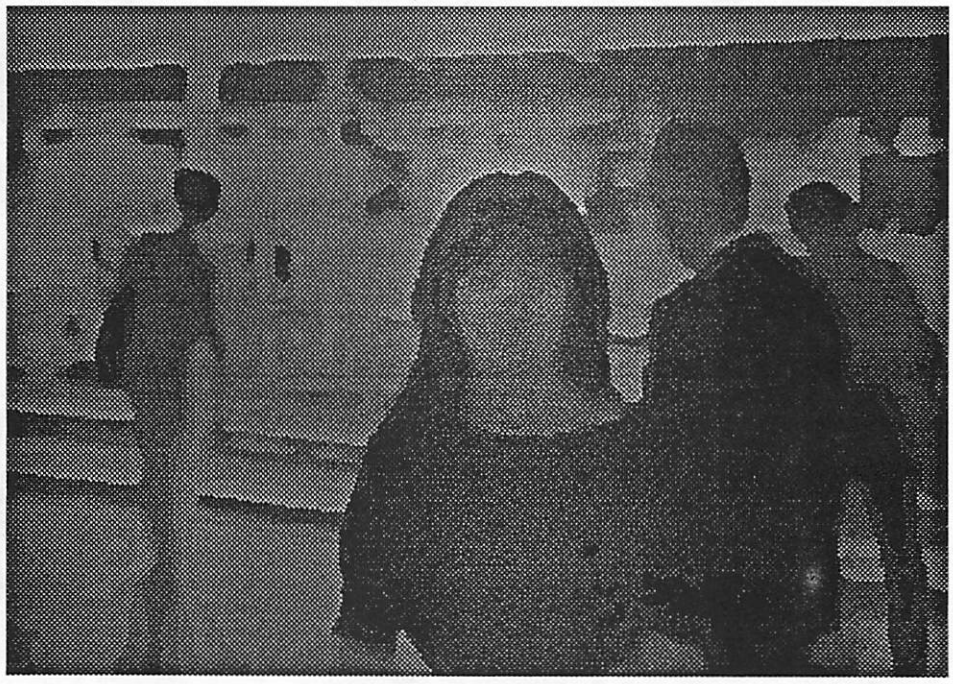


Fig. 5.6: (a)

Fig. 5.6: (b)

Fig. 5.6: (c)

Fig. 5.6: (d)

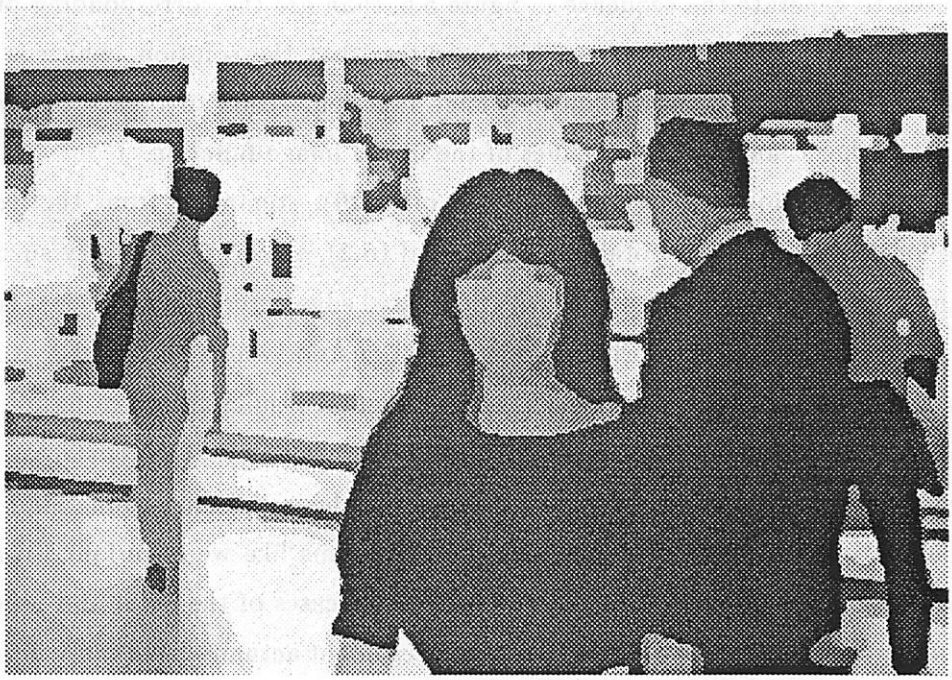


Fig. 5.6: (e)

Fig. 5.6: (f)

Figure 5.6: Estimated image after i iterations, when $\chi_0 = 0$. (a) $i = 25$. (b) $i = 50$. (c) $i = 100$. (d) $i = 200$. (e) $i = 400$. (f) $i = 800$.

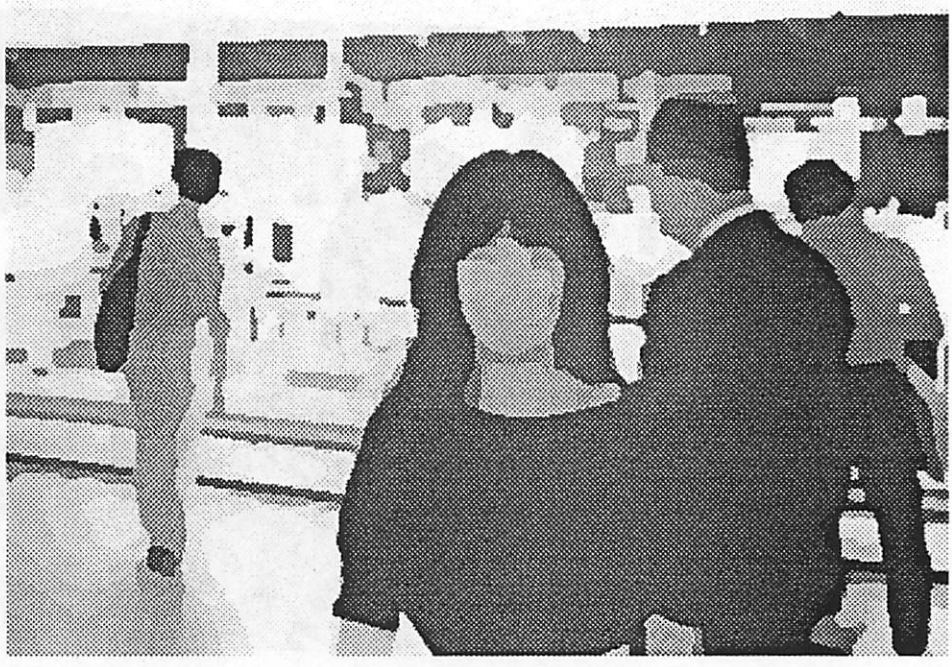
image, used to generate the sequence in figure 5.6, than are the corresponding pixel values (of the nonblobs) in figure 5.6 (f). This indicates, that the solution, which is implicitly selected by choosing a particular initial image function, tends to reflect the smoothness properties, rather than the actual values of the initial image function.

The nonuniqueness of the solutions of (5.26), stemming from the existence of multiple local minima of the total cost functional (5.2), should not be very surprising. In fact for most of the other existing regularization based edge detection methods it is relatively easy to construct examples of original image functions, for which the total cost functional exhibits this behavior. It is clear from the experimental results shown in figure 5.6, if not by intuition, that all the local minima of the total cost are potentially satisfactory solutions to the edge detection problem. Moreover by choosing the initial image function χ_0 to equal either the original image function or a constant, it seems like we have found a method of selecting those local minima, which correspond to the cases of the most and least detailed estimated images respectively. These extreme cases might actually be of more interest, than the solution corresponding to the global minimum.

For our observations regarding the parameter dependence of the solution, that is the influence of the edge cost coefficient λ and the pixel width h on the estimated image function z, we recall, that $r \doteq 1/h$ is a true scale-space parameter governing the spatial resolution of the edge detector, and that $\sqrt{\lambda}$, proportional to the edge enhancement threshold γ_0, controls its sensitivity in a linear fashion. Since the local differences of the (original) image function, unlike the discrete approximations of its derivatives, remain invariant under scale-space variations in terms of h, a more meaningful sensitivity parameter is in this context given by $s \doteq \sqrt{\lambda}h$, which is proportional to the corresponding local difference enhancement threshold $\gamma_0 h$. (The same conclusion would have been obtained, had we kept h constant and instead incorporated the explicit scale-space parameter μ in the total cost functional, as discussed in section 5.3.) Figure 5.7 shows an example of how the estimated image function (after 100 iterations) depends on the scale-space parameter r for a fixed sensitivity parameter ($s = \sqrt{20}$). Its dependence on the sensitivity parameter s for a fixed scale-space parameter ($r = \sqrt{50},$) is illustrated in figure 5.8.

In order to extract a set of edges from the estimated image function z, we followed the strategy outlined in section 5.3, and simply thresholded the gradient magnitude. Figure 5.9 shows the edges extracted from the estimated image function in figure 5.7 (b) using two different thresholds, one lower than, and the other one equally much higher than

(a)



(b)

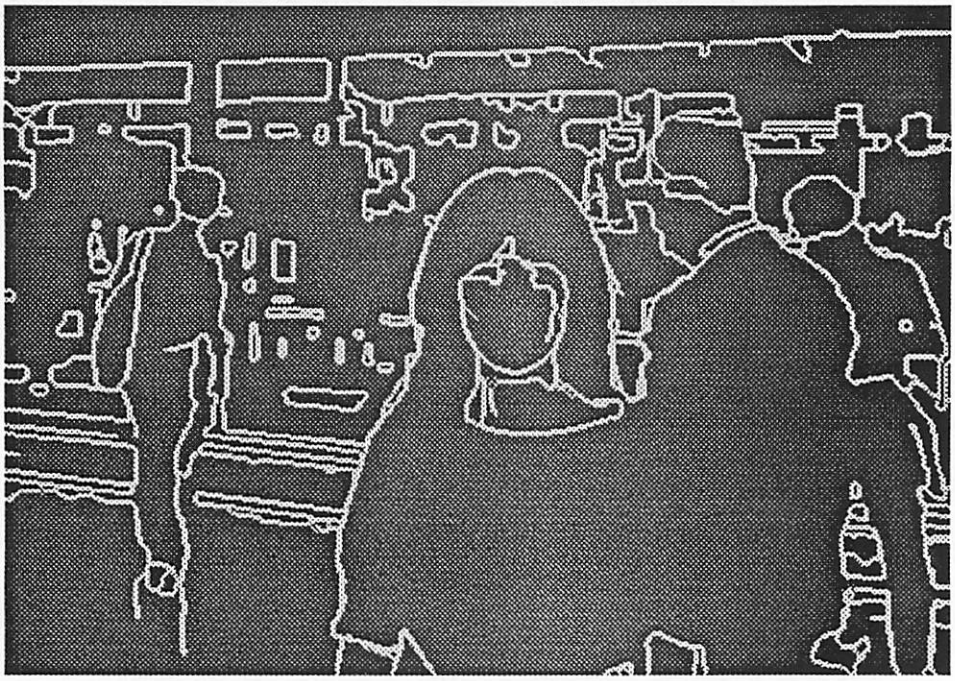
Figure 5.7: Estimated images for different values of the scale-space parameter r. (a) $r^2 = 12.5$. (b) $r^2 = 100$.

(a)

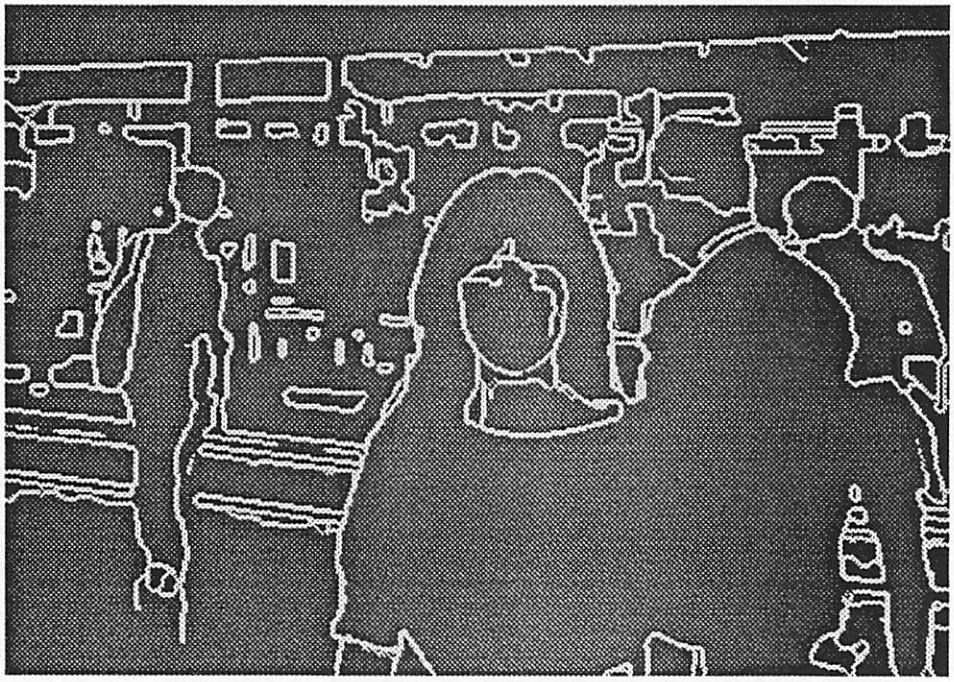
(b)

Figure 5.8: Estimated images for different values of the sensitivity parameter s. (a) $s^2 = 10$. (b) $s^2 = 40$.

the edge enhancement threshold γ_0. As predicted by the discussion in section 5.6, the experiments confirm, that the edge extraction is very robust with respect to changes in the threshold ϑ for a wide range of thresholds around γ_0. In fact, if one allows a couple of edge segments to change, the range in question in this case extends well beyond, that spanned by the three examples in the figure.



(a)



(b)

Figure 5.9: Extracted edges for different values of the threshold $\vartheta \doteq g^{-1}(\theta)$. (a) $\vartheta = 22$. (b) $\vartheta = 67$.

Chapter 6

Conclusions

At the time our efforts in edge detection began, almost all existing edge detection methods were based on the use of local operators (for deciding whether or not a given pixel belonged to an edge). Most such local methods yield edges that exhibit systematic errors, and all of them lack a mathematical problem formulation that relates the *entire* output of edges to the *entire* input—the original image function. These shortcomings motivate a global approach.

We have proposed and studied the properties of two paradigms for global edge detection. We have also developed general methods and more specific algorithms for solving the resulting computational problems. These algorithms were implemented in software and their performances demonstrated and evaluated through a number of experiments.

Both our paradigms are based on variational regularization, and expressed mathematically as the minimization of a cost functional depending on the edges as well as a piecewise smooth estimate of the true image function. The total cost consists in both cases of the sum of three separate subcosts—an edge cost penalizing the extent of the edges, a deviation cost promoting close approximation of the true image function by the estimated image function and a stabilizing cost favoring a smooth estimated image function.

The most essential difference between the two approaches lies in the representation of the edges. In the first paradigm the edges are represented by parametrized curves in $\mathbf{R}^2$. In the second paradigm, which led us to the biased anisotropic diffusion method, they are represented by a (strictly positive) real valued continuity control function defined on the entire image domain. Both approaches have their merits. While the global curve-represented edge detection method yields a more structured—higher level—description of the edges, the

biased anisotropic diffusion method is easier to implement and requires less computation. The edges obtained with the biased anisotropic diffusion method are in general also better looking than those obtained with our global curve-represented edge detector. It is important to note, however, that "better looking" is not necessarily better, unless the detected edges are subject to human inspection, (for example by the editor of a prominent journal.) A higher level description might be more valuable for further machine processing. It should also be noted that the global curve-represented edge detector most likely is open for more improvement than the biased anisotropic diffusion method, which does not require that initial edges are found externally.

Whether or not the proposed global edge detection methods are going to be useful in the future is not clear at this point. With present technology their relatively high computational cost is a serious disadvantage in comparison with the much faster local methods. However, highly specialized analog VLSI chips dedicated to solving partial differential equations similar to the Euler equations associated with the kind of cost functionals that our paradigms involve, are being developed in research laboratories. If this effort succeeds, global edge detection methods will become more tractable. Another possible avenue for speeding up the global edge detection methods would of course be to develop or possibly employ existing algorithms for solving the Euler equations with parallel processing.

6.1 Curve-Represented Edge Detection

In the curve-represented edge detection paradigm the edges are represented by parametrized curves. We have considered quite general spaces of such curves—restricted only by continuity, smoothness and regularity constraints. We have also considered the special case of uniform cubic B-spline curves in detail. The spline curves offer simpler—lower dimensional—parametrizations, which are particularly convenient for implementations. The curve-represented edge detection paradigm is modular in the sense that it includes a number of different edge and stabilizing costs, from which different linear combinations can be selected to form a variety of total cost functionals together with the deviation cost. Most of the edge costs apply to (edges represented by) general smooth regular parametrized curves. For edges represented by splines we have also proposed a couple of additional edge costs defined directly in terms of the associated control polygons. These edge costs are most often simpler to deal with than those that apply to the wider class of edges represented by

general parametrized curves. The different stabilizers essentially enforce different degrees of (piecewise) smoothness of the estimated image function.

In order to solve the global curve-represented edge detection problem, the total cost functional must be minimized. Following one of the standard approaches of calculus of variations we have somewhat heuristically derived a number of optimality conditions for the image segmentation, that is the edges, as well as the estimated image function. The optimality conditions regarding the edges are unfortunately but not surprisingly practically useless for solving the edge detection problem. In other words, there is no known method available for finding an image segmentation that satisfies these conditions. The edge optimality conditions do, however, bring us some valuable insight about the properties of any such optimal image segmentation. With the optimality conditions regarding the estimated image function, on the other hand, the situation is more or less reversed. These conditions dictate that the optimal estimated image function satisfies an (elliptic linear partial differential) Euler equation, which can be shown to have a unique solution. While this fact hardly provides much insight about the optimal estimated image function, (which is not of primary concern anyway,) it does support a computationally straight forward method for finding the unique optimal estimated image function, which minimizes the total cost for a given, not necessarily optimal, image segmentation.

Because of the heuristics involved and the complicated way in which the optimal edge conditions depend on the (optimal) estimated image function, the variational approach does not suffice to settle the issue of whether there exist an optimal image segmentation and an optimal estimated image function, which minimize the total cost functional. In order to answer that question we have taken a somewhat different route employing techniques and results from modern functional analysis. Our answer is affirmative, if yet valid only for a somewhat restricted space of image segmentations. Our analysis is moreover limited to a particularly simple choice of stabilizer. The chosen stabilizer is (just because of its simplicity), however, the most interesting one for processing of image functions representing brightness data.

Based on the heuristically derived variations of the total cost with respect to the edges and the optimality conditions regarding the estimated image function we have proposed a method for solving the global curve-represented edge detection problem. More precisely, the method finds an image segmentation and an estimated image function at which the total cost functional has a local minimum. It starts out with some initial guess

of what the optimal edges should be, and proceeds by adjusting these edges according to a steepest descent rule. In order to compute the appropriate adjustment of the edges, the optimal estimated image function for the present edges must be known. Each step of the steepest descent procedure therefore requires the solution of the Euler equation governing this optimal estimated image function. This can be done by way of standard numerical methods, for example Gauss-Seidel. Since the Euler equation is quite well-behaved, convergence to a unique solution can be guaranteed. The solution of the Euler equation does, however, require considerable computational resources, and is therefore—at least with present technology—the bottleneck of the entire edge detection method.

The global curve-represented edge detection method is most conveniently implemented for edges represented by splines. For such edges and a specific relatively simple total cost functional we have developed a detailed global edge detection algorithm. This algorithm has been implemented in software and subject to substantial experimentation. Our experimental results verify that the algorithm essentially operates as intended. They also provide examples of how the detected edges and the estimated image function depend on the edge and stabilizing cost coefficients as well as the control vertex density of the splines representing the edge segments.

6.1.1 Future Work

There are three more or less obvious major directions for further development of the global curve-represented edge detection method and the theory underlying it. First of all, the variational calculus, upon which the method is based, ought to be straightened out. Secondly, the existence theorem 3.10.3 can almost certainly be generalized to apply to a wider space of image segmentations than those satisfying the image segmentation constraints (I1), (I2), (R), (B1), (B2), (E1) and (E2) in section 3.11, or to total cost functionals with more general stabilizers. Finally, there are many algorithm and implementation alternatives that remain to be investigated.

The variational calculus in chapter 2 suffers from two basic weaknesses; it does not consider second variations, and it is heuristic. By including second variations in the analysis, one might be able to rule out certain spurious "solutions", that is saddles and local maxima, which satisfy the optimality conditions without corresponding to local minima of the total cost functional. A strict mathematical development of the variational calculus

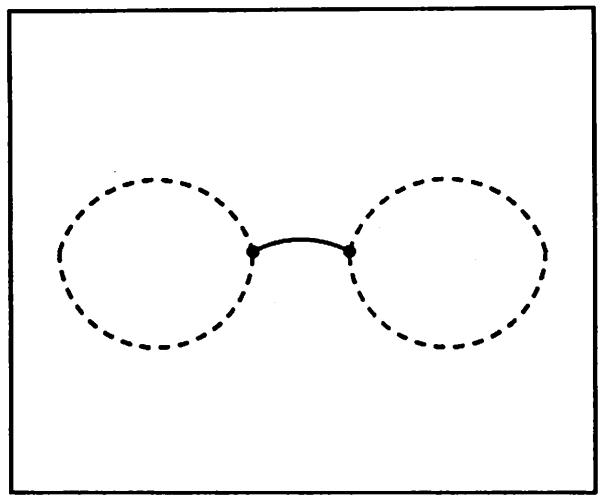


Figure 6.1: Edge segment (solid) with the same component of the continuity set on both sides.

would most likely necessitate mathematical sophistication and techniques comparable to those in chapter 3. If successful, such a development could support the relaxation of the intersection constraints (B1), (B2), (E1) and (E2) on the space of image segmentations to which the existence theorem 3.10.3 applies. It could also result in a full understanding of the theoretical problems with the free endpoints. This might in turn lead to a more general set of optimality conditions, which permit the existence of such endpoints, and to a theoretically satisfactory way of treating such endpoints in the computational methods for solving the global curve-represented edge detection problem.

We have just commented on one possible generalization of theorem 3.10.3, and that one is probably the easiest such generalization to achieve. Besides a relaxation of the intersection constraints it seems also quite possible to relax the interconnection constraints in section 3.11 a bit. A reasonable first attempt would be to replace the constraint (I2) by one that only prohibits free endpoints. This would for example allow edge segments that have the same component of the continuity set on both sides. An example of such an edge segment is depicted in figure 6.1. The ultimate goal would of course be to remove the constraint (I2) altogether, thus allowing edge segments with free endpoints. Whether this is to ambitious, and under which additional conditions, if any, this might be possible, is hard to say. Finally, the application domain of theorem 3.10.3 could be extended by generalizing the proofs in chapter 3 so as to incorporate cost functionals with stabilizers involving second

and higher order partial derivatives of the estimated image function. With the techniques that we have used, however, this would most likely require a restriction of the space of image segmentations so that the boundaries of the components of the continuity set are guaranteed to be smooth.

During the implementation of the global curve-represented edge detector described in section 4.2 lots of viable design alternatives were necessarily left unexplored. There are thus many interesting changes and possibilities for improvement of the implemented algorithm that are worth further investigation. One could for example develop and implement a version of the algorithm that enforces all the image segmentation constraints in section 3.11. One could also experiment with algorithms for edges represented by general parametrized curves rather than splines and for cost functionals involving different edge and stabilizing costs. In terms of tuning the existing algorithm for better performance it would probably make most sense to upgrade the preliminary edge detector described in section C.1. The first suggestion would be to improve its jump mechanism. It could for example be made to adapt according to some local property of the original image function. The edge initiation and termination thresholds t_i and t_t could also be made adaptive. The preliminary edge detector could furthermore be replaced altogether or preceded by some other "continuity control function-represented" edge detector. Two choices worth trying would be the Canny edge detector [4] and the biased anisotropic diffusion method described in chapter 5. Ultimately it would be desirable to furnish the global curve-represented edge detector with an outer loop that allows for deletion of existing edge segments, introduction of new edge segments, and other changes of the image segmentation configuration while the steepest descent edge adjustment procedure is in progress. The initial edge finder and hence the preliminary edge detector, which is part thereof, would then play a less significant role and thus not require much improvement. Finally, there is an obvious need for automating the selection of the edge and stabilizing cost coefficients λ and μ—the two fundamental parameters. Since the edge adjustment procedure, in which these parameters are active, operates on the output from the initial edge finder, such an attempt should also involve the related parameters of the preliminary edge detector.

6.2 Biased Anisotropic Diffusion

In our second paradigm the edges are represented by a continuity control function. Although the approach is based on variational regularization, the resulting method can also be viewed as a biased anisotropic diffusion method. This circumstance exemplifies the close connection between the regularization and diffusion approaches in early vision, and we hope that our analysis has shed some fruitful light on this interesting subject. Besides being of general interest, the coincidence of the two paradigms has also allowed us to analyze our variational edge detection method in the diffusion context. We have for example showed, that it shares the attractive edge enhancement property characteristic of the unbiased anisotropic diffusion method.

Unlike other existing regularization approaches to edge detection, our method is tailored to support calculus of variations, not only with respect to the estimated/reconstructed image function, but also with respect to the continuity control function representing the edges. This modification of the paradigm leads to substantial computational savings in comparison with many, if not all, of the other regularization methods, without impairing the performance. The sharpness of the edges, which is seemingly given up from the outset, is regained during the iteration by the edge enhancement mechanism. This was demonstrated by our theoretical analysis as well as by our experimental results.

The most notable difference between our method and other existing anisotropic diffusion methods is, that our method converges to a solution of interest. This fact removes the problem of deciding when to stop the diffusion process as well as that of actually stopping it. As our discussion has revealed, the removal of the latter of these two problems represents a major advantage for potential analog circuit implementations. The price, that one pays for this improvement, is that the estimated image functions for different values of the scale-space parameter no longer can be generated recursively.

For the solution of the variational edge detection problem we have proposed an algorithm as well as an analog circuit realization. For a practically limited range of parameter values the algorithm has further been found to be extremely well-behaved; it converges to a unique solution of the discretized problem, independently of the initial image function, that is the initial state of the iteration process. An important aspect, about the proposed circuit, is, that it does not require either capacitors or synchronous readout.

While our theoretical convergence analysis has some limitations, our experimental

results have clearly demonstrated that this method works very well for typical parameter values of interest for edge detection. The algorithm does indeed converge to a solution of interest, that is an estimated image function, which is remarkably robust with respect to the initial image function. Furthermore the edges, which are obtained by postprocessing the estimated image function with a rudimentary local edge detector—thresholding of the gradient—are insensitive to changes in the threshold—the goal of the regularization. In addition to the convergence and robustness issues our experiments have exhibited the dependence of the solution on the values of the scale-space and sensitivity parameters embedded in our paradigm.

6.2.1 Future Work

A number of theoretically relevant problems have been left open. The most important mathematical questions that need to be answered, are arguably whether or not a solution to the quasi-linear variational edge detection problem (5.10) exists, and under which conditions such a solution is unique. In the case of nonunique solutions it would moreover be most interesting to characterize the different solutions and somehow relate the properties of the solution to those of the corresponding initial image function χ in the biased anisotropic diffusion problem (5.16). The same problem is also relevant for the numerical methods that can be employed for solving the edge detection problem (5.10). Other important issues concerning the numerical methods are the traditional ones of consistency, convergence and stability.

Any result pertaining to the numerical method will of course depend on the particular method under consideration. Some methods might lend themselves to nice results while others do not. An effort along these lines will thus consist in part of finding the most tractable numerical method(s) to work with. The choice of numerical method does of course also involve the discretization of the image domain and the discrete approximations of the differential operators that figure in (5.10). For additional possibilities one can moreover experiment with different edge cost density functions.

Besides being of interest for further theoretical developments the choices of the numerical method and the edge cost density function also naturally affects the convergence rate. Since the edge detection problem (5.10) is nonlinear, it might be hard to obtain good theoretical bounds on this rate. The problem is, however, of immediate practical

importance. In order to improve the performance of the biased anisotropic diffusion method it might thus be worth while estimating the convergence rate empirically for a number of different numerical methods and edge cost density functions.

Appendix A

Proof of Theorem 3.8.2

The introduction of the concept of admissible image segments in section 3.8 made the proof of the central lemma 3.8.3 relatively simple. To show directly from definition 3.8.1, however, that a given subset of an image domain B is an admissible image segment of B, is far from trivial. For verification of the hypotheses of lemma 3.8.3 we therefore rely on theorem 3.8.2, which states that every Lipschitz domain $\Omega \subseteq B$ is an admissible image segment of B. In this appendix we prove this important theorem.

A.1 The Original Atlas

Let $\Omega \subseteq \mathbf{R}^2$ be a (bounded open) Lipschitz domain. Thence according to definition 3.5.1 there exist a finite collection $\{T_m\}_{m=1}^M$ of coordinate transformations, a collection $\{\phi_m : \Delta_m \doteq]a_m, b_m[\to \mathbf{R}\}_{m=1}^M$ of corresponding Lipschitz continuous functions, and a number $d > 0$ such that the maps

$$\begin{aligned} \Phi_m \;&:\; Q_m \doteq \Delta_m \times]-d, d[\to U_m \doteq \Phi_m(Q_m) \\ &:\; x \mapsto T_m(x_1, \phi_m(x_1) + x_2) \qquad m = 1, \ldots, M \end{aligned}$$

satisfy the conditions:

(i) $U_{m+} \doteq \Phi_m(Q_{m+}) \subseteq \Omega \qquad m = 1, \ldots, M$

(ii) $U_{m-} \doteq \Phi_m(Q_{m-}) \subseteq \complement\overline{\Omega} \qquad m = 1, \ldots, M$

(iii) $\bigcup_{m=1}^{M} U_{m0} \doteq \bigcup_{m=1}^{M} \Phi_m(Q_{m0}) = \partial\Omega$

where $Q_{m\pm} \doteq \mathbf{R}^2_{\pm} \cap Q_m$, $Q_{m0} \doteq \mathbf{R}^2_0 \cap Q_m$ and $U_{m0} \doteq \Phi_m(Q_{m0})$, $m = 1, \ldots, M$. Without further assumptions the original atlas $\{\Phi_m\}_{m=1}^M$ may be very complicated and inconvenient to work with. (This circumstance is of course the price one pays for the ease with which the Lipschitz property can be verified.) However, given any original atlas of the form above it is fortunately possible to generate a new atlas that satisfies a few additional conditions. Some crucial simplifying assumptions regarding the original atlas can thereby be justified as outlined below.

We will first without loss of generality assume that the intervals $\Delta_1, \ldots, \Delta_M$ have been foreshortened as necessary so that none of the Lipschitz charts $\Phi_1, \ldots, \Phi_M$ is redundant, or more precisely

$$U_{m0} \setminus \bigcup_{\substack{p=1 \\ p \neq m}}^{M} \overline{U_{p0}} \neq \emptyset \qquad m = 1, \ldots, M \tag{A.1}$$

Consequently each *boundary segment* U_{m0}, $m = 1, \ldots, M$, intersects exactly two other (by the Lipschitz condition necessarily distinct) boundary segments in $\{U_{p0}\}_{p=1}^M$. The indices of these boundary segments, each of which contains exactly one of the endpoints of U_{m0}, will be labeled l_m and r_m (for left and right respectively). Thus

$$\left.\begin{array}{l} \lim\limits_{x_1 \downarrow a_m} \Phi_m(x_1, 0) \in U_{l_m 0} \\ \lim\limits_{x_1 \uparrow b_m} \Phi_m(x_1, 0) \in U_{r_m 0} \end{array}\right\} \qquad m = 1, \ldots, M$$

The geometrical relationships between the boundary segments U_{m0}, $U_{l_m 0}$ and $U_{r_m 0}$ and their associated local coordinate systems, (which for each $x \in \mathbf{R}^2$ indicate the values of $T_m^{-1}(x)$, $T_{l_m}^{-1}(x)$ and $T_{r_m}^{-1}(x)$ respectively,) are shown in figure A.1. Since $r_{l_m} = l_{r_m} = m$, we also have

$$\left.\begin{array}{l} \lim\limits_{x_1 \uparrow b_{l_m}} \Phi_{l_m}(x_1, 0) \in U_{m0} \\ \lim\limits_{x_1 \downarrow a_{r_m}} \Phi_{r_m}(x_1, 0) \in U_{m0} \end{array}\right\} \qquad m = 1, \ldots, M$$

The x_{m1}-coordinates* of these two points will as indicated in the figure be labeled $\tilde{a}_m$ and $\tilde{b}_m$ respectively.

Obviously $\tilde{a}_m, \tilde{b}_m \in \Delta_m$. From (A.1) it also follows that $\tilde{b}_m > \tilde{a}_m$. Hence the closures of any pair of disjoint boundary segments in $\{U_{m0}\}_{m=1}^M$ are also disjoint. In other

*The x_{m1}- and x_{m2}-coordinates of any point $x \in \mathbf{R}^2$ are given by $[1 \quad 0]^T T_m^{-1}(x)$ and $[0 \quad 1]^T T_m^{-1}(x)$ respectively.

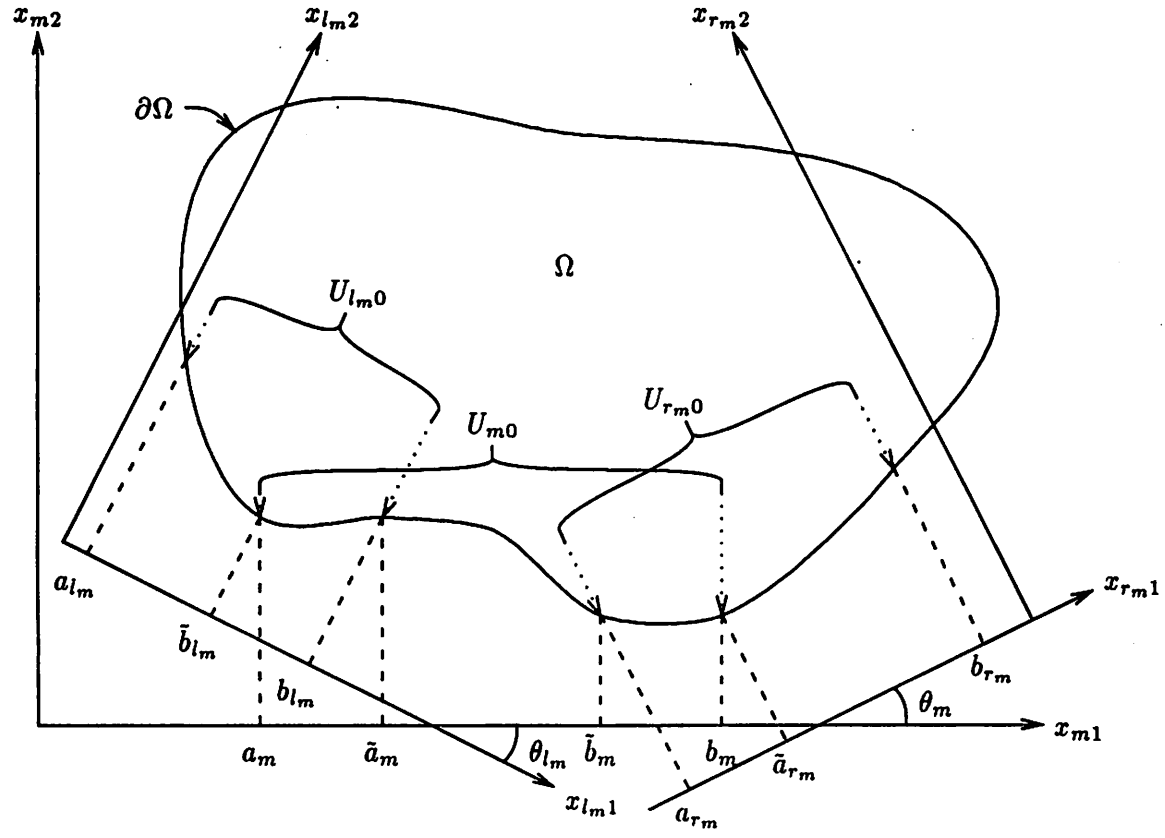


Figure A.1: Boundary segments U_m, U_{l_m}, U_{r_m} and their associated local coordinate systems.

words,

$$\overline{U_{m0}} \cap \overline{U_{p0}} = \emptyset \qquad p \in \{1,\ldots,M\} \setminus \{m, l_m, r_m\} \qquad m = 1,\ldots,M$$

Since the sets $\overline{U_{1,0}},\ldots,\overline{U_{M,0}}$ are compact, the minimum distance between any two disjoint such sets is strictly positive. We will therefore without loss of generality assume that the number $d > 0$ above is chosen sufficiently small that

$$U_m \cap U_p = \emptyset \qquad p \in \{1,\ldots,M\} \setminus \{m, l_m, r_m\} \qquad m = 1,\ldots,M \tag{A.2}$$

Finally we will make an assumption about the relative orientation of the local coordinate systems associated with overlapping boundary segments. As indicated in figure A.1, for each $m = 1,\ldots,M$, we let θ_m denote the orientation angle of the $x_{r_m 1}$-axis relative to the x_{m1}-axis. Let furthermore L_m denote the Lipschitz constant of the function ϕ_m. The maximum absolute angle that the tangent of the graph Γ_{ϕ_m} of ϕ_m forms with the x_{m1}-axis is then bounded by the constant

$$\eta_m \doteq \arctan L_m < \frac{\pi}{2} \tag{A.3}$$

By inserting an additional Lipschitz chart Φ_{M+1} with range inside the open set $U_m \cap U_{r_m}$, and thereafter again reducing the number d and foreshortening the domains Δ_m and Δ_{r_m} as necessary, one can always obtain a new atlas $\{\Phi_m\}_{m=1}^{M+1}$ of Lipschitz charts with the same properties as the old collection, but with the *original* angle θ_m replaced by two *new* angles θ_m and θ_{M+1} both half the size of the original θ_m. The Lipschitz constant L_{M+1} of the new function $\phi_{M+1} : \Delta_{M+1} \to \mathbf{R}$ introduced by this process can moreover easily be seen to satisfy the relation $L_{M+1} \leq L_m \vee L_{r_m}$. By repeating this insertion procedure sufficiently many times one can thus reduce the maximum absolute relative orientation angle between the local coordinate systems associated with overlapping boundary segments to an arbitrarily small value, without increasing the maximum value of the Lipschitz constants. We will therefore without loss of generality assume that

$$|\theta_m| + (\eta_m \vee \eta_{r_m}) < \frac{\pi}{2} \qquad m = 1,\ldots,M \tag{A.4}$$

In this section we have carefully used the subscript m to indicate that the indices l_m and r_m are those of the unique boundary segments in $\{U_{p0}\}_{p=1}^{M}$ that contain the left and right endpoints of U_{m0}. In the interest of avoiding too many subscripts we will henceforth most often let this dependence be implicit, and simply denote these indices by l and r respectively.

A.2 A Family of Modified Atlases

Equipped with the preliminaries from the previous section we are now ready to build the tools for generating a collection $\mathcal{F}_\Omega$ of interior set approximations of the domain Ω with the properties specified in definition 3.8.1. The idea is to let $\mathcal{F}_\Omega$ be an infinite subset of a collection $\{F_h\}_{h\in]0,H]}$ where F_h is obtained from Ω by basically shifting its boundary $\partial\Omega$ inwards a distance that tends to zero as $h \downarrow 0$. Technically this is done by for each $m = 1,\ldots,M$, introducing a set $\{\phi_{mh} : \Delta_m \to \mathbb{R}\}_{h\in]0,H]}$ of Lipschitz continuous functions with the properties that $\phi_{mh} \geq \phi_m$ and $\phi_{mh} \downarrow \phi$ uniformly (on Δ_m) as $h \downarrow 0$. The next step is to let the collection $\{\{\phi_{mh}\}_{m=1}^M\}_{h\in]0,H]}$ somehow induce a collection of modified Lipschitz chart atlases, which can be made to satisfy the conditions (i)–(iii) of definition 3.5.1 for the sets in $\mathcal{F}_\Omega$.

Since we must show that the collection $\mathcal{F}_\Omega$ satisfies condition (ii) of definition 3.8.1, it is desirable that the induced Lipschitz charts share the ranges of those in the original atlas, and that they and their inverses all satisfy a Lipschitz continuity condition with a uniform Lipschitz constant. These circumstances make the construction of $\{\{\phi_{mh}\}_{m=1}^M\}_{h\in]0,H]}$ and the induced Lipschitz charts somewhat more complicated than one at first might expect. First of all, the transformed graphs $T_m(\mathcal{F}_{\phi_m}$, $m = 1,\ldots,M$ must join so as to form simple closed curves. Secondly, the simple Lipschitz chart construction in definition 3.5.1 must be modified.

A.2.1 The Collection $\{\{\phi_{mh}\}_{m=1}^M\}_{h\in]0,H]}$

Consider the two boundary segments U_{m0} and U_{r0} and their associated local coordinate systems shown in figure A.2. Let for each $h > 0$, as there indicated $\tilde{b}_{mh}$ be the x_{m1}-coordinate of $\lim_{x_1\downarrow a_r} \Phi_r(x_1,h)$. From the figure we then see that

$$\tilde{b}_{mh} = \tilde{b}_m - h\sin\theta_m \to \tilde{b}_m \qquad \text{as } h \downarrow 0 \tag{A.5}$$

Since the interval Δ_m is open, this means that $\tilde{b}_{mh} \in \Delta_m$ for sufficiently small $h > 0$. For such values of h we define

$$\phi_{mh}(x_1) \doteq \phi_m(x_1) + h + \frac{[\tilde{c}_{mh} - \phi_m(\tilde{b}_{mh}) - h](x_1 - a_m)}{\tilde{b}_{mh} - a_m} \qquad x_1 \in]a_m, \tilde{b}_{mh}] \tag{A.6}$$

where $\tilde{c}_{mh} \doteq \phi_m(\tilde{b}_m) + h\cos\theta_m$ is the x_{m2}-coordinate of the left endpoint $\lim_{x_1\downarrow a_r} \Phi_r(x_1,h)$ of the curve segment $\Phi_r(\Delta_r,\{h\})$.

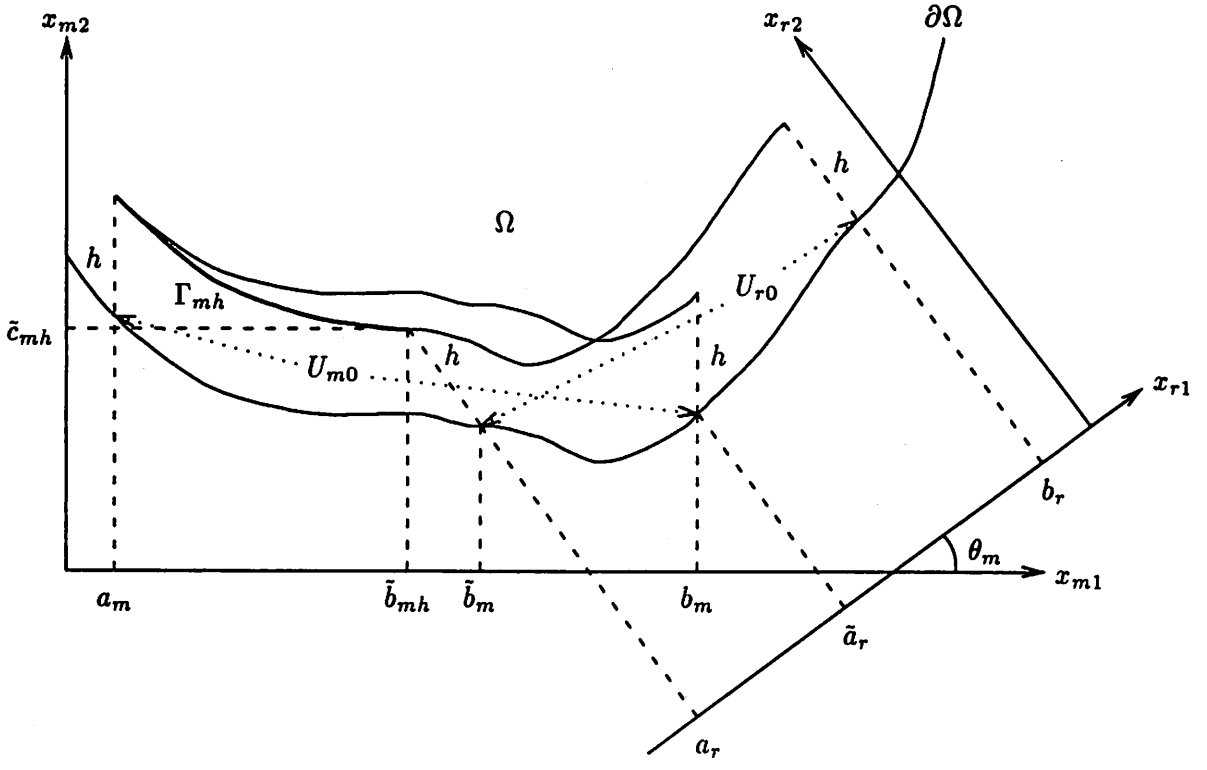


Figure A.2: Construction of the curve segment Γ_{mh} (heavy line).

From the angle condition (A.4) it follows, as one can see from figure A.2, that

$$0 < \tilde{c}_{mh} - \phi_m(\tilde{b}_{mh}) \leq L_m h |\sin\theta_m| + h\cos\theta_m < L_m h + h \tag{A.7}$$

Since

$$\lim_{x_1 \downarrow a_m} \phi_{mh}(x_1) = \lim_{x_1 \downarrow a_m} \phi_m(x_1) + h \tag{A.8}$$

$$\phi_{mh}(\tilde{b}_{mh}) = \tilde{c}_{mh} \tag{A.9}$$

and $\phi_{mh} - \phi_m$ is affine, (A.7) implies that

$$0 < \phi_{mh}(x_1) - \phi_m(x_1) < (L_m + 1)h \qquad \forall x_1 \in]a_m, \tilde{b}_{mh}] \tag{A.10}$$

Hence for sufficiently small $h > 0$ we have

$$\Gamma_{mh} \doteq \{T_m(x_1, \phi_{mh}(x_1)) : x_1 \in]a_m, \tilde{b}_{mh}]\} \subseteq U_{m+} \subseteq \Omega \tag{A.11}$$

The geometrical construction of this curve segment from the two boundary segments U_{m0} and U_{r0} is illustrated in figure A.2.

From (A.5) and (A.7) it follows that

$$\frac{\tilde{c}_{mh} - \phi_m(\tilde{b}_{mh}) - h}{\tilde{b}_{mh} - a_m} = O(h)$$

The function defined by (A.6) is therefore Lipschitz continuous with Lipschitz constant

$$\tilde{L}_{mh} = L_m + O(h) \tag{A.12}$$

By (A.3) and the differentiability of the arctan-function we also have

$$\tilde{\eta}_{mh} \doteq \arctan \tilde{L}_{mh} = \eta_m + O(h) \tag{A.13}$$

The procedure above can of course be carried out for each one of the functions $\phi_1, \ldots, \phi_M$. For sufficiently small $h > 0$ the angle condition (A.4) then also applies to $\tilde{\eta}_{mh}$ and $\tilde{\eta}_{rh}$, that is

$$|\theta_m| + (\tilde{\eta}_{mh} \vee \tilde{\eta}_{rh}) < \frac{\pi}{2} \tag{A.14}$$

This means that the (transformed) curve segment $T_m^{-1}(\Gamma_{rh})$ is the graph of a function on some interval $]\tilde{b}_{mh}, \tilde{d}_{mh}]$ (of the x_{m1}-axis) where $\tilde{d}_{mh}$ is the x_{m1}-coordinate of the right end-point $T_r(\tilde{b}_{rh}, \tilde{c}_{rh})$ of Γ_{rh}. According to (A.5) and (A.7) $\lim_{h \downarrow 0} T_r(\tilde{b}_{rh}, \tilde{c}_{rh}) = T_r(\tilde{b}_r, \phi_r(\tilde{b}_r))$ and by the nonredundancy condition (A.1) (applied to U_{r0}) we have $T_r(\tilde{b}_r, \phi_r(\tilde{b}_r)) \notin \overline{U_{m0}}$.

From the angle condition (A.4) it therefore follows that $\lim_{h \downarrow 0} \tilde{d}_{mh} > b_m$. Hence for sufficiently small $h > 0$, $T_m^{-1}(U_m \cap \Gamma_{rh})$ is the graph of a function $\varphi_{mh} :]\tilde{b}_{mh}, b_m[\rightarrow \mathbf{R}$. Let

$$\phi_{mh}(x_1) \doteq \varphi_{mh}(x_1) \qquad x_1 \in]\tilde{b}_{mh}, b_m[\tag{A.15}$$

Together (A.6) and (A.15) then define a function $\phi_{mh} : \Delta_m \rightarrow \mathbf{R}$.

We already know that ϕ_{mh} is Lipschitz continuous on $]a_m, \tilde{b}_{mh}]$ with Lipschitz constant given by (A.12). From (A.13), the angle condition (A.14) and the differentiability of the tan-function it also follows that ϕ_{mh} is Lipschitz continuous on $]\tilde{b}_{mh}, b_m[$ with Lipschitz constant

$$\tan(|\theta_m| + \tilde{\eta}_{rh}) = \tan(\theta_m + \eta_r) + O(h)$$

Finally, from (A.7), (A.8) (applied to ϕ_{rh}), (A.9) and (A.11) (applied to Γ_{rh}) we see that ϕ_{mh} is continuous at $\tilde{b}_{mh}$. Hence $\phi_{mh} : \Delta_m \rightarrow \mathbf{R}$ is Lipschitz continuous with Lipschitz constant

$$L_{mh} = [L_m \vee \tan(|\theta_m| + \eta_r)] + O(h) \tag{A.16}$$

By the definition of φ_{mh} and (A.11) it follows that $T_m(\mathcal{F}_{\varphi_{mh}}) \subseteq U_m \cap \Gamma_{rh} \subseteq U_m \cap \Omega = U_{m+}$. Thus

$$\phi_{mh}(x_1) = \varphi_{mh}(x_1) > \phi_m(x_1) \qquad \forall x_1 \in]\tilde{b}_{mh}, b_m[$$

From (A.10) (applied to ϕ_{rh}) and the angle condition (A.4) we also have that

$$\phi_{mh}(x_1) - \phi_m(x_1) < L_{mh} h \sin|\theta_m| + (L_m + 1)h \cos\theta_m \qquad \forall x_1 \in]\tilde{b}_{mh}, b_m[$$

By (A.16) the asymptotic behavior of ϕ_{mh} on $]a_m, \tilde{b}_{mh}]$, as dictated by (A.10), therefore generalizes to

$$0 < \phi_{mh}(x_1) - \phi_m(x_1) < (L_{mh} + L_m + 1)h = O(h) \qquad \forall x_1 \in \Delta_m \tag{A.17}$$

From the discussion above we conclude that there exist three finite constants $H > 0$, $L_\phi > \bigvee_{m=1}^{M} L_m$ and $K_\phi > 0$, such that for each $m = 1, \ldots, M$ and each $h \in]0, H]$, (A.6) and (A.15) define a function $\phi_{mh} : \Delta_m \rightarrow \mathbf{R}$ that satisfies the Lipschitz condition

$$|\phi_{mh}(y_1) - \phi_{mh}(x_1)| \leq L_\phi |y_1 - x_1| \qquad \forall x_1, y_1 \in \Delta_m \tag{A.18}$$

and the "localization" condition

$$0 < \phi_{mh}(x_1) - \phi_m(x_1) \leq K_\phi h < \frac{d}{2} \qquad \forall x_1 \in \Delta_m \tag{A.19}$$

A.2.2 The Induced Lipschitz Charts

Let for each $m = 1, \ldots, M$ and each $h \in]0, H]$

$$\Phi_{mh}(x) \doteq \begin{cases} T_m(x_1, \phi_{mh}(x_1) + \chi_{mh\pm}(x_1)x_2) & \text{if } x \in Q_{m\pm} \\ T_m(x_1, \phi_{mh}(x_1)) & \text{if } x \in Q_{m0} \end{cases} \tag{A.20}$$

where

$$\chi_{mh\pm}(x_1) \doteq 1 \pm \frac{\phi_m(x_1) - \phi_{mh}(x_1)}{d} \qquad x_1 \in \Delta_m \tag{A.21}$$

We thereby obtain a collection $\{\mathcal{A}_h \doteq \{\Phi_{mh}\}_{m=1}^M\}_{h\in]0,H]}$ of modified atlases. In view of (A.19) it is easy to check that the map $x_2 \mapsto \phi_{mh}(x_1) + \chi_{mh+}(x_1)x_2$ maps $]0, d[$ *onto* the nonempty interval $]\phi_{mh}(x_1), \phi_m(x_1)+d[$, and that $x_2 \mapsto \phi_{mh}(x_1)+\chi_{mh-}(x_1)x_2$ maps $]-d, 0[$ *onto* the nonempty interval $]\phi_m(x_1) - d, \phi_{mh}(x_1)[$. Hence the three sets

$$U_{mh+} \doteq \Phi_{mh}(Q_{m+}) = \{T_m(x) : \phi_{mh}(x_1) < x_2 < \phi_m(x_1) + d, x_1 \in \Delta_m\} \tag{A.22a}$$

$$U_{mh-} \doteq \Phi_{mh}(Q_{m-}) = \{T_m(x) : \phi_m(x_1) - d < x_2 < \phi_{mh}(x_1), x_1 \in \Delta_m\} \tag{A.22b}$$

$$U_{mh0} \doteq \Phi_{mh}(Q_{m0}) = \{T_m(x_1, \phi_{mh}(x_1)) : x_1 \in \Delta_m\} \tag{A.22c}$$

are, as shown in figure A.3, mutually disjoint and satisfy the conditions

$$U_{mh+} \cup U_{mh-} \cup U_{mh0} = U_m \tag{A.23}$$

$$U_{mh+} \cup U_{mh0} \subseteq U_{m+} \tag{A.24}$$

$$U_{mh-} \supseteq U_{m-} \cup U_{m0} \tag{A.25}$$

From the angle conditions (A.4) and (A.14) it moreover follows that

$$U_{mh+} \subseteq \complement U_{lh-} \cap \complement U_{rh-} \tag{A.26}$$

and

$$\partial U_{mh-} \subseteq \overline{U_{m-}} \cup U_{lh-} \cup U_{rh-} \cup U_{mh0} \cup U_{lh0} \cup U_{rh0} \tag{A.27}$$

Fact A.2.1 $\bigcup_{m=1}^M \partial U_{mh-} \subseteq \complement\Omega \cup \bigcup_{m=1}^M (U_{mh-} \cup U_{mh0})$.

Proof: Since Ω is open we have that $\overline{U_{m-}} \subseteq \overline{\complement\Omega} \subseteq \complement\Omega$, $m = 1, \ldots, M$. Moreover, by definition $l_m, r_m \in \{1, \ldots, M\}$, $m = 1, \ldots, M$. Hence (A.27) implies that

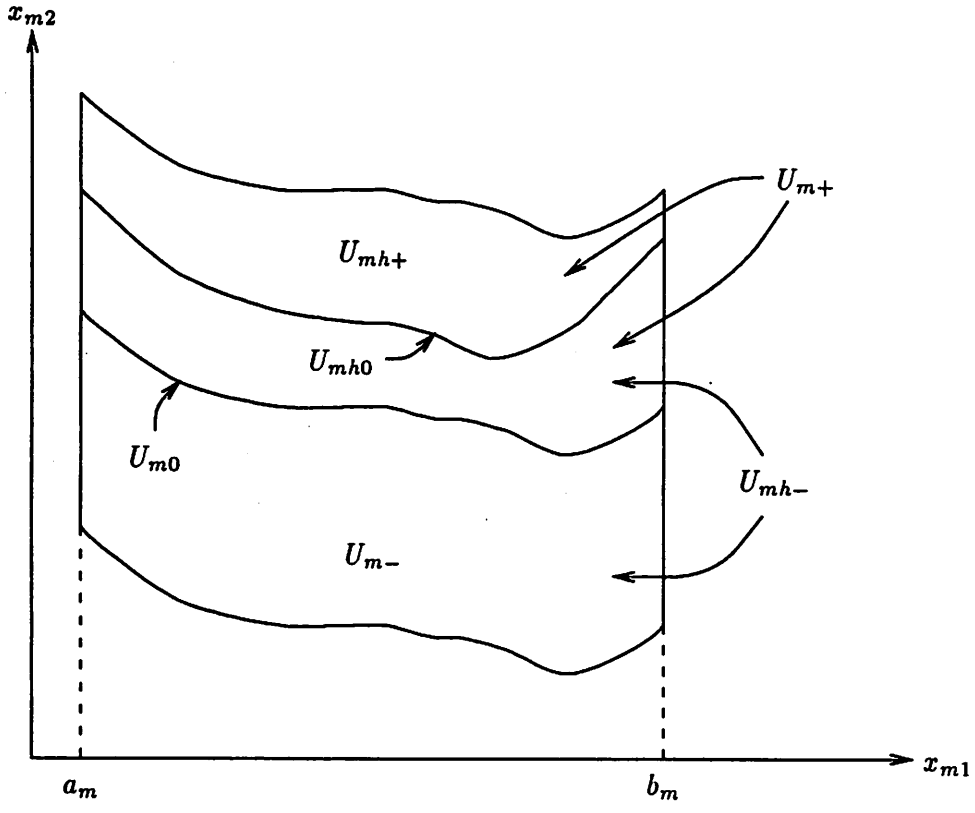


Figure A.3: The sets $U_{mh\pm}$ and U_{mh0} and their relationships with the sets $U_{m\pm}$ and U_{m0}.

$$\bigcup_{m=1}^{M} \partial U_{mh-} \subseteq \bigcup_{m=1}^{M} (\overline{U_{m-}} \cup U_{mh-} \cup U_{mh0}) \subseteq \complement\Omega \cup \bigcup_{m=1}^{M} (U_{mh-} \cup U_{mh0})$$

■

Next we want to show that the maps defined by (A.20) are Lipschitz charts. Since Φ_{mh} is defined differently on each of the three sets $Q_{m\pm}$ and Q_{m0}, the verification of the Lipschitz continuity condition must be separated into multiple cases. The same goes for Φ_{mh}^{-1}, which as it naturally turns out, is given by different expressions on each of the sets $U_{mh\pm}$ and U_{mh0}. By the following proposition, however, the Lipschitz condition can be verified without considering the nowhere dense sets Q_{m0} and U_{mh0}. The number of cases that have to be considered is thereby effectively reduced to only a couple.

Proposition A.2.2 *Let f be a continuous function from a subset X of a normed vector space V to a normed vector space, and let D be a dense subset of X. If $f|D$ is Lipschitz continuous with Lipschitz constant L, the same is also true for f.*

Proof: Let $x, y \in X$ and let $\epsilon > 0$. Since f is continuous and $X = \overline{D}$, $\exists x_D \in D \cap B_V(x, \epsilon)$ and $y_D \in D \cap B_V(y, \epsilon)$ such that $\|f(x_D) - f(x)\| < \epsilon$ and $\|f(y_D) - f(y)\| < \epsilon$. Hence

$$\begin{aligned}
&\|f(y) - f(x)\| \leq \\
&\quad \leq \|f(y) - f(y_D)\| + \|f(y_D) - f(x_D)\| + \|f(x_D) - f(x)\| \\
&\quad < 2\epsilon + L\|y_D - x_D\| \\
&\quad < 2\epsilon + L(\|y_D - y\| + \|y - x\| + \|x - x_D\|) \\
&\quad < L\|x - y\| + 2(1 + L)\epsilon
\end{aligned}$$

Since $\epsilon > 0$ was arbitrarily chosen, the proposition follows. ■

Fact A.2.3 *Each $\Phi \in \mathcal{A} \doteq \bigcup_{h \in]0,H]} \mathcal{A}_h$ is a Lipschitz chart. All the members of $\mathcal{A}$ and their inverses furthermore share a common Lipschitz constant $L_{\mathcal{A}} < \infty$.*

Proof: Let $m \in \{1, \ldots, M\}$, and let $h \in]0, H]$. The coordinate transformation T_m affects neither the properties of invertibility and differentiability nor the values of Lipschitz constants. It will therefore without loss of generality be assumed to be the identity map.

From (A.18) and (A.21) we see that ϕ_{mh} and $\chi_{mh\pm}$ are Lipschitz continuous with Lipschitz constants L_ϕ and $2L_\phi/d$ respectively. Next by (A.19)

$$\sup_{x_1\in\Delta_m}\phi_{mh}(x_1) < \bigvee_{m=1}^{M}\sup_{x_1\in\Delta_m}|\phi_m(x_1)| + \frac{d}{2} < \infty$$

and

$$\frac{1}{2} \le \inf_{x_1\in\Delta_m}\chi_{mh\pm}(x_1) \le \sup_{x_1\in\Delta_m}\chi_{mh\pm}(x_1) \le \frac{3}{2} \tag{A.28}$$

Since Q_m and U_m (being the range of a Lipschitz chart,) are bounded, proposition 3.5.2 therefore implies that the maps $\Phi_{mh\pm} \doteq \Phi_{mh}|Q_{m\pm}$ are Lipschitz charts with inverses given by

$$\Phi_{mh\pm}^{-1} : U_{mh\pm} \to Q_{m\pm} :\mapsto \begin{bmatrix} x_1 \\ \frac{x_2-\phi_{mh}(x_1)}{\chi_{mh\pm}(x_1)} \end{bmatrix} \tag{A.29}$$

and that $\exists$ a constant L, *independent of* m *and* h, such that

$$\|\Phi_{mh\pm}(y) - \Phi_{mh\pm}(x)\| \le L\|y-x\| \qquad \forall x,y \in Q_{m\pm} \tag{A.30a}$$

$$\|\Phi_{mh\pm}^{-1}(y) - \Phi_{mh\pm}^{-1}(x)\| \le L\|y-x\| \qquad \forall x,y \in U_{mh\pm} \tag{A.30b}$$

The map $\Phi_{mh0} \doteq \Phi_{mh}|Q_{m0}$ is of course also invertible with the simple projection

$$\Phi_{mh0}^{-1} : U_{mh0} \to Q_{m0} : x \mapsto \begin{bmatrix} x_1 \\ 0 \end{bmatrix} \tag{A.31}$$

Since the sets U_{mh+}, U_{mh-} and U_{mh0} are mutually disjoint, it thus follows that Φ_{mh} has an inverse given by

$$\Phi_{mh}^{-1}(x) = \begin{cases} \Phi_{mh\pm}^{-1}(x) & \text{if } x \in U_{mh\pm} \\ \Phi_{mh0}^{-1}(x) & \text{if } x \in U_{mh0} \end{cases}$$

Since $\Phi_{mh\pm}$ and $\Phi_{mh\pm}^{-1}$ are differentiable a.e., and the sets Q_{m0} and U_{mh0} are of zero measure (in $\mathbb{R}^2$), it also follows that Φ_{mh} and Φ_{mh}^{-1} are differentiable a.e. It finally remains to prove the claim about the Lipschitz constant $L_{\mathcal{A}}$. Since ϕ_{mh} is continuous and $\chi_{mh\pm}$ satisfy (A.28), it follows directly from (A.20) that Φ_{mh} is continuous on Q_{m0}, and from (A.29) and (A.31) that Φ_{mh}^{-1} is continuous on U_{mh0}. Consider first the (slightly more complicated) map Φ_{mh}^{-1}. Let $x,y \in U_m$. If $x,y \in U_{mh+}$ or $x,y \in U_{mh-}$, then $\|\Phi_{mh}^{-1}(y) - \Phi_{mh}^{-1}(x)\| \le L\|y-x\|$ by (A.30b). If instead $x \in U_{mh+}$ and $y \in U_{mh-}$, then the line segment between x and y must by the intermediate value theorem contain a point $w \in U_{mh0}$. Let $\epsilon > 0$. Since Φ_{mh}^{-1} is continuous

at w and $U_{mh0} \subseteq \partial U_{mh+} \cap \partial U_{mh-}$, $\exists w_+ \in U_{mh+} \cap B_{\mathbb{R}^2}(w,\epsilon)$ and $w_- \in U_{mh-} \cap B_{\mathbb{R}^2}(w,\epsilon)$ such that $\|\Phi_{mh}^{-1}(w_\pm) - \Phi_{mh}^{-1}(w)\| < \epsilon$. Hence by (A.30)

$$\begin{aligned}
&\|\Phi_{mh}^{-1}(y) - \Phi_{mh}^{-1}(x)\| \leq \\
&\quad \leq \|\Phi_{mh}^{-1}(y) - \Phi_{mh}^{-1}(w_-)\| + \|\Phi_{mh}^{-1}(w_-) - \Phi_{mh}^{-1}(w)\| + \|\Phi_{mh}^{-1}(w) - \Phi_{mh}^{-1}(w_+)\| \\
&\qquad + \|\Phi_{mh}^{-1}(w_+) - \Phi_{mh}^{-1}(x)\| \\
&\quad < L(\|y - w_-\| + \|w_+ - x\|) + 2\epsilon \\
&\quad < L(\|y - w\| + \|w - x\| + 2\epsilon) + 2\epsilon \\
&\quad = L\|y - x\| + 2(L+1)\epsilon
\end{aligned}$$

Since $\epsilon > 0$ was arbitrarily chosen, it once again follows that $\|\Phi_{mh}^{-1}(y) - \Phi_{mh}^{-1}(x)\| \leq L\|y - x\|$. We thus conclude that $\Phi_{mh}^{-1}|U_{mh+} \cup U_{mh-}$ is Lipschitz continuous with Lipschitz constant L. Substituting $Q_{m\pm}$, Q_{m0} and Φ_{mh} for $U_{mh\pm}$, U_{mh0} and Φ_{mh}^{-1} respectively in the argument above we likewise find that the same is true for $\Phi_{mh}|Q_{m+} \cup Q_{m-}$. Since $U_{mh+} \cup U_{mh-}$ is dense in U_m, and $Q_{m+} \cup Q_{m-}$ is dense in Q_m, the fact then follows by proposition A.2.2. ■

A.3 A Collection of Interior Set Approximations

In this section we use the modified atlas collection $\{\mathcal{A}_h\}_{h\in]0,H]}$ to construct a collection $\mathcal{F}_\Omega$ of interior set approximations of the domain Ω. We also prove a few results about the properties of the sets in $\mathcal{F}_\Omega$. From these we conclude that $\mathcal{F}_\Omega$ satisfies the conditions (regarding $\mathcal{F}_G$) in definition 3.8.1. The proof of theorem 3.8.2 then follows.

Let $h \in]0, H]$, and define the set

$$F_h \doteq \Omega \setminus \bigcup_{m=1}^{M} \overline{U_{mh-}}$$

Since Ω is bounded and open, so is obviously its interior set approximation F_h. In addition F_h satisfies the conditions listed in the following five facts.

Fact A.3.1 *For each $m = 1, \ldots, M$ the following inclusions hold:*

(i) $U_{mh+} \subseteq F_h$

(ii) $U_{mh-} \subseteq \complement\overline{F_h}$

(iii) $U_{mh0} \subseteq \partial F_h$

Proof: Let $m \in \{1,\ldots,M\}$. From (A.24) we know that $U_{mh+} \subseteq U_{m+} \subseteq \Omega$. By the decomposition (A.23) and the separation condition (A.2) we also have

$$U_{mh+} \subseteq U_m \subseteq \bigcap_{\substack{p=1 \\ p\notin\{m,l,r\}}}^{M} \complement U_p \subseteq \bigcap_{\substack{p=1 \\ p\notin\{m,l,r\}}}^{M} \complement U_{ph-}$$

Since U_{mh+} and U_{mh-} are disjoint, it is moreover true that $U_{mh+} \subseteq \complement U_{mh-}$. Finally by (A.26), $U_{mh+} \subseteq \complement U_{lh-} \cap \complement U_{rh-}$. Combining these four inclusion relations and recalling that U_{mh+} (being homeomorphic to Q_{m+}) and Ω are open we get

$$U_{mh+} = U^{\circ}_{mh+} \subseteq \left(\Omega \cap \bigcap_{p=1}^{M} \complement U_{ph-}\right)^{\circ} = \Omega^{\circ} \cap \bigcap_{p=1}^{M} \left(\complement U_{ph-}\right)^{\circ} = \Omega \setminus \bigcup_{p=1}^{M} \overline{U_{ph-}} = F_h$$

This proves (i). Next we note that

$$F_h \doteq \Omega \setminus \bigcup_{p=1}^{M} \overline{U_{ph-}} \subseteq \complement \overline{U_{mh-}} = \left(\complement U_{mh-}\right)^{\circ}$$

Since U_{mh-} (being homeomorphic to Q_{m-}) is open, we therefore have $\overline{F_h} \subseteq \overline{\left(\complement U_{mh-}\right)^{\circ}} \subseteq \complement U_{mh-}$, which proves (ii). Finally, as the sets $U_{mh\pm}$ and U_{mh0} are images of $Q_{m\pm}$ and Q_{m0} respectively under the same homeomorphism (Φ_{mh}), we have that $U_{mh0} \subseteq \overline{U_{mh+}} \cap \overline{U_{mh-}}$. Hence by (i) and (ii), $U_{mh0} \subseteq \overline{F_h} \cap \overline{\complement F_h} \subseteq \overline{F_h} \cap \complement F_h$, from which (iii) follows. ■

Fact A.3.2 $\overline{F_h} \subseteq \Omega$.

Proof: By its definition $F_h \subseteq \Omega$. Hence $\overline{F_h} \subseteq \overline{\Omega}$. From fact A.3.1 (ii) and (A.25) it moreover follows that

$$\complement \overline{F_h} \supseteq \bigcup_{m=1}^{M} U_{mh-} \supseteq \bigcup_{m=1}^{M} U_{m0} = \partial\Omega$$

Thus $\overline{F_h} \subseteq \overline{\Omega} \cap \complement\partial\Omega \subseteq \Omega$. ■

Fact A.3.3 $\partial F_h \subseteq \bigcup_{m=1}^{M} U_{mh0}$.

Proof: By fact A.3.1 (ii)

$$\bigcup_{m=1}^{M} U_{mh-} \subseteq \mathbb{C}\overline{F_h}$$

Since Ω is open, fact A.2.1 therefore implies that

$$\begin{aligned}
\partial F_h &= \\
&= \partial\left(\Omega \setminus \bigcup_{m=1}^{M} \overline{U_{mh-}}\right) \\
&\subseteq \partial\Omega \cup \bigcup_{m=1}^{M} \partial U_{mh-} \\
&\subseteq \mathbb{C}\Omega \cup \bigcup_{m=1}^{M} (U_{mh-} \cup U_{mh0}) \\
&\subseteq \mathbb{C}\Omega \cup \mathbb{C}\overline{F_h} \cup \bigcup_{m=1}^{M} U_{mh0}
\end{aligned}$$

However, from fact A.3.2 we also see that $\mathbb{C}\Omega \subseteq \mathbb{C}\overline{F_h} \subseteq \mathbb{C}\partial F_h$, whence the fact follows. ■

Fact A.3.4 *The set F_h is a Lipschitz domain.*

Proof: Define the maps

$$\mathrm{X}_m : \Delta_m \times \left]-\frac{d}{2}, \frac{d}{2}\right[\to \mathbb{R}^2 : x \mapsto T_m(x_1, \phi_{mh}(x_1) + x_2) \qquad m = 1, \ldots, M$$

Then by (A.19) and fact A.3.1 (i) and (ii)

$$\mathrm{X}_m\left(\Delta_m \times \left]0, \frac{d}{2}\right[\right) \subseteq U_{mh+} \subseteq F_h \qquad m = 1, \ldots, M$$

and

$$\mathrm{X}_m\left(\Delta_m \times \left]-\frac{d}{2}, 0\right[\right) \subseteq U_{mh-} \subseteq \mathbb{C}\overline{F_h} \qquad m = 1, \ldots, M$$

Furthermore by (A.20), fact A.3.1 (iii) and fact A.3.3

$$\bigcup_{m=1}^{M} \mathrm{X}_m(\Delta_m \times \{0\}) = \bigcup_{m=1}^{M} U_{mh0} = \partial F_h$$

Since F_h is bounded and open, the fact thence follows directly from definition 3.5.1. ■

Fact A.3.5 *The collection $\{F_h, U_1, \ldots, U_M\}$ is an open covering of $\overline{\Omega}$.*

Proof: From fact A.2.1 and the decomposition (A.23) we first note that

$$\bigcup_{m=1}^{M} \overline{U_{mh-}} = \bigcup_{m=1}^{M} (\partial U_{mh-} \cup U_{mh-}) \subseteq \complement\Omega \cup \bigcup_{m=1}^{M} (U_{mh-} \cup U_{mh0}) \subseteq \complement\Omega \cup \bigcup_{m=1}^{M} U_m$$

Hence

$$F_h = \Omega \setminus \bigcup_{m=1}^{M} \overline{U_{mh-}} \supseteq \Omega \cap \complement\left(\complement\Omega \cup \bigcup_{m=1}^{M} U_m\right) = \Omega \setminus \bigcup_{m=1}^{M} U_m$$

which implies that

$$\Omega \subseteq F_h \cup \bigcup_{m=1}^{M} U_m$$

Moreover,

$$\partial\Omega = \bigcup_{m=1}^{M} U_{m0} \subseteq \bigcup_{m=1}^{M} U_m$$

Since F_h and $U_1, \ldots, U_M$ are open, the fact then follows. ■

Consider now the entire collection $\{F_h\}_{h\in]0,H]}$ where F_h for each h is defined as above. Since Ω is open, fact A.3.2 implies that $\overline{F_H} \subset\subset \Omega$, whence $\varrho \doteq \rho(\partial F_H, \partial\Omega) > 0$. From fact A.3.3 we know that $\partial F_H = \bigcup_{m=1}^{M} U_{mH0}$. Likewise $\partial\Omega = \bigcup_{m=1}^{M} U_{m0}$. Since the coordinate transformations $T_1, \ldots, T_M$ preserve distances, (A.19) and (A.22c) therefore imply that

$$K_\phi H \geq \bigwedge_{m=1}^{M} \inf_{x_1 \in \Delta_m} \phi_{mH}(x_1) - \phi_m(x_1) \geq \varrho$$

Let $\tilde{H} \doteq \varrho / K_\phi$. Then obviously $\tilde{H} \leq H$. From (A.19) and (A.22b) it also follows that

$$U_{mh-} \subseteq U_{mH-} \qquad \forall h \in]0, \tilde{H}] \qquad m = 1, \ldots, M \tag{A.32}$$

Let furthermore $\mathcal{F}_\Omega \doteq \{F_h\}_{h\in]0,\tilde{H}]}$. We then have the following two facts.

Fact A.3.6 *There exists a uniformly bounded collection* $\{P_F \in \mathcal{L}(\mathcal{H}^1(F), \mathcal{H}^1(\mathbf{R}^2))\}_{F\in\mathcal{F}_\Omega}$ *of extension operators.*

Proof: Let $\mathcal{U} \doteq \{U_m\}_{m=0}^{M}$ where $U_0 \doteq F_H$. From fact A.3.5 we recall that $\mathcal{U}$ is an open covering of $\overline{\Omega}$. Let thence $\Psi \doteq \{\psi_m\}_{m=0}^{M}$ be a C^∞-partition of unity for $\overline{\Omega}$ subordinate to $\mathcal{U}$. Then let $h \in]0, \tilde{H}]$. According to fact A.3.4 the bounded open set F_h is a Lipschitz domain. From fact A.2.3 we also know that $\mathcal{A}_h$ is an atlas of Lipschitz charts, and that the members of $\mathcal{A}_h$ and their inverses share a common Lipschitz constant $L_{\mathcal{A}} < \infty$, which is independent

of h. The facts A.3.1 and A.3.3 moreover imply that $\mathcal{A}_h$ satisfies the conditions (i)–(iii) of definition 3.5.1 for the domain F_h. Next, from (A.32) we see that

$$U_0 = F_H = \Omega \setminus \bigcup_{m=1}^{M} \overline{U_{mH-}} \subseteq \Omega \setminus \bigcup_{m=1}^{M} \overline{U_{mh-}} = F_h$$

Since $\mathcal{U}$ is an open covering of $\overline{\Omega}$, fact A.3.2 furthermore shows that $\mathcal{U}$ is an open covering of $\overline{F_h}$ as well, which in turn means that Ψ is also a C^∞-partition of unity for $\overline{F_h}$ subordinate to $\mathcal{U}$. We note in particular that $\mathcal{U}$ and hence Ψ are independent of h. Altogether the observations above imply that the hypotheses of theorem 3.7.7 (with F_h playing the role of Ω) are satisfied. Hence $\exists$ an extension operator $P_{F_h} \in \mathcal{L}(\mathcal{H}^1(F_h), \mathcal{H}^1(\mathbf{R}^2))$ with norm $\|P_{F_h}\|_{\mathcal{L}(\mathcal{H}^1(F_h),\mathcal{H}^1(\mathbf{R}^2))}$ bounded (above) by a constant that *only* depends on the three constants M, $L_{\mathcal{A}}$ and $M_\Psi \doteq \bigvee_{m=0}^{M} \bigvee_{l=0}^{1} \sup_{x\in\mathbf{R}^2} \|\psi_m^{(l)}(x)\|$. Since M, $L_{\mathcal{A}}$ and M_Ψ are all independent of h, this completes the proof. ■

Fact A.3.7 $\lim_{h\downarrow 0} m(\Omega \setminus F_h) = 0$.

Proof: Let $p \in \{1, \ldots, M\}$. From (A.22b) and (A.23) we see that

$$\Omega \cap U_{ph-} = U_{p+} \cap U_{ph-} = \{T_p(x) : \phi_p(x_1) < x_2 < \phi_{ph}(x_1), x_1 \in \Delta_p\}$$

Since $m(\partial U_{ph-}) = 0$ and the Jacobian determinant $|J_{T_p}| \equiv 1$, using (A.19) we therefore obtain

$$m(\Omega \cap \overline{U_{ph-}}) = m(\Omega \cap U_{ph-}) = \int_{\Delta_p} (\phi_{ph} - \phi_p)\, dx_1 \leq K_\phi h (b_p - a_p)$$

By the definition of F_h and the subadditivity of the Lebesgue measure it hence follows that

$$m(\Omega \setminus F_h) = m\left(\bigcup_{p=1}^{M} (\Omega \cap \overline{U_{ph-}})\right) \leq K_\phi h \sum_{p=1}^{M} (b_p - a_p) \to 0 \qquad \text{as } h \downarrow 0$$

■

Proof of Theorem 3.8.2: Let Ω be a subset of class $C^{0,1}$ of an image domain B, and define as above $\mathcal{F}_\Omega$ to be a collection of open interior set approximations of Ω. The conditions (i), (ii) and (iii) of definition 3.8.1 then follow by fact A.3.2, fact A.3.6 and fact A.3.7 respectively. ■

Appendix B

Proofs of Results in Section 3.11

A couple of the results in section 3.11 were stated essentially without proof. In this appendix we present proofs of these results along with some of their preliminaries. In section B.1 we prove fact 3.11.6. In section B.2 we prove fact 3.11.7.

B.1 Proof of Fact 3.11.6

Throughout this section we assume, as in the hypotheses of fact 3.11.6, that $l, N \in \mathbf{N}$. We need to show, that the image segmentations in $C^l(\Sigma)^{2N}$, which satisfy the interconnection, regularity and intersection constraints (I1), (I2), (R1), (B1), (B2), (E1) and (E2) presented in section 3.11 is closed in $C^l(\Sigma)^{2N}$. Since intersections of closed sets are closed, these constraints need not be dealt with all at once. We therefore define Γ_I to be the set of image segmentations in $C^l(\Sigma)^{2N}$, which satisfy the interconnection constraints (I1) and (I2). Likewise we define Γ_{R1}, Γ_{B1}, Γ_{B2}, Γ_{E1} and Γ_{E2}, to be the sets of image segmentations in $C^l(\Sigma)^{2N}$, which satisfy the constraints (R1), (B1), (B2), (E1) and (E2) respectively. We will show, that each of the sets Γ_I, Γ_{R1}, Γ_{B1}, Γ_{B2}, Γ_{E1} and Γ_{E2} is closed in $C^l(\Sigma)^{2N}$. Fact 3.11.6 then readily follows.

Fact B.1.1 *The set Γ_I is closed in $C^l(\Sigma)^{2N}$.*

Proof: Let $\Gamma_{\mathcal{I}}$ be the set of all image segmentations in $C^l(\Sigma)^{2N}$ satisfying the interconnection constraints associated with the directed graph $\mathcal{I}$ according to (I1). Suppose $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C^l(\Sigma)^{2N} \setminus \Gamma_{\mathcal{I}}$. Then $\exists$ two joining endpoints $(n,s),(p,t) \in E_N$, such

that $\gamma_n(s) \neq \gamma_p(t)$, or $\exists\ (n,s) \in \mathcal{N}_{\partial B}$, such that $\gamma_n(s) \notin \partial B$. In the first case

$$B_{C^l(\Sigma)^{2N}}\left(\gamma, \frac{\|\gamma_n(s) - \gamma_p(t)\|}{2\sqrt{2}}\right) \subseteq C^l(\Sigma)^{2N} \setminus \Gamma_{\mathcal{I}}$$

In the second case the point $\gamma_n(s)$ is separated from the closed set ∂B by a Euclidean distance $\delta > 0$. Hence

$$B_{C^l(\Sigma)^{2N}}\left(\gamma, \frac{\delta}{\sqrt{2}}\right) \subseteq C^l(\Sigma)^{2N} \setminus \Gamma_{\mathcal{I}}$$

In either case $\exists$ a nonempty open $C^l(\Sigma)^{2N}$-ball centered at γ, which is contained in $\complement\Gamma_{\mathcal{I}}$. Thus $\Gamma_{\mathcal{I}}$ is closed in $C^l(\Sigma)^{2N}$. Since $N < \infty$, the space E_N is finite, and therefore the number of possible interconnections of N edge segments is also finite. In particular the number of such interconnections identified with directed graphs, which satisfy (I2), is finite. Since

$$\Gamma_I = \bigcup_{\mathcal{I} \text{ satisfies (I2)}} \Gamma_{\mathcal{I}}$$

and the union of finitely many closed sets is closed, the fact follows. ■

Fact B.1.2 *The set Γ_{R1} is closed in $C^l(\Sigma)^{2N}$.*

Proof: Suppose $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C^l(\Sigma)^{2N} \setminus \Gamma_{R1}$. By (R1) this is equivalent to, that $\exists$ $n \in \{1,\ldots,N\}$ and $\sigma \in \Sigma$, such that $\|\dot{\gamma}_n(\sigma)\| < \omega$. Since $l \geq 1$, this implies that

$$B_{C^l(\Sigma)^{2N}}\left(\gamma, \frac{\omega - \|\dot{\gamma}_n(\sigma)\|}{\sqrt{2}}\right) \subseteq C^l(\Sigma)^{2N} \setminus \Gamma_{R1}$$

Hence Γ_{R1} is closed in $C^l(\Sigma)^{2N}$. ■

For the proofs, that the intersection constraints are closed, we will need the following elementary preliminary result.

Proposition B.1.3 *Let $v, w, x, y \in \mathbf{R}^n \setminus \{0\}$. Then the following is true:*

(i) $$\left|\frac{|v_k|}{\|v\|} - \frac{|w_k|}{\|w\|}\right| \leq \left\|\frac{v}{\|v\|} - \frac{w}{\|w\|}\right\| \leq \frac{2\|v-w\|}{\|w\|} \qquad k = 1,\ldots,n$$

(ii) $$\left|\frac{v^T x}{\|v\|\|x\|} - \frac{w^T y}{\|w\|\|y\|}\right| \leq \frac{2\|v-w\|}{\|w\|} + \frac{2\|x-y\|}{\|y\|}$$

Proof: Let $k \in \{1,\ldots,n\}$. Since

$$\begin{aligned}
\left|\frac{|v_k|}{\|v\|} - \frac{|w_k|}{\|w\|}\right| \leq \\
&\leq \left|\frac{v_k}{\|v\|} - \frac{w_k}{\|w\|}\right| \\
&\leq \left\|\frac{v}{\|v\|} - \frac{w}{\|w\|}\right\| \\
&= \frac{\|\,\|w\|v - \|v\|w\,\|}{\|v\|\|w\|} \\
&\leq \frac{|\,\|w\| - \|v\|\,|\,\|v\| + \|v\|\|v-w\|}{\|v\|\|w\|} \\
&\leq \frac{2\|v-w\|}{\|w\|}
\end{aligned}$$

(i) follows, and therefore

$$\left|\frac{v^T x}{\|v\|\|x\|} - \frac{w^T y}{\|w\|\|y\|}\right| \leq \left\|\frac{v}{\|v\|} - \frac{w}{\|w\|}\right\| + \left\|\frac{x}{\|x\|} - \frac{y}{\|y\|}\right\| \leq \frac{2\|v-w\|}{\|w\|} + \frac{2\|x-y\|}{\|y\|}$$

which proves (ii). ■

Fact B.1.4 *The set Γ_{B1} is closed in $C^l(\Sigma)^{2N}$.*

Proof: Suppose $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C^l(\Sigma)^{2N} \setminus \Gamma_{B1}$, that is $\exists\, n \in \{1,\ldots,N\}$ and $\sigma \in \Sigma$, such that both the conditions (i) and (ii) or both the conditions (iii) and (iv) of the constraint (B1) are violated. Suppose (i) and (ii) are violated, or equivalently that

$$\gamma_{n1}(\sigma) \notin [a+\delta_0, b-\delta_0] \tag{B.1}$$

and

$$|\dot{\gamma}_{n2}(\sigma)| > (1-\delta_1)\|\dot{\gamma}_n(\sigma)\| \tag{B.2}$$

Since the set $[a+\delta_0, b-\delta_0]$ is closed, (B.1) implies, that it is separated from $\gamma_{n1}(\sigma)$ by a distance $r_0 > 0$ (in $\mathbf{R}$). Hence

$$\beta_{n1}(\sigma) \notin [a+\delta_0, b-\delta_0] \quad \forall \beta \in B_{C^l(\Sigma)^{2N}}(\gamma, r_0) \tag{B.3}$$

From (B.2) we see, that $\dot{\gamma}_n(\sigma) \neq 0$, so (B.2) is actually equivalent to

$$\frac{|\dot{\gamma}_{n2}(\sigma)|}{\|\dot{\gamma}_n(\sigma)\|} > 1 - \delta_1$$

and then by proposition B.1.3 (i)

$$\frac{|\dot{\beta}_{n2}(\sigma)|}{\|\dot{\beta}_n(\sigma)\|} > 1 - \delta_1$$

$\forall \beta \in C^l(\Sigma)^{2N}$, for which

$$\dot{\beta}_n(\sigma) \neq 0 \tag{B.4}$$

and

$$\frac{2\|\dot{\beta}_n(\sigma) - \dot{\gamma}_n(\sigma)\|}{\|\dot{\gamma}_n(\sigma)\|} < \frac{|\dot{\gamma}_{n2}(\sigma)|}{\|\dot{\gamma}_n(\sigma)\|} - 1 + \delta_1 \tag{B.5}$$

Now (B.5) implies that

$$\begin{aligned}
&|\dot{\beta}_{n2}(\sigma)| \geq \\
&\quad \geq |\dot{\gamma}_{n2}(\sigma)| - \|\dot{\gamma}_n(\sigma) - \dot{\beta}_n(\sigma)\| \\
&\quad > |\dot{\gamma}_{n2}(\sigma)| - \frac{|\dot{\gamma}_{n2}(\sigma)| - (1-\delta_1)\|\dot{\gamma}_n(\sigma)\|}{2} \\
&\quad = \frac{|\dot{\gamma}_{n2}(\sigma)| + (1-\delta_1)\|\dot{\gamma}_n(\sigma)\|}{2} \\
&\quad > 0
\end{aligned}$$

Thus (B.5) is a stronger condition, than (B.4), which can hence be neglected. Since $l \geq 1$, we therefore have that

$$\frac{|\dot{\beta}_{n2}(\sigma)|}{\|\dot{\beta}_n(\sigma)\|} > 1 - \delta_1 \qquad \forall \beta \in B_{C^l(\Sigma)^{2N}}(\gamma, r_1) \tag{B.6}$$

where by (B.2)

$$r_1 \doteq \frac{|\dot{\gamma}_{n2}(\sigma)| - (1-\delta_1)\|\dot{\gamma}_n(\gamma)\|}{2\sqrt{2}} > 0$$

From (B.3) and (B.6) we now see that

$$B_{C^l(\Sigma)^{2N}}(\gamma, r_0 \wedge r_1) \subseteq C^l(\Sigma)^{2N} \setminus \Gamma_{B1}$$

If instead of (i) and (ii), conditions (iii) and (iv) are violated, a similar proof shows, that $\exists$ an open nonempty $C^l(\Sigma)^{2N}$-ball centered at γ and contained in $C^l(\Sigma)^{2N} \setminus \Gamma_{B1}$. Hence Γ_{B1} is closed in $C^l(\Sigma)^{2N}$. ■

Fact B.1.5 *The set Γ_{B2} is closed in $C^l(\Sigma)^{2N}$.*

Proof: Suppose $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C^l(\Sigma)^{2N} \setminus \Gamma_{B2}$. By (B2) this is equivalent to, that $\exists$ $(n,s) \in E_N \setminus \mathcal{N}_{\partial B}$ and $x \in \partial B$, such that $\|\gamma_n(s) - x\| < \delta_0$. Hence

$$B_{C^l(\Sigma)^{2N}}\left(\gamma, \frac{\delta_0 - \|\gamma_n(s) - x\|}{\sqrt{2}}\right) \subseteq C^l(\Sigma)^{2N} \setminus \Gamma_{B2}$$

which shows, that Γ_{B2} is closed in $C^l(\Sigma)^{2N}$. ■

Fact B.1.6 *The set Γ_{E1} is closed in $C^l(\Sigma)^{2N}$.*

Proof: Suppose $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C^l(\Sigma)^{2N} \setminus \Gamma_{E1}$. By (E1) this is equivalent to, that $\exists$ $n,p \in \{1,\ldots,N\}$ and $(\sigma,\tau) \in T_{np}$, such that $\|\gamma_n(\sigma) - \gamma_p(\tau)\| < \delta_0$. Hence

$$B_{C^l(\Sigma)^{2N}}\left(\gamma, \frac{\delta_0 - \|\gamma_n(\sigma) - \gamma_p(\tau)\|}{2\sqrt{2}}\right) \subseteq C^l(\Sigma)^{2N} \setminus \Gamma_{E1}$$

which shows, that Γ_{E1} is closed in $C^l(\Sigma)^{2N}$. ■

Fact B.1.7 *The set Γ_{E2} is closed in $C^l(\Sigma)^{2N}$.*

Proof: Suppose $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C^l(\Sigma)^{2N} \setminus \Gamma_{E2}$. By (E2) this is equivalent to, that $\exists$ two joining endpoints $(n,s),(p,t) \in E_N$, such that

$$(-1)^{s+t}\dot{\gamma}_n(\sigma)^T\dot{\gamma}_p(\tau) > (1-\delta_1)\|\dot{\gamma}_n(\sigma)\|\|\dot{\gamma}_p(\tau)\| \tag{B.7}$$

From (B.7) we note, that $\dot{\gamma}_n(\sigma) \neq 0$ and $\dot{\gamma}_p(\tau) \neq 0$, so (B.7) is equivalent to

$$(-1)^{s+t}\frac{\dot{\gamma}_n(\sigma)^T\dot{\gamma}_p(\tau)}{\|\dot{\gamma}_n(\sigma)\|\|\dot{\gamma}_p(\tau)\|} > 1-\delta_1$$

and then by proposition B.1.3 (ii)

$$(-1)^{s+t}\frac{\dot{\beta}_n(\sigma)^T\dot{\beta}_p(\tau)}{\|\dot{\beta}_n(\sigma)\|\|\dot{\beta}_p(\tau)\|} > 1-\delta_1$$

$\forall \beta \in C^l(\Sigma)^{2N}$, for which

$$\dot{\beta}_n(\sigma) \neq 0 \tag{B.8a}$$

$$\dot{\beta}_p(\tau) \neq 0 \tag{B.8b}$$

and

$$\frac{2\|\dot{\beta}_n(\sigma)-\dot{\gamma}_n(\sigma)\|}{\|\dot{\gamma}_n(\sigma)\|}+\frac{2\|\dot{\beta}_p(\tau)-\dot{\gamma}_p(\tau)\|}{\|\dot{\gamma}_p(\tau)\|}<(-1)^{s+t}\frac{\dot{\gamma}_n(\sigma)^T\dot{\gamma}_p(\tau)}{\|\dot{\gamma}_n(\sigma)\|\|\dot{\gamma}_p(\tau)\|}-1+\delta_1 \qquad \text{(B.9)}$$

Now (B.9) implies that

$$\begin{aligned}
&\|\dot{\beta}_n(\sigma)\| \geq \\
&\quad \geq \|\dot{\gamma}_n(\sigma)\|-\|\dot{\gamma}_n(\sigma)-\dot{\beta}_n(\sigma)\| \\
&\quad > \|\dot{\gamma}_n(\sigma)\|-\frac{(-1)^{s+t}\dot{\gamma}_n(\sigma)^T\dot{\gamma}_p(\tau)-(1-\delta_1)\|\dot{\gamma}_n(\sigma)\|\|\dot{\gamma}_p(\tau)\|}{2\|\dot{\gamma}_p(\tau)\|} \\
&\quad \geq \left(1-\frac{\delta_1}{2}\right)\|\dot{\gamma}_n(\sigma)\| \\
&\quad > 0
\end{aligned}$$

and similarly

$$\|\dot{\beta}_p(\tau)\|>\left(1-\frac{\delta_1}{2}\right)\|\dot{\gamma}_p(\tau)\|>0$$

Thus (B.9) is a stronger condition, than (B.8), which can hence be neglected. Since $l \geq 1$, we therefore have that

$$B_{C^l(\Sigma)^{2N}}(\gamma,r)\subseteq C^l(\Sigma)^{2N}\setminus\Gamma_{E2}$$

where by (B.7)

$$r\doteq\frac{(-1)^{s+t}\dot{\gamma}_n(\sigma)^T\dot{\gamma}_p(\tau)-(1-\delta_1)\|\dot{\gamma}_n(\sigma)\|\|\dot{\gamma}_p(\tau)\|}{4\sqrt{2}\,(\|\dot{\gamma}_n(\sigma)\|\vee\|\dot{\gamma}_p(\tau)\|)}>0$$

Hence Γ_{E2} is closed in $C^l(\Sigma)^{2N}$. ■

The proof of fact 3.11.6 is now trivial.

Proof of Fact 3.11.6: Let Γ be the set of image segmentations in $C^l(\Sigma)^{2N}$, which satisfy all of the interconnection, regularity and intersection constraints (I1), (I2), (R1), (B1), (B2), (E1) and (E2). Then

$$\Gamma=\Gamma_I\cap\Gamma_{R1}\cap\Gamma_{B1}\cap\Gamma_{B2}\cap\Gamma_{E1}\cap\Gamma_{E2}$$

Hence fact 3.11.6 follows from the facts B.1.1 - B.1.7. ■

B.2 Proof of Fact 3.11.7

Throughout this section we assume, that $\gamma = [\gamma_1^T \cdots \gamma_N^T]^T \in C_{l,N}(h, r, \omega, \delta_0, \delta_1, v)$, where the constants $l, N, h, r, \omega, \delta_0, \delta_1$ and v satisfy the hypotheses of fact 3.11.7, that is $l, N \in \mathbf{N}$, $h \in]0, 1]$, $r, \omega, \delta_0 > 0$, $\delta_1, v \in]0, 1[$ and

$$v < \left(\frac{\delta_1 \omega}{2\sqrt{2}r}\right)^{\frac{1}{H}}$$

where $H = h$ if $l = 1$, and $H = 1$ otherwise. We have to show, that the image segmentation γ is admissible. By theorem 3.8.2 it is sufficient to show, that all the components of the corresponding continuity set C_γ are of class $C^{0,1}$. This in turn can be done, by verifying the conditions of the Lipschitz domain characterization given on page 78.

B.2.1 Function Graph Representations

We begin with three results, each of which considers a connected subset of one of the following types of sets:

1. A single edge segment.
2. A union of two joining edge segments.
3. A union of an edge segment and a line segment of the image domain boundary ∂B.

We will show, that each such subset is a curve segment, which can be (re)parametrized by its coordinate along some axis in $\mathbf{R}^2$, so that it is congruent to the graph of a Lipschitz continuous function of the form $\phi :]a, b[\to \mathbf{R}$. It is easy to see, that these curve segments are subsets of ∂C_γ. Later we will also see, that every point on the boundary of C_γ, and therefore every point on the boundary of any of its components, belongs to some curve segment of this kind. From this we will be able to prove, that every component of C_γ is of class $C^{0,1}$.

Fact B.2.1 *Let $n \in \{1, \ldots, N\}$, and assume, that $B_{\mathbf{R}}(\sigma, \epsilon) \subseteq \Sigma$ for some $\epsilon \in]0, v]$. Define the function*

$$f : B_{\mathbf{R}}(\sigma, \epsilon) \to \mathbf{R} : \varsigma \mapsto u^T[\gamma_n(\varsigma) - \gamma_n(\sigma)]$$

where u is the unit vector pointing in the direction of $\dot{\gamma}_n(\sigma)$. Then f is strictly increasing, and the curve segment $\gamma_n(B_{\mathbf{R}}(\sigma, \epsilon))$ is congruent to the graph of a Lipschitz continuous function $\phi :]a, b[\to \mathbf{R}$.

Proof: Since $\gamma_n \in K^2_{l,h,r}$, and $l \geq 1$, we have $\forall \varsigma \in B_{\mathbf{R}}(\sigma, \epsilon)$ that

$$\begin{aligned}
\dot{f}(\varsigma) = & \\
&= u^T \dot{\gamma}_n(\sigma) + u^T[\dot{\gamma}_n(\varsigma) - \dot{\gamma}_n(\sigma)] \\
&\geq \|\dot{\gamma}_n(\sigma)\| - \|\dot{\gamma}_n(\varsigma) - \dot{\gamma}_n(\sigma)\| \\
&\geq \omega - \sqrt{2}r|\varsigma - \sigma|^H \\
&> \omega - \sqrt{2}r\upsilon^H \\
&> 0
\end{aligned}$$

Hence f is continuous and strictly increasing, and thus a bijection from $B_{\mathbf{R}}(\sigma, \epsilon)$ to $]a, b[$, where

$$\begin{aligned}
a &\doteq u^T[\gamma_n(\sigma - \epsilon) - \gamma_n(\sigma)] \\
b &\doteq u^T[\gamma_n(\sigma + \epsilon) - \gamma_n(\sigma)]
\end{aligned}$$

We therefore have the situation depicted in figure B.1, from which it is evident that

$$\gamma_n(B_{\mathbf{R}}(\sigma, \epsilon)) \cong {}_F\phi$$

where

$$\phi :]a, b[\rightarrow \mathbf{R} : y_1 \mapsto u_\perp^T[\gamma_n(f^{-1}(y_1)) - \gamma_n(\sigma)]$$

Since $\|\dot{\gamma}_n(\varsigma)\| \leq \sqrt{2}r \ \forall \varsigma \in \Sigma$, we also see, that ϕ is Lipschitz continuous with Lipschitz constant

$$L_\phi = \sup_{\varsigma \in B_{\mathbf{R}}(\sigma,\epsilon)} \frac{u_\perp^T \dot{\gamma}_n(\varsigma)}{u^T \dot{\gamma}_n(\varsigma)} \leq \frac{\sqrt{2}r}{\omega - \sqrt{2}r\upsilon^H} < \infty$$

■

Fact B.2.2 *Let $\epsilon \in]0, \upsilon]$, and let $(n, s), (p, t) \in E_N$ be two joining endpoints. Define a function $f :]-\epsilon, \epsilon[$ by*

$$f(\varsigma) \doteq \begin{cases} u^T[\gamma_p(t - (-1)^t\varsigma) - \gamma_p(t)] & \text{if } \varsigma \in]-\epsilon, 0[\\ u^T[\gamma_n(s + (-1)^s\varsigma) - \gamma_n(s)] & \text{if } \varsigma \in [0, \epsilon[\end{cases}$$

where u is the unit vector pointing in the direction of

$$(-1)^s \frac{\dot{\gamma}_n(s)}{\|\dot{\gamma}_n(s)\|} - (-1)^t \frac{\dot{\gamma}_p(t)}{\|\dot{\gamma}_p(t)\|}$$

Then f is strictly increasing, and the curve segment $\gamma_n(\Sigma \cap B_{\mathbf{R}}(s, \epsilon)) \cup \gamma_p(\Sigma \cap B_{\mathbf{R}}(t, \epsilon))$ is congruent to the graph of a Lipschitz continuous function $\phi :]a, b[\rightarrow \mathbf{R}$.

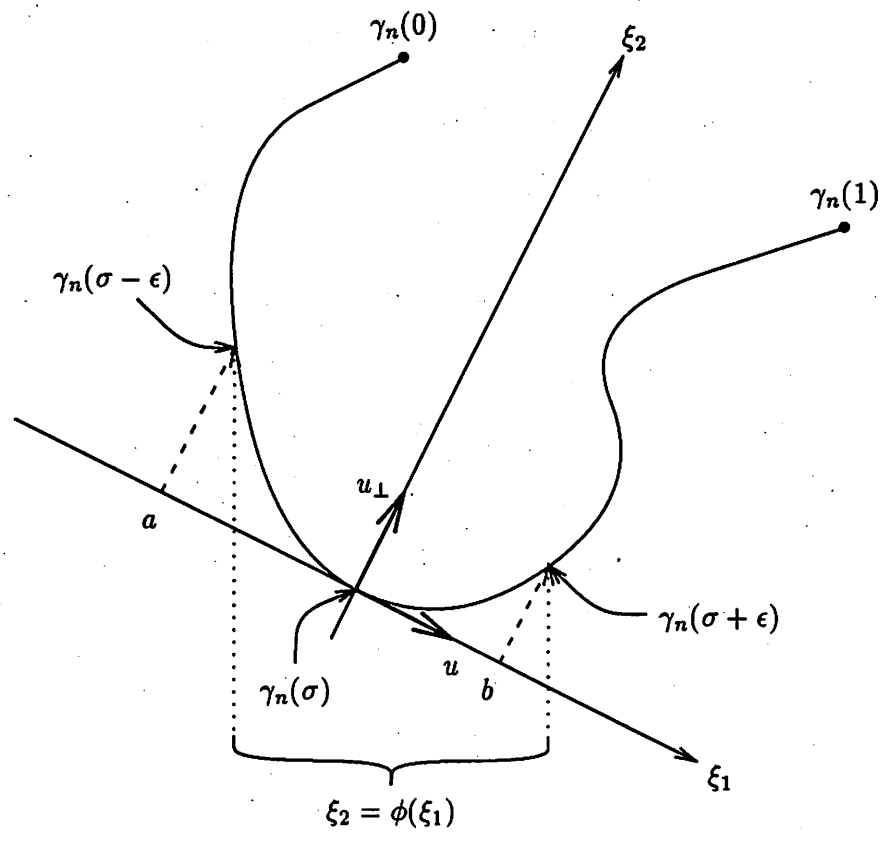


Figure B.1: Function graph representation of subset of single edge segment.

Proof: Let

$$v \doteq (-1)^s \frac{\dot{\gamma}_n(s)}{\|\dot{\gamma}_n(s)\|}$$

and

$$w \doteq (-1)^t \frac{\dot{\gamma}_p(t)}{\|\dot{\gamma}_p(t)\|}$$

Since $(n,s) \bowtie (p,t)$, the edge segment intersection constraint (E2) implies that

$$(-1)^{s+t} \dot{\gamma}_n(s)^T \dot{\gamma}_p(t) \leq (1-\delta_1)\|\dot{\gamma}_n(s)\|\|\dot{\gamma}_p(t)\|$$

Since moreover $\gamma_p \in K^2_{l,h,r}$, and $l \geq 1$, we have $\forall \varsigma \in]-\epsilon, 0[$ that

$$\begin{aligned}
&(v-w)^T \frac{d}{d\varsigma}[\gamma_p(t-(-1)^t\varsigma) - \gamma_p(t)] = \\
&\quad = (w-v)^T(-1)^t\dot{\gamma}_p(t) + (w-v)^T(-1)^t[\dot{\gamma}_p(t-(-1)^t\varsigma) - \dot{\gamma}_p(t)] \\
&\quad \geq \|\dot{\gamma}_p(t)\| - (-1)^{s+t}\frac{\dot{\gamma}_n(s)^T\dot{\gamma}_p(t)}{\|\dot{\gamma}_n(s)\|} - \|w-v\|\|\dot{\gamma}_p(t-(-1)^t\varsigma) - \dot{\gamma}_p(t)\| \\
&\quad \geq \|\dot{\gamma}_p(t)\| - (1-\delta_1)\|\dot{\gamma}_p(t)\| - 2\sqrt{2}r|\varsigma|^H \\
&\quad > \delta_1\omega - 2\sqrt{2}rv^H \\
&\quad > 0
\end{aligned}$$

Hence $v - w \neq 0$, so u is actually well-defined, and

$$\dot{f}(\varsigma) > \frac{\delta_1\omega - 2\sqrt{2}rv^H}{\|v-w\|} > 0 \qquad \forall \varsigma \in]-\epsilon, 0[$$

Since $\gamma_n \in K^2_{l,h,r}$, it similarly follows that

$$\dot{f}(\varsigma) > \frac{\delta_1\omega - 2\sqrt{2}rv^H}{\|v-w\|} > 0 \qquad \forall \varsigma \in]0, \epsilon[$$

In addition, by the continuity of γ_n and γ_p we have

$$\lim_{\varsigma \downarrow 0} f(\varsigma) = f(0) = \lim_{\varsigma \uparrow 0} f(\varsigma) = 0$$

Hence f is continuous and strictly increasing. For the same reasons as in the proof of the previous fact we therefore have that

$$\gamma_n(\Sigma \cap B_{\mathbf{R}}(s,\epsilon)) \cup \gamma_p(\Sigma \cap B_{\mathbf{R}}(t,\epsilon)) \cong {}_f\!\phi$$

where $\phi :]a,b[\rightarrow \mathbf{R}$ is Lipschitz continuous. If we define $u_\perp \in \mathbf{R}^2$ as in figure B.2, the

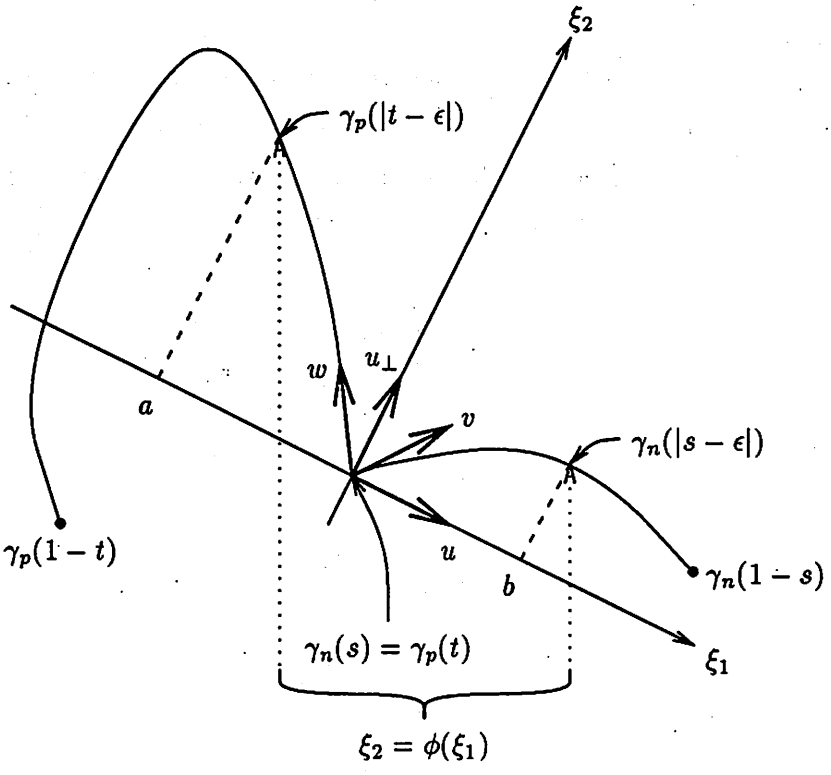


Figure B.2: Function graph representation of subset of two joining edge segments.

function ϕ is defined by

$$\phi(y_1) \doteq \begin{cases} u_\perp^T[\gamma_p(t-(-1)^t f^{-1}(y_1)) - \gamma_p(t)] & \text{if } y_1 \in]a,0[\\ u_\perp^T[\gamma_n(s+(-1)^s f^{-1}(y_1)) - \gamma_n(s)] & \text{if } y_1 \in [0,b[\end{cases}$$

where

$$\begin{aligned} a &\doteq u^T[\gamma_p(|t-\epsilon|) - \gamma_p(t)] \\ b &\doteq u^T[\gamma_n(|s-\epsilon|) - \gamma_n(s)] \end{aligned}$$

■

For the third parametrization result concerning boundary segments, composed by intersecting portions of the edge segments and the image domain boundary, we need an initial parametrization of the image domain boundary segments $\{a\} \times [c,d]$, $\{b\} \times [c,d]$, $[a,b] \times \{c\}$ and $[a,b] \times \{d\}$. We therefore define the functions $\gamma_n : \Sigma \to \mathbf{R}^2$, $n = -3, \ldots, 0$, by

$$\begin{aligned} \gamma_{-3}(\sigma) &\doteq \begin{bmatrix} a \\ c+(d-c)\sigma \end{bmatrix} \\ \gamma_{-2}(\sigma) &\doteq \begin{bmatrix} b \\ c+(d-c)\sigma \end{bmatrix} \\ \gamma_{-1}(\sigma) &\doteq \begin{bmatrix} a+(b-a)\sigma \\ c \end{bmatrix} \\ \gamma_0(\sigma) &\doteq \begin{bmatrix} a+(b-a)\sigma \\ d \end{bmatrix} \end{aligned}$$

and observe, that $\bigcup_{n=-3}^{0} \gamma_n(\Sigma) = \partial B$. We then have the following.

Fact B.2.3 *Assume that* $\gamma_n(\sigma) = \gamma_p(\tau)$ *for some* $n \in \{1,\ldots,N\}$, $p \in \{-3,\ldots,0\}$ *and* $\sigma, \tau \in \Sigma$. *Let* $s,t \in \{0,1\}$, $\epsilon_1 \in]0,v]$ *and* $\epsilon_2 > 0$ *be chosen, so that*

$$\Sigma_n \doteq \sigma + (-1)^s[0,\epsilon_1[\subseteq \Sigma$$

and

$$\Sigma_p \doteq \tau + (-1)^t[0,\epsilon_2[\subseteq \Sigma$$

Then the curve segment $\gamma_n(\Sigma_n) \cup \gamma_p(\Sigma_p)$ *is congruent to the graph of a Lipschitz continuous function* $\phi :]a,b[\to \mathbf{R}$.

Proof: Let

$$v \doteq (-1)^s \frac{\dot{\gamma}_n(\sigma)}{\|\dot{\gamma}_n(\sigma)\|}$$

and

$$w \doteq (-1)^t \frac{\dot{\gamma}_p(\tau)}{\|\dot{\gamma}_p(\tau)\|}$$

and define a function $f :]-\epsilon_2, \epsilon_1[\rightarrow \mathbf{R}$ by

$$f(\varsigma) \doteq \begin{cases} (v-w)^T[\gamma_p(\tau - (-1)^t\varsigma) - \gamma_p(\tau)] & \text{if } \varsigma \in]-\epsilon_2, 0[\\ (v-w)^T[\gamma_n(\sigma + (-1)^s\varsigma) - \gamma_n(\sigma)] & \text{if } \varsigma \in [0, \epsilon_1[\end{cases}$$

As in the proof of the previous fact it is then sufficient to show, that $\dot{f}$ exists, and is bounded below by a strictly positive constant on $]-\epsilon_2, 0[\cup]0, \epsilon_1[$, and that f is continuous at the origin. Suppose $p = 0$. Then $\gamma_{n2}(\sigma) = d$, and

$$\dot{\gamma}_p(\varsigma) = \begin{bmatrix} \|\dot{\gamma}_p(\tau)\| \\ 0 \end{bmatrix} = \begin{bmatrix} b-a \\ 0 \end{bmatrix} \qquad \forall \varsigma \in \Sigma$$

Hence by boundary intersection constraint (B1)

$$|\dot{\gamma}_n(\sigma)^T\dot{\gamma}_p(\tau)| = |\dot{\gamma}_{n1}(\sigma)|\|\dot{\gamma}_p(\tau)\| \leq (1-\delta_1)\|\dot{\gamma}_n(\sigma)\|\|\dot{\gamma}_p(\tau)\|$$

For $\varsigma \in]-\epsilon_2, 0[$ we thus have

$$\begin{aligned} \dot{f}(\varsigma) = & \\ &= (w-v)^T(-1)^t\dot{\gamma}_p(\tau) \\ &\geq \|\dot{\gamma}_p(\tau)\| - \frac{|\dot{\gamma}_n(\sigma)^T\dot{\gamma}_p(\tau)|}{\|\dot{\gamma}_n(\sigma)\|} \\ &\geq \|\dot{\gamma}_p(\tau)\| - (1-\delta_1)\|\dot{\gamma}_p(\tau)\| \\ &= \delta_1(b-a) \\ &> 0 \end{aligned}$$

Since $\gamma_n \in K^2_{l,h,r}$, and $l \geq 1$, for $\varsigma \in]0, \epsilon_1[$ we also have that

$$\begin{aligned} \dot{f}(\varsigma) = & \\ &= (v-w)^T(-1)^s\dot{\gamma}_n(\sigma) + (v-w)^T(-1)^s[\dot{\gamma}_n(\sigma + (-1)^s\varsigma) - \dot{\gamma}_n(\sigma)] \\ &\geq \|\dot{\gamma}_n(\sigma)\| - \frac{|\dot{\gamma}_p(\tau)^T\dot{\gamma}_n(\sigma)|}{\|\dot{\gamma}_p(\tau)\|} - \|v-w\|\|\dot{\gamma}_n(\sigma + (-1)^s\varsigma) - \dot{\gamma}_n(\sigma)\| \\ &\geq \|\dot{\gamma}_n(\sigma)\| - (1-\delta_1)\|\dot{\gamma}_n(\sigma)\| - 2\sqrt{2}r|\varsigma|^H \\ &> \delta_1\omega - 2\sqrt{2}rv^H \\ &> 0 \end{aligned}$$

Finally by the continuity of γ_n and γ_p we have

$$\lim_{\varsigma \uparrow 0} f(\varsigma) = f(0) = \lim_{\varsigma \downarrow 0} f(\varsigma) = 0$$

This completes the proof for $p = 0$. For $p \in \{-3, -2, -1\}$ the proofs are almost identical. ■

B.2.2 Edge Segment Intersections and Simple Curves

Intersections

The main purpose of the three results above is to demonstrate the existence of a sufficiently large family of subsets of ∂C_γ, which are congruent to graphs of Lipschitz continuous functions of the form $\phi :]a, b[\rightarrow \mathbf{R}$. The parametrizations in facts B.2.1 and B.2.2 have a second important consequence, however, as they allow us to characterize all possible edge segment intersections.

Fact B.2.4 *The edge segments $\gamma_n(\Sigma)$, $n = 1, \ldots, N$, intersect (at and) only at joining endpoints. In other words if $\gamma_n(\sigma) = \gamma_p(\tau)$, then either $\sigma, \tau \in \{0, 1\}$ and $(n, \sigma) \bowtie (p, \tau)$, or $(n, \sigma) = (p, \tau)$.*

Proof: Assume that $\gamma_n(\sigma) = \gamma_p(\tau)$ for some $n, p \in \{1, \ldots, N\}$ and $\sigma, \tau \in \Sigma$. Then by edge segment intersection constraint (E1)

$$(\sigma, \tau) \in \Sigma^2 \setminus T_{np}$$

Hence by (3.87) $\exists\ s, t \in \{0, 1\}$, such that $(\sigma, \tau) \in \Upsilon_{np}(s, t)$, and then by (3.86) either $(n, s) = (p, t)$ and $|\sigma - \tau| < v$, or $(n, s) \bowtie (p, t)$ and $|\sigma - s| + |\tau - t| < v$. In the former case the injectivity of the function f in fact B.2.1 implies, that $\sigma = \tau$, and therefore $(n, \sigma) = (p, \tau)$. In the latter case the injectivity of the function f in fact B.2.2 implies, that $\sigma = s$ and $\tau = t$, and therefore $(n, \sigma) \bowtie (p, \tau)$. ■

The fact above has a few interesting consequences: Together with the boundary intersection constraint (B2) it implies, that the image segmentation γ, satisfying the hypotheses in fact 3.11.7, satisfies the interconnection constraints imposed by exactly one interconnection graph. For the rest of this section this graph will be referred to as the

(unique) interconnection graph associated with the image segmentation γ, and denoted by $\mathcal{I}_\gamma$.

Another interesting consequence of fact B.2.4 is, that no edge segments intersections can take place on the image domain boundary. Indeed if there is an edge segment intersection at $\gamma_n(\sigma) = \gamma_p(\tau)$, $n,p \in \{1,\ldots,N\}$, $\sigma,\tau \in \Sigma$, then $(n,\sigma),(p,\tau) \in E_N$ and $(n,\sigma) \bowtie (p,\tau)$. By boundary intersection constraint (B2) it then follows, that $\gamma_n(\sigma) \notin \partial B$.

A third consequence of fact B.2.4 ultimately has to do with the existence of simple curves in the discontinuity set D_γ. For this discussion it will be helpful to define a few new concepts.

Simple Curves

Consider an interconnection graph $\mathcal{I}$. If a branch $\mathcal{B}$ in $\mathcal{I}$ is directly connected to a node $\mathcal{N}$ in $\mathcal{I}$, we say, that the pair $(\mathcal{B},\mathcal{N})$ is a *link* in $\mathcal{I}$. Thus all links are of the form $(\mathcal{B}(n),\mathcal{N}(n,s))$, $(n,s) \in E_N$. Clearly $(\mathcal{B}(n),\mathcal{N}(n,s)) = (\mathcal{B}(p),\mathcal{N}(p,t))$, iff $n = p$ and $(n,s) \sim (p,t)$. Hence the surjective map $(n,s) \mapsto (\mathcal{B}(n),\mathcal{N}(n,s))$ is also one-to-one, except at endpoints of edge segments, which are closed, or both begin and end on ∂B. At such endpoints it is "two-to-one", still mapping endpoints of distinct edge segments to distinct links. We therefore write $(\mathcal{B},\mathcal{N}) \rightarrowtail (n,s)$, (also using the reverse symbol $\rightarrowtail$, when appropriate,) to indicate, that $(\mathcal{B},\mathcal{N}) = (\mathcal{B}(n),\mathcal{N}(n,s))$. A finite sequence $\mathcal{P} \doteq \langle \mathcal{L}_q \rangle_{q=1}^{Q}$, $(Q \in \mathbb{N},)$ of links in $\mathcal{I}$ is said to be a *path* in $\mathcal{I}$, if $\mathcal{N}_q$ is directly connected to $\mathcal{B}_{q+1}$ for $q = 1,\ldots,Q-1$. If in addition $\mathcal{N}_Q$ is directly connected to $\mathcal{B}_1$, we say, that $\mathcal{P}$ is *closed*. The path $\mathcal{P}$, whether closed or not, is said to be *simple*, if the branches $\mathcal{B}_1,\ldots,\mathcal{B}_Q$ and the nodes $\mathcal{N}_1,\ldots,\mathcal{N}_Q$ are all distinct. Finally if a subsequence of $\mathcal{P}$ is also a path in $\mathcal{I}$, it is referred to as a subpath of $\mathcal{P}$.

Proposition B.2.5 *Let $\mathcal{P} \doteq \langle \mathcal{L}_q \rangle_{q=1}^{Q}$ be a closed path in an edge segment interconnection graph. Then there exists a closed subpath $\mathcal{P}' \doteq \langle \mathcal{L}'_q \rangle_{q=1}^{Q'}$ of $\mathcal{P}$, such that $\mathcal{L}'_1 = \mathcal{L}_1$ and $\mathcal{N}'_{Q'} = \mathcal{N}_Q$. If $\mathcal{B}_1 \notin \{\mathcal{B}_2,\ldots,\mathcal{B}_Q\}$, then $\mathcal{P}'$ is simple. If $\mathcal{N}_Q \notin \{\mathcal{N}_1,\ldots,\mathcal{N}_{Q-1}\}$, and $\mathcal{B}_Q \neq \mathcal{B}_1$, then $\mathcal{P}'$ is again simple, and moreover $\mathcal{L}'_{Q'} = \mathcal{L}_Q$.*

Proof: Apply the following algorithm to $\mathcal{P}$.

For $q = 2, \ldots, Q$:

If $\exists\ p < q$, such that $\mathcal{N}_p = \mathcal{N}_q$,

delete $\mathcal{L}_{p+1}, \ldots, \mathcal{L}_q$ from $\mathcal{P}$.

Let $\mathcal{P}' \doteq \langle \mathcal{L}'_q \rangle_{q=1}^{Q'}$ consist of the remaining (less than once deleted) links in the order inherited from $\mathcal{P}$. Then $\mathcal{P}'$ is clearly a closed path with $\mathcal{L}'_1 = \mathcal{L}_1$, $\{\mathcal{L}'_2, \ldots, \mathcal{L}'_{Q'}\} \subseteq \{\mathcal{L}_2, \ldots, \mathcal{L}_Q\}$ and $\mathcal{N}'_{Q'} = \mathcal{N}_Q$. Moreover the nodes $\mathcal{N}'_1, \ldots, \mathcal{N}'_{Q'}$ are distinct. Unless $Q' = 2$, this implies, that the sets $\{\mathcal{N}'_{q-1}, \mathcal{N}'_q\}$, $q = 1, \ldots, Q'$, (where we interpret $\mathcal{N}'_0 \doteq \mathcal{N}'_{Q'}$,) and hence the branches $\mathcal{B}'_1, \ldots, \mathcal{B}'_{Q'}$ are distinct as well. Thus $\mathcal{P}'$ is simple, whenever $Q' \neq 2$. Suppose first, that $\mathcal{B}_1 \notin \{\mathcal{B}_2, \ldots, \mathcal{B}_Q\}$, and that $Q' = 2$. Since $\mathcal{B}'_1 = \mathcal{B}_1$, and $\mathcal{B}'_2 \in \{\mathcal{B}_2, \ldots, \mathcal{B}_Q\}$, it then follows, that $\mathcal{B}'_1 \neq \mathcal{B}'_2$, and hence that $\mathcal{P}'$ is simple. Suppose instead, that $\mathcal{N}_Q \notin \{\mathcal{N}_1, \ldots, \mathcal{N}_{Q-1}\}$, and that $\mathcal{B}_Q \neq \mathcal{B}_1$. Then $\mathcal{L}_Q$ does not get deleted. Thus $\mathcal{L}'_{Q'} = \mathcal{L}_Q$. If $Q' = 2$, we therefore have, that $\mathcal{B}'_1 = \mathcal{B}_1 \neq \mathcal{B}_Q = \mathcal{B}'_2$, and again we conclude, that $\mathcal{P}'$ is simple. ■

Fact B.2.6 *Assume that $\langle \mathcal{L}_q \rangle_{q=1}^{Q} \leftarrowtail \langle (n_q, s_q) \rangle_{q=1}^{Q}$ is a simple closed path in the edge segment interconnection graph $\mathcal{I}_\gamma$, and that Γ is a nonempty connected subset of $B \cap P$, where*

$$P \doteq \bigcup_{q=1}^{Q} \gamma_{n_q}(\Sigma)$$

Then there exists a simple curve S in $\mathbf{R}^2$, such that $\overline{\Gamma} \subseteq S \subseteq \overline{B} \cap P$ with one of the following two additional properties:

(i) *S is a closed curve in B.*

(ii) *S intersects ∂B at and only at its distinct endpoints.*

Proof: Suppose $\mathcal{N}_{\partial B} \notin \{\mathcal{N}_1, \ldots, \mathcal{N}_Q\}$. Then

$$(n_q, s_q) \bowtie (n_{q+1}, 1 - s_{q+1}) \quad q = 1, \ldots, Q-1$$
$$(n_Q, s_Q) \bowtie (n_1, 1 - s_1)$$

By fact B.2.4 the edge segments $\gamma_{n_q}(\Sigma)$, $q = 1, \ldots, Q$, intersect at and only at joining endpoints. Since the branches $\mathcal{B}_1, \ldots, \mathcal{B}_Q$ and the nodes $\mathcal{N}_1, \ldots, \mathcal{N}_Q$ are distinct, it therefore follows, that P is a simple closed curve. In particular P is a closed connected set. If $\partial B \cap P = \emptyset$, as $B \cap P \supseteq \Gamma \neq \emptyset$, it therefore follows, that $P \subseteq B$. Let $S \doteq P$. Then (i)

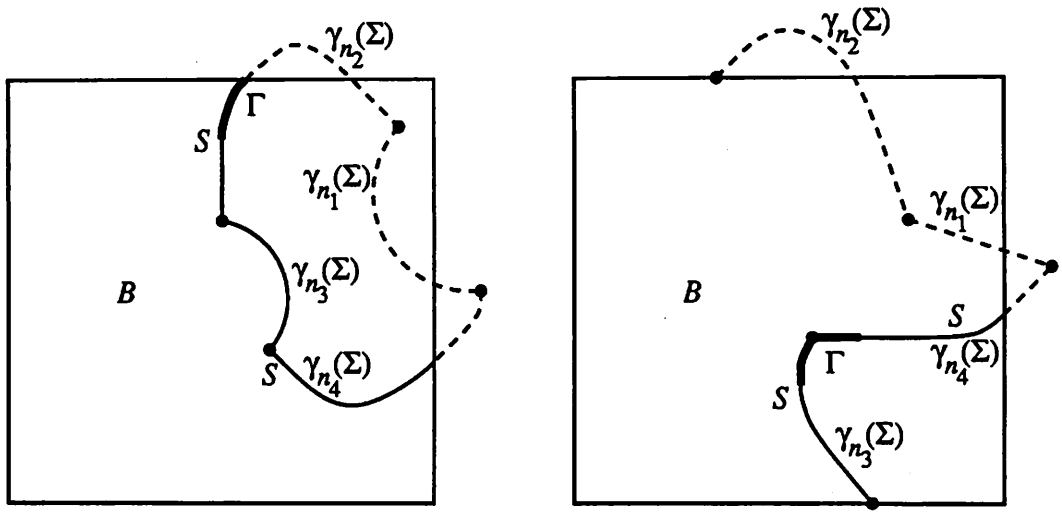


Figure B.3: Simple curve S (solid line) constructed from simple closed path (with $Q = 4$ links) in interconnection graph, such that $\overline{\Gamma}$ (heavy line) $\subseteq S \subseteq \overline{B} \cap \bigcup_{q=1}^{Q} \gamma_{n_q}(\Sigma)$ (entire curve). (a) Path without boundary node. (b) Path with boundary node.

is satisfied, and $\overline{\Gamma} \subseteq S \subseteq \overline{B} \cap P$. If on the other hand $\partial B \cap P \neq \varnothing$, the set $B \cap P$ can have several components, one of which must contain the connected set Γ. By boundary intersection constraint (B2)

$$\gamma_{n_q}(s) \notin \partial B \qquad s = 0, 1 \qquad q = 1, \ldots, Q$$

Thus from boundary intersection constraint (B1) we see, that at every $x \in \partial B \cap P$, P has a tangent, which forms a nonzero angle with the line segment(s) of ∂B passing through x, that is P is transversal to B. The closure of each component of $B \cap P$ is therefore a simple curve with distinct endpoints on ∂B as shown in figure B.3 (a). Let S be the closure of the component containing Γ. Then (ii) is satisfied, and

$$\overline{\Gamma} \subseteq S \subseteq \overline{B \cap P} \subseteq \overline{B} \cap P$$

Suppose instead, that $\mathcal{N}_{\partial B} \in \{\mathcal{N}_1, \ldots, \mathcal{N}_Q\}$. Then $\exists$ a unique $p \in \{1, \ldots, Q\}$, such that $\mathcal{N}_p = \mathcal{N}_{\partial B}$. Shift the path with respect to its index q according to

$$\langle (n'_q, s'_q) \rangle_{q=1}^{Q} \rightarrowtail \langle \mathcal{L}'_q \rangle_{q=1}^{Q} \doteq \langle \mathcal{L}_{p+1}, \ldots, \mathcal{L}_Q, \mathcal{L}_1, \ldots, \mathcal{L}_p \rangle$$

so that $\mathcal{N}'_Q = \mathcal{N}_{\partial B}$. Then

$$(n'_q, s'_q) \bowtie (n'_{q+1}, 1 - s'_{q+1}) \qquad q = 1, \ldots, Q-1$$
$$(n'_Q, s'_Q), (n'_1, 1 - s'_1) \in \mathcal{N}_{\partial B}$$

Thus by fact B.2.4

$$P = \bigcup_{q=1}^{Q} \gamma_{n'_q}(\Sigma)$$

is a simple curve with *distinct* endpoints $\gamma_{n'_1}(1 - s'_1), \gamma_{n'_Q}(s'_Q) \in \partial B$. Again P is a closed set, and the closure of each component of $B \cap P$ is a simple curve with distinct endpoints on ∂B, as shown in figure B.3 (b). Thus as before we let S be the closure of the component containing Γ. Then (ii) is satisfied, and $\overline{\Gamma} \subseteq S \subseteq \overline{B} \cap P$. ■

B.2.3 Lipschitz Property Verification

In verifying the Lipschitz domain characterizing conditions on page 78 for a boundary point of a domain, specified only as a component of the complement (relative to the image domain) of the union of unknown edge segments, it is relatively easy to show, that the boundary point lies on a boundary segment, which is congruent to the graph of a Lipschitz continuous function. In fact this task has essentially already been taken care of by the facts B.2.1 – B.2.3. The hard part is to show, that the two different sides of this boundary segment are locally contained in the domain and its complement respectively. The main hurdle is indeed to show, that these two sides are not in the same component, that is that they are separated by the edge segments. One way to approach this problem is to search for a simple closed curve in the boundary of the continuity set, such that one of the two sides, just mentioned, is locally inside this curve, and the other side is locally outside. The following two facts address this problem.

Fact B.2.7 $\partial C_\gamma = \partial B \cup (B \cap D_\gamma)$.

Proof: Since the discontinuity set D_γ is a null set (in $\mathbf{R}^2$), it has empty interior. Hence $\overline{\complement D_\gamma} = \mathbf{R}^2$. Since the image domain B is open, we therefore have that $B = B \cap \overline{\complement D_\gamma} \subseteq \overline{B \cap \complement D_\gamma}$. It thus follows that $\overline{B} \subseteq \overline{B \cap \complement D_\gamma} \subseteq \overline{B}$, whence

$$\overline{C_\gamma} = \overline{B \setminus D_\gamma} = \overline{B \cap \complement D_\gamma} = \overline{B}$$

Since $\overline{\complement C_\gamma} = \overline{\complement B \cup D_\gamma} = \overline{\complement B} \cup \overline{D_\gamma}$, and D_γ is closed, we therefore obtain

$$\partial C_\gamma = \overline{B} \cap (\overline{\complement B} \cup D_\gamma) = \partial B \cup (\overline{B} \cap D_\gamma) = \partial B \cup (B \cap D_\gamma)$$

■

Fact B.2.8 *Let G be a component of the continuity set C_γ, and let $x \in \partial G$. Then there exist a curve segment Γ, a Lipschitz continuous function $\phi :]a,b[\rightarrow \mathbf{R}$ and an open connected set $\Omega \subseteq \mathbf{R}^2$, such that*

(i) $\Gamma \cong_{f_\phi}$

(ii) *$\partial\Omega$ is a simple closed curve.*

(iii) $\Omega \supseteq G$

(iv) $x \in \Gamma \subseteq \partial\Omega \subseteq \partial C_\gamma$

(v) $\rho(x, \Omega \cap \partial C_\gamma) > 0$

Proof: Since G is a component of C_γ, it follows that $\partial G \subseteq \partial C_\gamma$. By fact B.2.7 the point x must thus belong to the disjoint union $(\partial B \setminus D_\gamma) \cup (\partial B \cap D_\gamma) \cup (B \cap D_\gamma)$. For the rest of this proof we will distinguish between the three possible cases corresponding to $x \in \partial B \setminus D_\gamma$, $x \in \partial B \cap D_\gamma$ and $x \in B \cap D_\gamma$ respectively.

Case 1: $x \in \partial B \setminus D_\gamma$

Since the image domain B is rectangular, it is trivial to show the existence of a curve segment Γ and a Lipschitz continuous function $\phi :]a,b[\rightarrow \mathbf{R}$, such that $\Gamma \cong_{f_\phi}$ and $x \in \Gamma \subseteq \partial B \subseteq \partial C_\gamma$. From fact B.2.7 we see that $B \cap \partial C_\gamma \subseteq D_\gamma$. Since D_γ is compact, and $x \notin D_\gamma$, we therefore have

$$\rho(x, B \cap \partial C_\gamma) \geq \rho(x, D_\gamma) > 0$$

Moreover $B \supseteq C_\gamma \supseteq G$. Since B is an open connected set, and ∂B is a simple closed curve, this completes the proof of case 1.

Case 2: $x \in \partial B \cap D_\gamma$

Assume without loss of generality, that $x = \gamma_1(\sigma)$, $\sigma \in \Sigma$. It is possible, that γ_1 is

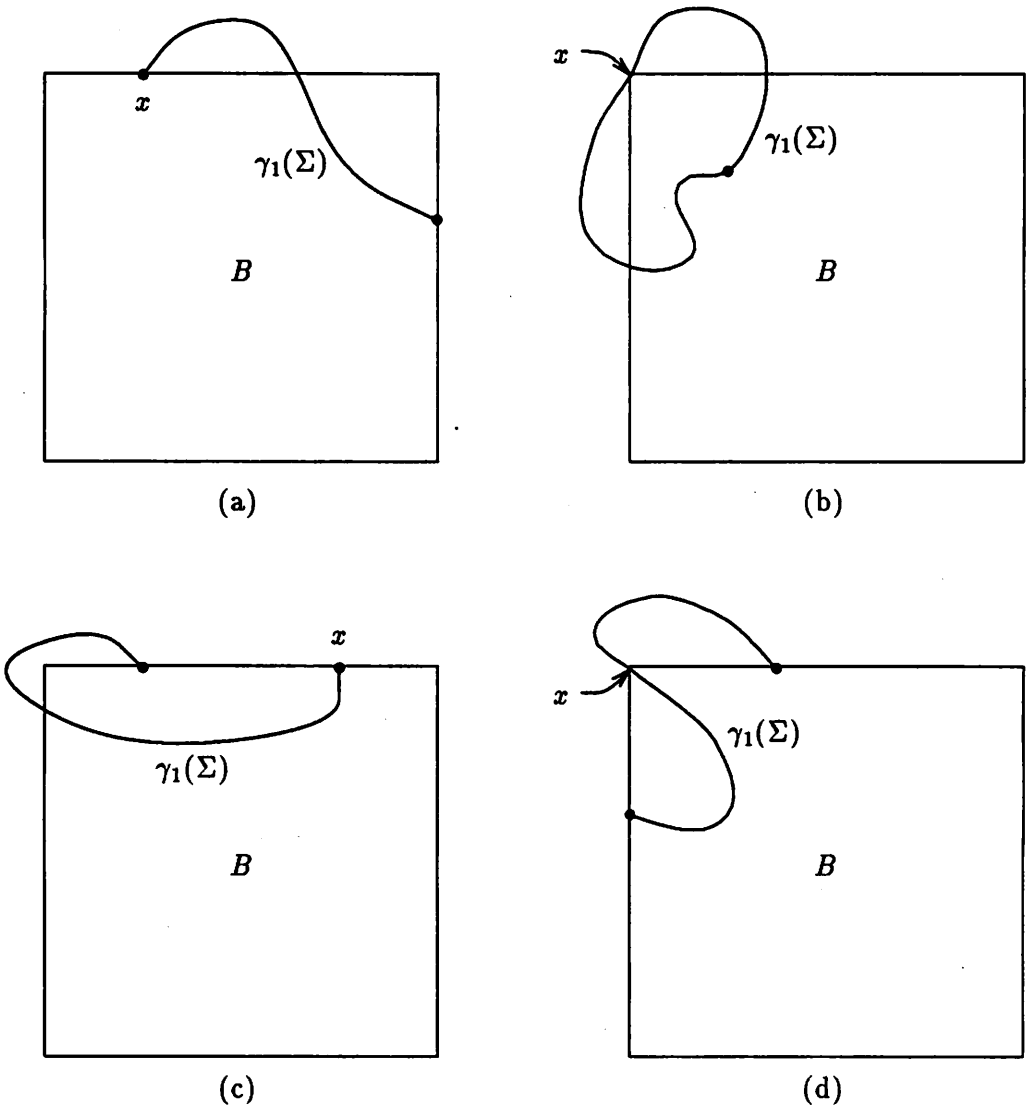


Figure B.4: Intersection of image domain boundary ∂B and edge segment $\gamma_1(\Sigma)$ at point $x = \gamma_1(\sigma)$. (a)-(b) γ_1 locally outside B at σ. (c)-(d) γ_1 *not* locally outside B at σ.

locally outside B at σ, as in figure B.4 (a) and (b). The opposite is also possible, as the configurations in figure B.4 (c) and (d) show. The analysis of case 2 is therefore naturally broken up into two separate subcases.

Subcase 2a: $\exists\ \epsilon > 0$, such that $\gamma_1(\Sigma \cap B_{\mathbb{R}}(\sigma, \epsilon)) \subseteq \complement B$.

Define the compact set

$$D(\epsilon) \doteq \gamma_1(\Sigma \setminus B_{\mathbb{R}}(\sigma, \epsilon)) \cup \bigcup_{n=2}^{N} \gamma_n(\Sigma) \tag{B.10}$$

The edge segments $\gamma_n(\Sigma)$, $n = 1, \ldots, N$, do not intersect on ∂B, so $x \notin D(\epsilon)$. Moreover, from fact B.2.7 we have

$$B \cap \partial C_\gamma \subseteq D_\gamma \setminus \gamma_1(\Sigma \cap B_{\mathbb{R}}(\sigma, \epsilon)) \subseteq D(\epsilon)$$

Hence

$$\rho(x, B \cap \partial C_\gamma) \geq \rho(x, D(\epsilon)) > 0$$

The rest of the proof of this subcase is identical to that of case 1.

Subcase 2b: $\gamma_1(\Sigma \cap B_{\mathbb{R}}(\sigma, \epsilon)) \not\subseteq \complement B \ \ \forall \epsilon > 0$.

Since B is convex, and $\gamma_1(\Sigma)$ forms nonzero angles with ∂B at x, $\exists\ s \in \{0, 1\}$ and $\epsilon_1 \in]0, v]$, such that

$$\Sigma_+ \subseteq \Sigma$$

and

$$\gamma_1(\Sigma_+^\circ) \subseteq B \tag{B.11a}$$

$$\gamma_1(\Sigma \cap \Sigma_-) \subseteq \complement B \tag{B.11b}$$

where

$$\Sigma_\pm \doteq \sigma \pm (-1)^s[0, \epsilon_1[$$

Now by interconnection constraint (I2) the number of components of the interconnection graph $\mathcal{I}_\gamma$ is unchanged, if the branch $\mathcal{B}(1)$ is removed. Hence the node(s) $\mathcal{N}(1, 0)$ and $\mathcal{N}(1, 1)$ (possibly equal) are connected in $\mathcal{I}_\gamma \setminus \{\mathcal{B}(1)\}$, which in turn implies, that $\exists$ a closed path in $\mathcal{I}_\gamma$, with $\mathcal{B}(1)$ figuring in and only in its first link. Thus by proposition B.2.5 this path has a simple closed subpath $\langle \mathcal{L}_q \rangle_{q=1}^{Q} \leftarrowtail \langle (n_q, s_q) \rangle_{q=1}^{Q}$ (in $\mathcal{I}_\gamma$), such that $\mathcal{B}_1 = \mathcal{B}(1)$, that is $\gamma_{n_1} = \gamma_1$. Let

$$P \doteq \bigcup_{q=1}^{Q} \gamma_{n_q}(\Sigma)$$

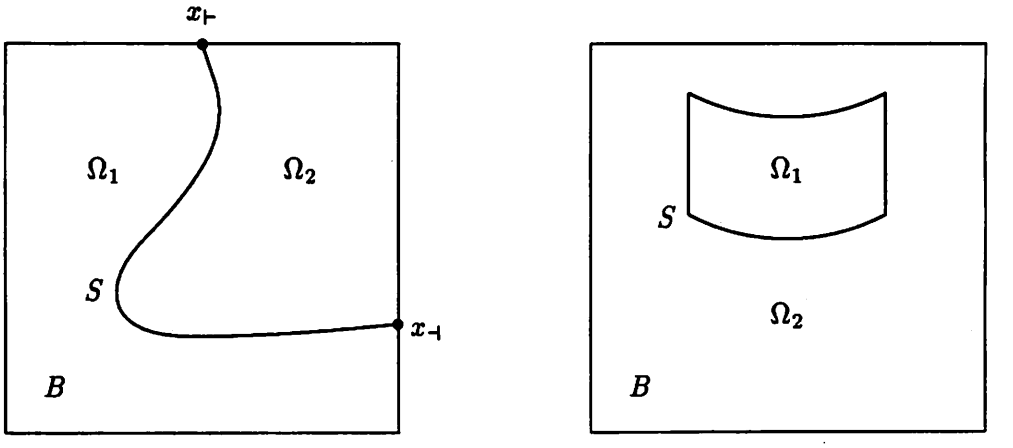


Figure B.5: Components Ω_1 and Ω_2 of $B \setminus S$ for simple curve S. (a) S intersects ∂B at and only at its distinct endpoints $x_{\vdash}$ and $x_{\dashv}$. (b) S is a closed curve in B.

Then $\gamma_1(\Sigma_+^\circ)$ is a connected subset of $B \cap P$. From fact B.2.6 it therefore follows, that $\exists$ a simple curve S intersecting ∂B at and only at its distinct endpoints $x_{\vdash}$ and $x_{\dashv}$, (one of which is x,) such that

$$\gamma_1(\Sigma_+) \subseteq S \subseteq \overline{B} \cap P \subseteq \overline{B} \cap D_\gamma \tag{B.12}$$

Hence S separates $B \setminus S$ into two open components Ω_1 and Ω_2 with simple closed curve boundaries satisfying

$$\begin{aligned} \partial\Omega_1 \cap \partial\Omega_2 &= S \\ \partial\Omega_1 \cup \partial\Omega_2 &= \partial B \cup S \end{aligned}$$

as in figure B.5 (a). Since $S \subseteq D_\gamma$, we have, that $G \subseteq C_\gamma \subseteq B \setminus S$. The connected set G must therefore be contained in one of the components of $B \setminus S$. Assume without loss of generality, that $G \subseteq \Omega_1$. Since $x_{\vdash} \neq x_{\dashv}$, $\exists\ p \in \{-3, \ldots, 0\}$, $\tau \in \Sigma$, $t \in \{0, 1\}$ and $\epsilon_2 > 0$, such that

$$\gamma_p(\tau) = \gamma_1(\sigma)$$

$$\Sigma_p \subseteq \Sigma$$

and

$$\gamma_p(\Sigma_p) \subseteq \partial\Omega_1 \tag{B.13}$$

where

$$\Sigma_p \doteq \tau + (-1)^t[0, \epsilon_2[$$

Thus by fact B.2.3 $\exists$ a Lipschitz continuous function $\phi :]a, b[\rightarrow \mathbf{R}$, such that

$$\Gamma \doteq \gamma_1(\Sigma_+) \cup \gamma_p(\Sigma_p) \cong_{F\phi} \tag{B.14}$$

Next from fact B.2.7, (B.12) and the properties of Ω_1 we see that

$$S \subseteq \partial\Omega_1 \subseteq \partial B \cup S \subseteq \partial B \cup (\overline{B} \cap D_\gamma) = \partial C_\gamma$$

Hence by (B.12), (B.13) and (B.14)

$$x \in \Gamma \subseteq \partial\Omega_1 \subseteq \partial C_\gamma$$

Finally, as $\Omega_1 \subseteq B$, from (B.11b) and (B.12) we have

$$\gamma_1(\Sigma \cap B_{\mathbf{R}}(\sigma, \epsilon_1)) = \gamma_1(\Sigma \cap \Sigma_-) \cup \gamma_1(\Sigma_+) \subseteq \complement B \cup \partial\Omega_1 \subseteq \complement\Omega_1$$

From fact B.2.7 it therefore follows that

$$\Omega_1 \cap \partial C_\gamma = \Omega_1 \cap D_\gamma \subseteq D(\epsilon_1)$$

where $D(\epsilon_1)$ is defined as in (B.10). Thus as before

$$\rho(x, \Omega_1 \cap \partial C_\gamma) \geq \rho(x, D(\epsilon_1)) > 0$$

which completes the proof of case 2.

Case 3: $x \in B \cap D_\gamma$

As in the previous case we assume without loss of generality, that $x = \gamma_1(\sigma)$, $\sigma \in \Sigma$. Since the edge segments are allowed to intersect in B, it is however possible, that x also belongs to some edge segment other than $\gamma_1(\Sigma)$. For this reason the analysis is once again naturally broken up into subcases.

Subcase 3a: $\sigma \in \Sigma^\circ$

Since $x \in B$, $\exists\ \epsilon \in]0, v]$, such that

$$B_{\mathbf{R}}(\sigma, \epsilon) \subseteq \Sigma$$

$$\gamma_1(B_{\mathbf{R}}(\sigma, \epsilon)) \subseteq B$$

Thus by fact B.2.1 $\exists$ a Lipschitz continuous function $\phi :]a, b[\rightarrow \mathbf{R}$, such that

$$\Gamma \doteq \gamma_1(B_{\mathbf{R}}(\sigma, \epsilon)) \cong_{F\phi}$$

For exactly the same reasons as in subcase 2b $\exists$ a simple closed path $\langle \mathcal{L}_q \rangle_{q=1}^{Q}$ in $\mathcal{I}_\gamma$, such that $\mathcal{B}_1 = \mathcal{B}(1)$. Again we let P be the union of the edge segments corresponding to the branch(es) of the path. Then Γ is a connected subset of $B \cap P$. From fact B.2.6 it therefore follows, that $\exists$ a simple curve S, either closed in B, or intersecting ∂B at and only at its distinct endpoints, and such that

$$\Gamma \subseteq S \subseteq \overline{B} \cap P \subseteq \overline{B} \cap D_\gamma \tag{B.15}$$

Whether S is closed or not, from figure B.5 it is clear, that it as before separates $B \setminus S$ into two components Ω_1 and Ω_2 with the same properties as those considered in subcase 2b. In particular $\Omega_1 \supseteq G$. Replacing (B.12) in the analysis of subcase 2b by (B.15), we then find that

$$x \in \Gamma \subseteq \partial\Omega_1 \subseteq \partial C_\gamma$$

This also implies that

$$\Omega_1 \cap \partial C_\gamma = \Omega_1 \cap D_\gamma \subseteq D_\gamma \setminus \Gamma \subseteq D(\epsilon)$$

where $D(\epsilon)$ is defined as in (B.10). By fact B.2.4 $x \notin D(\epsilon)$, so again we obtain

$$\rho(x, \Omega_1 \cap \partial C_\gamma) \geq \rho(x, D(\epsilon)) > 0$$

Subcase 3b: $\sigma \in \partial\Sigma$

As in subcase 2b we see, that the nodes $\mathcal{N}(1,0)$ and $\mathcal{N}(1,1)$ are connected in $\mathcal{I}_\gamma \setminus \{\mathcal{B}(1)\}$. Hence $\mathcal{N}(1,0)$ and $\mathcal{N}(1,1)$ are equal or both directly connected to some branch(es) other that $\mathcal{B}(1)$. In either case $\mathcal{N}(1,\sigma)$ must contain at least one endpoint besides $(1,\sigma)$. Since $\gamma_1(\sigma) = x \in B$, we know, that $(1,\sigma) \notin \mathcal{N}_{\partial B}$, so the endpoints in $\mathcal{N}(1,\sigma)$ are mutually joining, and therefore

$$\gamma_n(s) = x \qquad \forall (n,s) \sim (1,\sigma) \tag{B.16}$$

Thus $\exists\ \epsilon \in]0, v]$, such that

$$\gamma_n(\Sigma_s) \subseteq B \qquad \forall (n,s) \sim (1,\sigma) \tag{B.17}$$

where

$$\Sigma_s \doteq \Sigma \cap B_{\mathbf{R}}(s,\epsilon) \qquad s = 0,1$$

Moreover the curve segments $\gamma_n(\Sigma_s)$, $(n,s) \sim (1,\sigma)$, emit from x in a more or less radial fashion, or to be more precise, the functions

$$f_{ns} : [0,\epsilon[\to \mathbf{R} : \varsigma \mapsto \|\gamma_n(s+(-1)^s\varsigma) - \gamma_n(s)\| \qquad (n,s) \sim (1,\sigma)$$

are strictly increasing. Indeed, as $\gamma \in K^{2N}_{l,h,r}$, for any $n \in \{1,\ldots,N\}$ and $(\varsigma,\tau) \in \Upsilon_{nn}(s,s)$ we have

$$\begin{aligned}
&\dot\gamma_n(\varsigma)^T\dot\gamma_n(\tau) \geq \\
&\quad \geq \|\dot\gamma_n(\varsigma)\|\,(\|\dot\gamma_n(\varsigma)\| - \|\dot\gamma_n(\tau) - \dot\gamma_n(\varsigma)\|) \\
&\quad \geq \omega(\omega - \sqrt{2}r|\tau-\varsigma|^H) \\
&\quad > \omega(\omega - \sqrt{2}rv^H) \\
&\quad > 0
\end{aligned}$$

Hence for $(n,s) \sim (1,\sigma)$

$$\begin{aligned}
&\frac{d}{d\varsigma}\frac{f^2_{ns}(\varsigma)}{2} = \\
&\quad = [\gamma_n(s+(-1)^s\varsigma) - \gamma_n(s)]^T\dot\gamma_n(s+(-1)^s\varsigma)(-1)^s \\
&\quad = \int_0^\varsigma \dot\gamma_n(s+(-1)^s\tau)^T\dot\gamma_n(s+(-1)^s\varsigma)\,d\tau \\
&\quad > 0 \qquad \forall\varsigma \in]0,\epsilon[
\end{aligned}$$

Since the functions f_{ns}, $(n,s) \sim (1,\sigma)$, are positive and continuous (at the origin), it therefore follows, that they are strictly increasing. Define the compact set

$$D_* \doteq \bigcup_{n=1}^{N} \gamma_n(\Sigma \setminus (\Sigma_0 \cup \Sigma_1)) \cup \bigcup_{(n,s)\in E_N\setminus\mathcal{N}(1,\sigma)} \gamma_n(\Sigma_s)$$

Then $\complement B \cup D_*$ is closed, and by fact B.2.4 $x \notin \complement B \cup D_*$, so

$$\varrho \doteq \rho(x, \complement B \cup D_*) > 0$$

Let $U \doteq B_{\mathbf{R}^2}(x,\varrho)$. By assumption $C_{l,N}(h,r,\omega,\delta_0,\delta_1,v) \neq \varnothing$, (as it contains γ.) Thus $\omega \leq \sqrt{2}r$, so

$$\epsilon \leq v \leq v^H < \frac{\delta_1\omega}{2\sqrt{2}r} < \frac{1}{2}$$

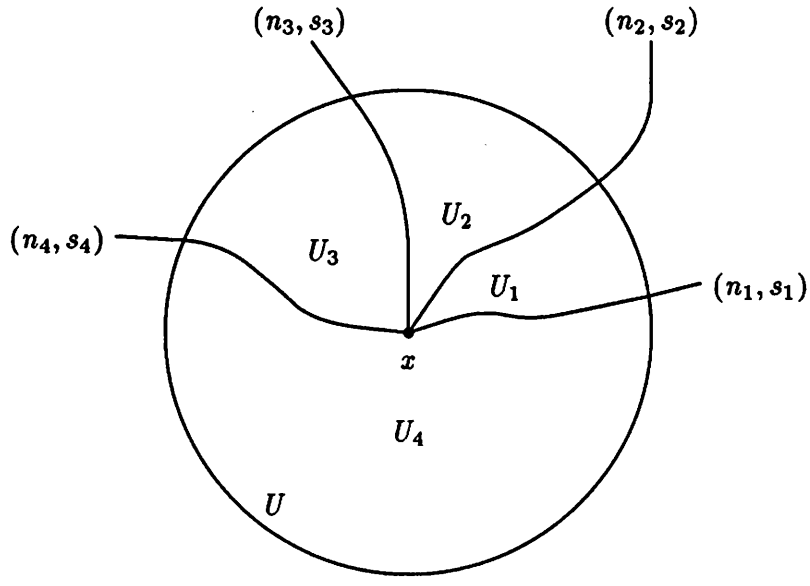


Figure B.6: Curve segments $\gamma_{n_i}(\Sigma_{s_i})$ labeled by (n_i, s_i), $i = 1, \ldots, I$, separating neighborhood U of common endpoint x into $I = 4$ components $U_1, \ldots, U_I$.

whence $B_{\mathbb{R}}(0, \epsilon) \cap B_{\mathbb{R}}(1, \epsilon) = \emptyset$. For each $(n, s) \sim (1, \sigma)$ we therefore conclude that

$$\gamma_n(s + (-1)^s \epsilon) \in D_* \subseteq \complement U$$

and since f_{ns} is strictly increasing, it follows, that $\gamma_n(\Sigma_s)$ intersects the circle ∂U exactly once. Moreover as the curve segments $\gamma_n(\Sigma_s)$, $(n, s) \sim (1, \sigma)$ intersect only at their common endpoint x, their intersections with ∂U are distinct. Let $\mathcal{N}(1, \sigma) = \{(n_i, s_i)\}_{i=1}^I$ be numbered counterclockwise with respect to these intersection points as in the example in figure B.6, and let

$$\Gamma_i \doteq U \cap \gamma_{n_i}(\Sigma_{s_i}) \qquad i = 1, \ldots, I \tag{B.18}$$

By construction the sets $\bigcup_{(n,s)\sim(1,\sigma)} \gamma_n(\Sigma_s)$ and D_* do not share any endpoints, (either $(n, s) \sim (1, \sigma)$ or $(n, s) \in E_N \setminus \mathcal{N}(1, \sigma)$,) so by fact B.2.4 they are disjoint. Since $U \subseteq B$ and $D_* \subseteq \complement U$, fact B.2.7 therefore yields

$$U \cap \partial C_\gamma = U \cap (D_\gamma \setminus D_*) = U \cap \bigcup_{(n,s)\sim(1,\sigma)} \gamma_n(\Sigma_s) = \bigcup_{i=1}^{I} \Gamma_i \tag{B.19}$$

From the properties of the curve segments $\gamma_n(\Sigma_s)$, $(n, s) \sim (1, \sigma)$, we see, that $\Gamma_1, \ldots, \Gamma_I$ are simple curve segments from x to ∂U, intersecting only at their common endpoint x.

Since $I \geq 2$, the set $U \cap \partial C_\gamma$ therefore separates the set $U \setminus \partial C_\gamma$ into I open components $U_1, \ldots, U_I$ according to figure B.6. Now $G \subseteq \complement \partial C_\gamma$, and U is a neighborhood of $x \in \partial G$, so

$$G \cap (U \setminus \partial C_\gamma) = G \cap U \neq \emptyset$$

Thus G intersects at least one of the components $U_1, \ldots, U_I$. Assume without loss of generality, that $G \cap U_1 \neq \emptyset$. We are now ready to proceed with steps similar to those, we used in the proofs of the previous subcases. First of all since $\epsilon \in]0, v]$, and $(n_1, s_1) \bowtie (n_2, s_2)$, by fact B.2.2 $\exists$ a Lipschitz continuous function $\phi :]a, b[\rightarrow \mathbf{R}$, such that

$$\Gamma \doteq \gamma_{n_1}(\Sigma_{s_1}) \cup \gamma_{n_2}(\Sigma_{s_2}) \cong_{F_\phi} \tag{B.20}$$

Next, as $(1, \sigma) \notin \mathcal{N}_{\partial B}$, by interconnection constraint (I2) the number of components of $\mathcal{I}_\gamma$ is unchanged, if the node $\mathcal{N}(1, \sigma)$ is removed. Hence the branch(es) $\mathcal{B}(n_1)$ and $\mathcal{B}(n_2)$ (possibly equal) are connected in $\mathcal{I}_\gamma \setminus \{\mathcal{N}(1, \sigma)\}$, which in turn implies, that $\exists$ a closed path in $\mathcal{I}_\gamma$, in which $\mathcal{N}(1, \sigma)$ figures exactly once. Moreover from figure B.6 it is clear, that this path can be chosen, so that its first and last links are $(\mathcal{B}(n_2), \mathcal{N}(n_2, 1 - s_2)) \rightarrowtail (n_2, 1 - s_2)$ and $(\mathcal{B}(n_1), \mathcal{N}(1, \sigma)) \rightarrowtail (n_1, s_1)$ respectively. Thus by proposition B.2.5 $\exists$ a simple closed (sub)path $\langle \mathcal{L}_q \rangle_{q=1}^{Q} \rightarrowtail \langle (p_q, t_q) \rangle_{q=1}^{Q}$ in $\mathcal{I}_\gamma$, such that $(p_1, t_1) = (n_2, 1 - s_2)$ and $(p_Q, t_Q) = (n_1, s_1)$. Let as before P be the union of the edge segments corresponding to the branch(es) of this subpath. Then by (B.16), (B.17) and (B.20) Γ is a connected subset of $B \cap P$. As in subcase 3a we therefore conclude, that $\exists$ a simple curve S, such that

$$\Gamma \subseteq S \subseteq \overline{B} \cap P \subseteq \overline{B} \cap D_\gamma \tag{B.21}$$

and which separates $B \setminus S$ into two components Ω_1 and Ω_2 with the same properties as those considered in the subcases 2b and 3a. In particular $\Omega_1 \supseteq G$. Just as in subcase 3a we then find that

$$x \in \Gamma \subseteq \partial \Omega_1 \subseteq \partial C_\gamma \tag{B.22}$$

Finally it remains to prove, that $\rho(x, \Omega_1 \cap \partial C_\gamma) > 0$. This can be done by showing, that $U \cap \Omega_1 \cap \partial C_\gamma = \emptyset$. Since $x \in \Gamma_1$, from (B.19) we see that

$$U = (U \cap \partial C_\gamma) \cup (U \setminus \partial C_\gamma) = \bigcup_{i=1}^{I} \Gamma_i \cup \bigcup_{i=1}^{I} U_i = \Gamma_1 \cup \Gamma_2 \cup U_1 \cup V$$

where

$$V \doteq \bigcup_{i=3}^{I} (\Gamma_i \setminus \{x\}) \cup \bigcup_{i=2}^{I} U_i$$

For $\Gamma_1 \cup \Gamma_2$ (B.18), (B.20) and (B.21) imply that

$$\Gamma_1 \cup \Gamma_2 \subseteq S = \partial\Omega_1 \cap \partial\Omega_2 \tag{B.23}$$

Next for the component U_1 of $U \setminus \partial C_\gamma$ from (B.22) we obtain

$$U_1 \subseteq U \setminus \partial\Omega_1 \subseteq B \setminus S$$

The connected set U_1 must therefore be contained in one of the two components of $B \setminus S$. Moreover

$$\Omega_1 \cap U_1 \supseteq G \cap U_1 \neq \varnothing$$

so it must be the case that

$$U_1 \subseteq \Omega_1 \tag{B.24}$$

The remaining part V of U can be treated in a similar way: By our definition of a simple closed path in an interconnection graph, it follows that

$$\{(p_q, t_q)\}_{q=1}^{Q} \cap \mathcal{N}_Q = \{(p_Q, t_Q)\} = \{(n_1, s_1)\}$$

and hence that

$$\{(p_q, 1-t_q)\}_{q=1}^{Q} \cap \mathcal{N}_Q = \{(p_1, 1-t_1)\} = \{(n_2, s_2)\}$$

Thus

$$\Big(\{p_q\}_{q=1}^{Q} \times \{0,1\}\Big) \cap \mathcal{N}_Q = \{(n_1, s_1), (n_2, s_2)\}$$

Since $\mathcal{N}_Q = \mathcal{N}(p_Q, t_Q) = \mathcal{N}(n_1, s_1) = \mathcal{N}(1, \sigma)$, we therefore conclude that

$$n_i \notin \{p_q\}_{q=1}^{Q} \qquad i = 3, \ldots I$$

Hence by fact B.2.4 and (B.21)

$$\Gamma_i \setminus \{x\} \subseteq \gamma_{n_i}(\Sigma^\circ) \subseteq \complement P \subseteq \complement S \qquad i = 3, \ldots, I$$

from which it follows, that $V \subseteq U \setminus S \subseteq B \setminus S$. Since V is (path-)connected, it must then be contained in Ω_1 or Ω_2. However U is a neighborhood of x, and by (B.21) and (B.22) $x \in S \subseteq \partial\Omega_2$, so (B.23) and (B.24) imply, that $\Omega_2 \cap V = \Omega_2 \cap U \neq \varnothing$, whence

$$V \subseteq \Omega_2 \tag{B.25}$$

From (B.23) – (B.25) we now find, that $U \cap \Omega_1 = U_1 \subseteq \complement \partial C_\gamma$, and therefore $\Omega_1 \cap \partial C_\gamma \subseteq \complement U$. Hence

$$\rho(x, \Omega_1 \cap \partial C_\gamma) \geq \rho(x, \complement U) = \varrho > 0$$

This completes the proof of case 3. ■

We are now ready to complete the proof of fact 3.11.7.

Proof of Fact 3.11.7: We have to show, that the image segmentation γ (satisfying the conditions stated in the beginning of this section) is admissible. By theorem 3.8.2 it is sufficient to show, that every component of the continuity set C_γ is a Lipschitz domain. We will show this, by verifying, that the Lipschitz domain characterizing conditions given on page 78 are satisfied at an arbitrary point on the boundary of an arbitrary component of C_γ. Thus let G be a component of C_γ, and let $x \in \partial G$. By fact B.2.8 and the Jordan Curve Theorem $\exists$ a curve segment Γ, a Lipschitz continuous function $\chi :]a, b[\rightarrow \mathbf{R}$ and an open set $\Omega \subseteq \mathbf{R}^2$, such that

(i) $\Gamma \cong f_\chi$

(ii) $\partial\Omega = \partial\complement\overline{\Omega}$ is a simple closed curve.

(iii) $\Omega \supseteq G$

(iv) $x \in \Gamma \subseteq \partial\Omega \subseteq \partial C_\gamma$

(v) $\rho(x, \Omega \cap \partial C_\gamma) > 0$

Let T be the unique rigid coordinate transformation, for which $T(f_\chi) = \Gamma$, and define the map

$$\mathrm{X} :]a, b[\times \mathbf{R} \rightarrow \mathbf{R}^2 : y \mapsto T(y_1, \chi(y_1) + y_2)$$

By proposition 3.5.2 X is a homeomorphism, so $\exists$ a unique $y_x \in]a, b[\times \{0\}$, such that $\mathrm{X}(y_x) = x$. Since $\chi :]a, b[\rightarrow \mathbf{R}$ is continuous, f_χ and hence Γ are connected curve segments, which do not contain their endpoints. Thus $\partial\Omega \setminus \Gamma \not\ni x$ is closed, and therefore $\rho(x, \partial\Omega \setminus \Gamma) > 0$. Since $\rho(x, \Omega \cap \partial C_\gamma) > 0$ as well, this implies, that $\exists\ d > 0$, such that $Q \doteq]-d, d[^2 \subseteq (]a, b[\times \mathbf{R}) - y_x$, and $U \doteq \mathrm{X}(y_x + Q)$ is an open neighborhood of x, which intersects neither $\partial\Omega \setminus \Gamma$ nor $\Omega \cap \partial C_\gamma$. Define the three sets

$$U_{\pm} \doteq \mathrm{X}(\mathbf{R}^2_{\pm} \cap (y_x + Q)) \tag{B.26a}$$

$$U_0 \doteq \mathrm{X}(\mathbf{R}^2_0 \cap (y_x + Q)) \ni x \tag{B.26b}$$

Then

$$\mathrm{X}^{-1}(U_{\pm}) = \mathbf{R}^2_{\pm} \cap (y_x + Q)$$

and

$$\mathrm{X}^{-1}(\Gamma) =]a, b[\times\{0\}$$

are disjoint. Hence $\Gamma \cap U_{\pm} = \varnothing$, and therefore

$$\partial\Omega \cap U_{\pm} = (\partial\Omega \setminus \Gamma) \cap U_{\pm} \subseteq (\partial\Omega \setminus \Gamma) \cap U = \varnothing$$

Since $U_{\pm}$, the continuous image of a connected set, is connected, this means, that either $U_{\pm} \subseteq \Omega$ or $U_{\pm} \subseteq \complement\overline{\Omega}$. Moreover

$$U_0 \subseteq \mathrm{X}(\mathbf{R}^2_0 \cap (]a, b[\times\mathbf{R})) = \mathrm{X}(]a, b[\times\{0\}) = \Gamma \subseteq \partial\Omega$$

Since U is a neighborhood of $x \in \partial\Omega = \partial\complement\overline{\Omega}$, and Ω is open, this implies that

$$\Omega \cap (U_+ \cup U_-) = \Omega \cap (U \setminus U_0) \supseteq \Omega \cap U \cap \complement\partial\Omega = \Omega \cap U \neq \varnothing$$

and likewise

$$\complement\overline{\Omega} \cap (U_+ \cup U_-) \supseteq \complement\overline{\Omega} \cap U \neq \varnothing$$

Hence one of the sets $U_{\pm}$ is contained in Ω, and the other one is contained in $\complement\overline{\Omega}$. Suppose $U_+ \subseteq \Omega$. Then

$$U_- \subseteq \complement\overline{\Omega} \subseteq \complement\overline{G} \tag{B.27}$$

For U_+ we note the following: Since U is a neighborhood of $x \in \partial G$,

$$G \cap U_+ = G \cap \Omega \cap U = G \cap U \neq \varnothing$$

Moreover

$$\partial G \cap U_+ \subseteq \partial C_\gamma \cap \Omega \cap U = \varnothing$$

Since U_+ is connected, we therefore conclude that

$$U_+ \subseteq G \tag{B.28}$$

For U_0 finally, as X is a homeomorphism, from (B.27) and (B.28) we have that

$$U_0 \subseteq \overline{U_+} \cap \overline{U_-} \subseteq \overline{G} \cap \overline{\complement G} \subseteq \partial G \tag{B.29}$$

Define the Lipschitz continuous function

$$\phi :]-d, d[\rightarrow \mathbf{R} : y_1 \mapsto \chi(y_{x1} + y_1)$$

and let

$$\Phi : Q \rightarrow \mathbf{R}^2 : y \mapsto T_+(y_1, \phi(y_1) + y_2)$$

where T_+ is the coordinate transformation defined by

$$T_+(y) \doteq T(y_x + y) \qquad y \in \mathbf{R}^2$$

Then

$$\begin{aligned} \Phi(y) &= \\ &= T_+(y_1, \chi(y_{x1} + y_1) + y_2) \\ &= T(y_{x1} + y_1, \chi(y_{x1} + y_1) + y_{x2} + y_2) \\ &= \mathrm{X}(y_x + y) \qquad \forall y \in Q \end{aligned}$$

Hence

$$\Phi(\mathbf{R}^2_{\pm} \cap Q) = \mathrm{X}(y_x + (\mathbf{R}^2_{\pm} \cap Q)) = U_{\pm}$$

and

$$\Phi(\mathbf{R}^2_0 \cap Q) = \mathrm{X}(y_x + (\mathbf{R}^2_0 \cap Q)) = U_0$$

From (B.26) – (B.29) it then follows that

$$\Phi(\mathbf{R}^2_+ \cap Q) \subseteq G \tag{B.30a}$$

$$\Phi(\mathbf{R}^2_- \cap Q) \subseteq \complement\overline{G} \tag{B.30b}$$

$$x \in \Phi(\mathbf{R}^2_0 \cap Q) \subseteq \partial G \tag{B.30c}$$

Suppose instead, that $U_+ \not\subseteq \Omega$. Then $U_+ \subseteq \complement\overline{\Omega}$ and $U_- \subseteq \Omega$, so the roles of U_+ and U_- in the proof above are reversed. In this case we define

$$\phi :]-d, d[\rightarrow \mathbf{R} : y_1 \mapsto -\chi(y_{x1} - y_1)$$

and

$$\Phi : Q \to \mathbf{R}^2 : y \mapsto T_-(y_1, \phi(y_1) + y_2)$$

where T_- is the coordinate transformation defined by

$$T_-(y) \doteq T(y_x - y) \qquad y \in \mathbf{R}^2$$

Again ϕ is Lipschitz continuous, and

$$\begin{aligned}\Phi(y) &= \\ &= T_-(y_1, -\chi(y_{x1} - y_1) + y_2) \\ &= T(y_{x1} - y_1, \chi(y_{x1} - y_1) + y_{x2} - y_2) \\ &= \mathrm{X}(y_x - y) \qquad \forall y \in Q\end{aligned}$$

Since $-Q = Q$, we now obtain

$$\Phi(\mathbf{R}^2_{\pm} \cap Q) = \mathrm{X}(y_x - (\mathbf{R}^2_{\pm} \cap Q)) = \mathrm{X}(y_x + (\mathbf{R}^2_{\mp} \cap Q)) = U_{\mp}$$

and

$$\Phi(\mathbf{R}^2_0 \cap Q) = \mathrm{X}(y_x - (\mathbf{R}^2_0 \cap Q)) = \mathrm{X}(y_x + (\mathbf{R}^2_0 \cap Q)) = U_0$$

Hence the conditions (B.30a) – (B.30c) are again satisfied, and the fact follows. ■

Appendix C

Initial Edge Finder Operation

The initial edge finder was introduced in section 4.2 as the first subroutine of the global curve-represented edge detector there presented. As part of that presentation we also discussed the initial edge finder's basic output data structure, which was noted to consist of a vertex list and a spline list. In this appendix we take a look at the mechanism by which that data structure is generated.

Throughout our description of the initial edge finder operation we will frequently consider distances between points in the image domain. All such distances are ∞-norm distances. In other words, the distance between the two points $x, y \in \mathbf{R}^2$ is given by $|x_1 - y_1| \vee |x_2 - y_2|$. By a slight abuse of language the distance between two pixels will always be understood to mean the be distance between the sites of those pixels.

At a high level of description the initial edge finder performs the following sequence of operations:

1. Detect preliminary edges and junctions.
2. Select initial junctions.
3. Form splines and select initial intermediate vertices.
4. Compute (spline) type variables.

The first of these four steps, which is by far the most involved, is a pure preprocessing stage; it does not generate any of the initial edge finder output. The second step generates part of the vertex list. The third step generates most of the spline list and the remaining part of

the vertex list. The fourth step finally completes the spline list. We will next take a closer look at each of these processing stages.

C.1 Preliminary Edges and Junctions

Before the image segmentation configuration and the initial control vertices can be found, a preliminary set of edges and junctions must be obtained with some external *preliminary edge detector.* Since the final location of the edges will be determined by the global steepest descent procedure, the preliminary edge detector does not need to be very sophisticated; a simple local edge detector is good enough. It is, however, important that the output data structures are well suited for selecting the image segmentation configuration and the initial control vertices. For this reason we wrote our own preliminary edge detector. The procedure is built around a simple contour tracing scheme, which traces smooth contours of high brightness gradient magnitude (of the original image function ζ). It does not produce high quality edges, but it is well tailored for its purpose of helping to generate a starting point for the steepest descent procedure.

C.1.1 Data Structures

The primary output from the preliminary edge detector consists of a list of *preliminary edges* and a list of *preliminary junctions.* Each preliminary edge is by itself a list of (sites of) eight-connected pixels located along one of the traced contours. A preliminary junction is just a single pixel (site), which is always the first or the last of one of the preliminary edges.

As a secondary output the preliminary edge detector generates an $X_1 \times X_2$ array of so called *edge status records*—one for each pixel in the (original) image. Each edge status record holds three pieces of information:

1. If the pixel belongs to one of the preliminary edges, it is marked (on the edge status record) as an *edge pixel.*

2. If the pixel is within a certain distance r_e, referred to as the *edge zone radius*, from an edge pixel, then it is marked as an *edge zone pixel.*

3. If the pixel is within a certain distance r_j, referred to as the *junction zone radius*, from one or more preliminary junctions, then the edge status record points to one of

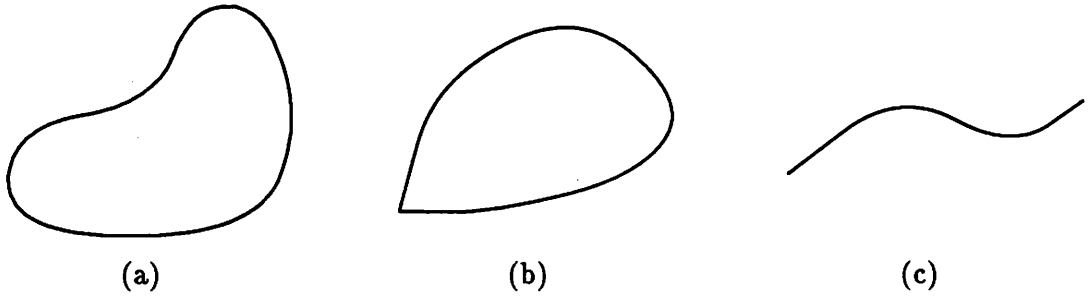


Figure C.1: Contours. (a) Smoothly closed. (b) Nonsmoothly closed. (c) Open.

these preliminary junctions. The pixel is moreover said to belong to the *(junction) zone* of that preliminary junction.

While the junction zone and edge information is used by the later stages of the initial edge finder, the edge zone information is only used internally by the preliminary edge detector—primarily as part of the mechanism for forming preliminary junctions. The edge zone also serves to prevent multiple preliminary edges along "wide contours" of high brightness gradient magnitude. Thinning of the preliminary edges is therefore unnecessary.

The lists constituting the preliminary edges can be circular or linear. The circular lists are referred to as *closed edges*. They correspond to traced contours that close smoothly upon themselves. Such a contour is depicted in figure C.1 (a). A closed edge may or may not intersect the zone of any preliminary junction. If it does, it will later contribute to the generation of a number of open splines. If not, it will result in a single closed spline. The linear lists are referred to as *open edges*. They correspond to traced contours that close nonsmoothly upon themselves or to open contours. These kinds of contours are shown in figure C.1 (b) and (c). An open edge both begins and ends with junction zone pixels. In the rare case of a nonsmoothly closed contour these pixels will belong to the zone of the same preliminary junction.

C.1.2 Preliminary Edge Detector Operation

Edge Initiation Condition

The outermost loop of the preliminary edge detector scans the original image column by column. At each pixel outside the currently detected edge zone it computes

an estimate of the brightness gradient $\nabla\zeta$ using a 3×3 window. If the magnitude of this gradient exceeds a certain fixed threshold t_i referred to as the *edge initiation threshold*, the scanning is temporarily interrupted and a decision is made to begin the tracing of a new contour. In order to suppress short spurious preliminary edges caused by noise and other disturbances, the threshold t_i should be chosen a bit higher than the expected (lowest) value of $\|\nabla\zeta^T\|$ along the contours of interest. The influence of short spurious preliminary edges is further repressed by the spline formation mechanism to be discussed later in this appendix.

"Contour Center" Location

Due to smooth shading as well as blurring and sampling in the image formation process many "contours" of high brightness gradient magnitude are not all that sharp, but rather a few pixels wide. Consequently the pixel that triggers the tracing of a new contour might not be very representative for the actual edge. Before starting tracing a new contour, the preliminary edge detector therefore attempts to locate the "center of the contour" by searching along a line parallel to the brightness gradient at the triggering pixel. The search begins at the triggering pixel and continues pixel by pixel in the direction of increasing brightness gradient magnitude. It terminates as soon as the first local maximum of the brightness gradient magnitude is encountered. The pixel site of this maximum is recorded as the first pixel of the new preliminary edge.

Edge Extension Outside the Edge Zone

After the first pixel of a new preliminary edge has been recorded, the preliminary edge detector begins tracing a smooth contour of high brightness gradient magnitude. The tracing, which of course starts out at the first pixel of the new preliminary edge, is carried out by making a sequence of short *jumps* from one pixel to another. The mechanism for selecting these jumps will be described shortly. Each time a jump has been made, (the list constituting) the new preliminary edge is extended by a sequence of eight-connected pixels along the line segment joining the centers of the two pixels between which the jump took place.

As the contour tracing proceeds and the new preliminary edge is being extended, the edge status array is also being updated. The extension pixels themselves are marked

as edge pixels. In addition all the pixels within the distance of an edge zone radius r_e from any of the extension pixels are marked as edge zone pixels. The edge zone pixels are of course just the pixels inside the union of all the $(2r_e + 1) \times (2r_e + 1)$ windows centered at any of the extension pixels. The main purpose of collecting this edge status information is to facilitate the detection of the events when the new preliminary edge is being extended into (eight-connected) contact with or into the vicinity of the pixels of a previously detected preliminary edge. Since the extension pixels are always in contact with and/or in the vicinity of the most recently detected pixel(s) of the new preliminary edge, the updates of the edge status array must be delayed, or else their main purpose will fail. The most simple minded approach would be to avoid updating the edge status array while a contour is being traced. (Computationally this would not cause any problem. One would just have to retrace each new preliminary edge upon the completion of its detection.) The simple minded method would, however, prevent the desired detection of the event when the new preliminary edge is being extended into (the vicinity off) its own tail. The necessary delay of the edge status array updates is therefore implemented by shifting the extension pixels through a short FIFO buffer before marking them and their neighboring pixels as edge and edge zone pixels respectively.

The Jump Mechanism

Most of the properties of the preliminary edges are determined by the mechanism that selects the jumps by which the contours are traced. In our description of this mechanism the pixel to which the previous jump was made, that is the most recently recorded pixel of the new preliminary edge, will be called the *(present) frontier pixel.* The pixel to which the next jump is going to be made will be called the *(present) destination pixel.* The direction of the vector from the center of the frontier pixel to the center of the destination pixel will be referred to as the *jump direction.* The distance between the frontier and the destination pixels will be called the *jump distance.* Figure C.2 shows a window of pixel sites in which a jump of distance three is taking place. The most recently recorded pixel sites of the new preliminary edge are highlighted with wide filled circles. The sites of the extension pixels are marked with empty circles. The jump and its direction are indicated by the arrow pointing from the frontier pixel site F to the destination pixel site D.

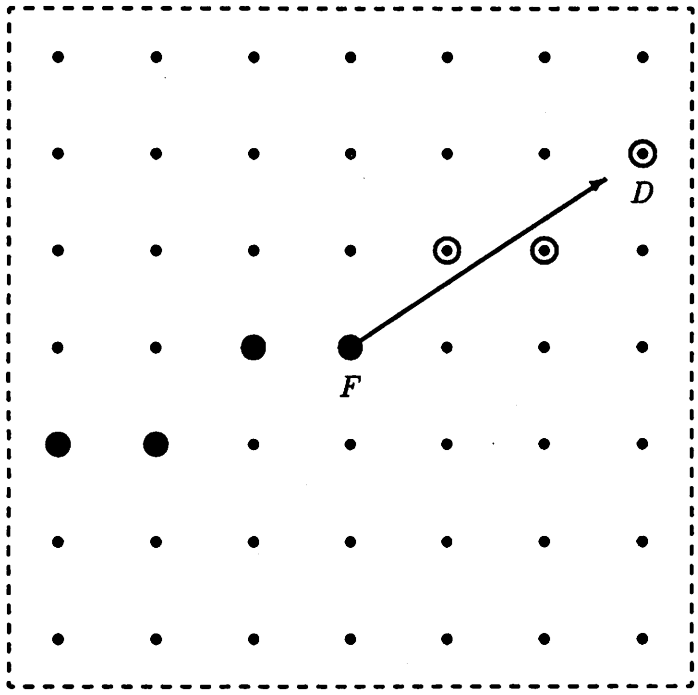


Figure C.2: Contour tracing jump from frontier pixel F to destination pixel D.

Brightness Gradient Requirement. The main objective of the jump mechanism is that the brightness gradient magnitude $\|\nabla\zeta^T\|$ remains high along the traced contour. This is achieved by insisting that $\|\nabla\zeta^T\|$ at the destination pixel exceeds a certain threshold t_t referred as the *edge termination threshold.* If no such destination pixel can be found, the contour tracing is terminated at the present frontier pixel.

Possible Destination Pixels. For reliable detection of L-junctions, of which a couple of examples are depicted in figure C.3, it is important that the preliminary edges are kept reasonably smooth. This is accomplished by restricting the set of possible destination pixels so that the directions of successive jumps do not differ too much. The set of possible destination pixels for each jump thus depends on the previous jump direction. (If no previous jump has been made, a direction normal to the brightness gradient is substituted for the previous jump direction.)

Consider the frontier pixel site F and the previous jump direction indicated by the arrow in figure C.4. The pixel sites at equal distances from the frontier pixel site are joined by the dashed squares S_d, $d = 1, 2, \ldots$. On each of these squares there is one pixel site that

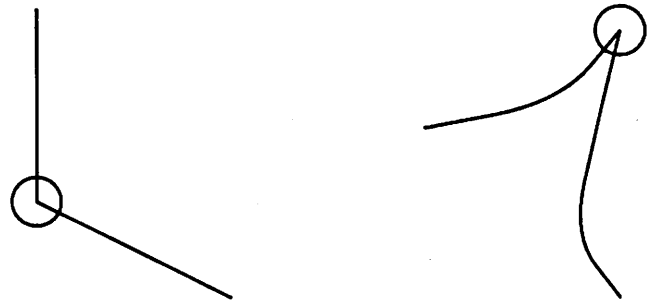


Figure C.3: L-junctions.

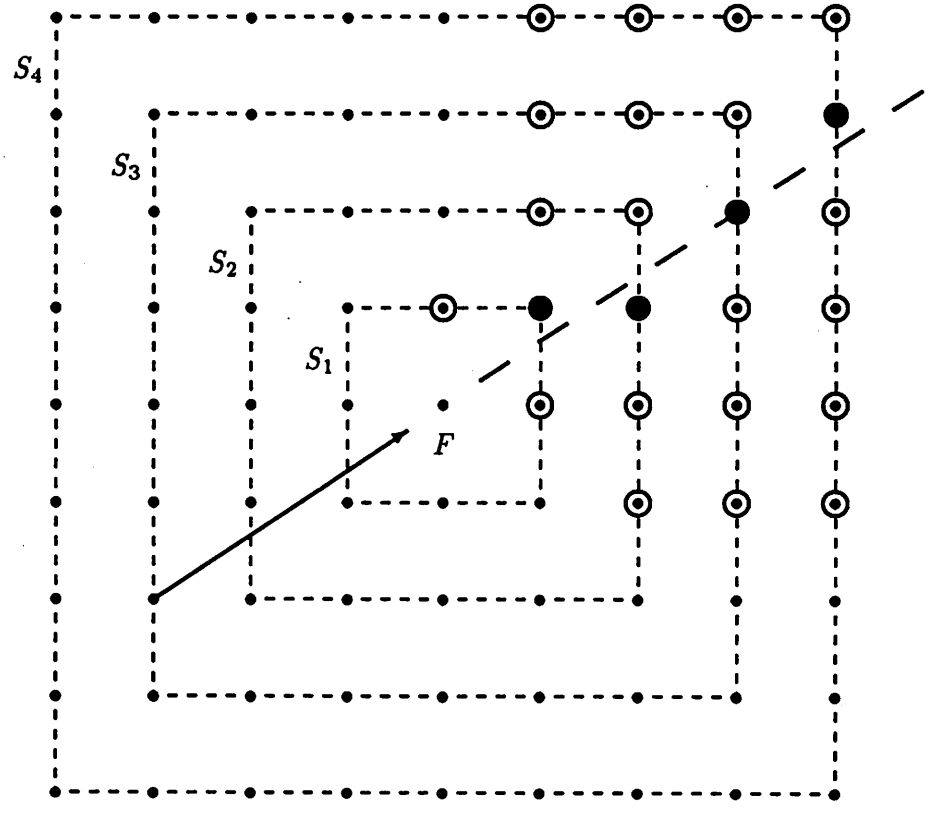


Figure C.4: Possible destination pixel sites (wide circles). Filled wide circles indicate extrapolation sites.

is closest to the dashed ray extrapolating the previous jump direction arrow. These pixel sites known as *extrapolation sites* are indicated by wide filled circles. For some previous jump directions (other than that in the figure) two pixel sites on the same square would both qualify as extrapolation sites. In such a situation an arbitrary choice is made. There is thus always exactly one extrapolation site on each square. The set of possible destination pixel sites at distance d from the frontier pixel site consists by definition of the unique extrapolation site on S_d together with its d closest neighbors (indicated by wide empty circles) in each of the two directions along S_d. At each distance $d \in \mathbb{N}$ from the frontier pixel there are thus $2d + 1$ possible destination pixels.

A possible destination pixel at which $\|\nabla\zeta^T\| > t_t$, is said to qualify as a *destination candidate.* Hypothetical jumps to destination candidates are referred to as *jump candidates.*

Noise Suppression. Short gaps in the contours of high brightness gradient magnitude are most likely caused by noise, insignificant small scale information or the lack of context dependent adaptation of the edge termination threshold. In order to bridge such undesirable gaps the jump distances are allowed to vary from jump to jump. The possible jump distances range from a minimum of unity to a maximum set by a constant parameter $d_m \in \mathbb{N}$. If any destination candidate can be found within a distance d_m from the frontier pixel, a jump is made. There are no requirements on the brightness gradient at the extension pixels that are filled in along the line segment joining the frontier pixel and the destination pixel.

Jump Distance Considerations. Long jumps have two significant advantages over short jumps. Most importantly, the long jumps offer a richer variety of jump directions. They therefore give the preliminary edges a less "boxy" appearance. Secondly, the long jumps require slightly fewer brightness gradient evaluations and comparisons per extension pixel. The jump mechanism therefore favors long jumps by always selecting a jump candidate of maximal distance $d_j \leq d_m$. Consequently the possible destination pixels furthest away from the frontier pixel are examined first. The possible destination pixels closer than d_j from the frontier pixel need obviously not be found at all.

Contour Direction. As a final consideration the jump direction should not only warrant smoothness of the preliminary edge, but preferably also approximate the underlying contour direction. When there are more than one jump candidates of the same distance to

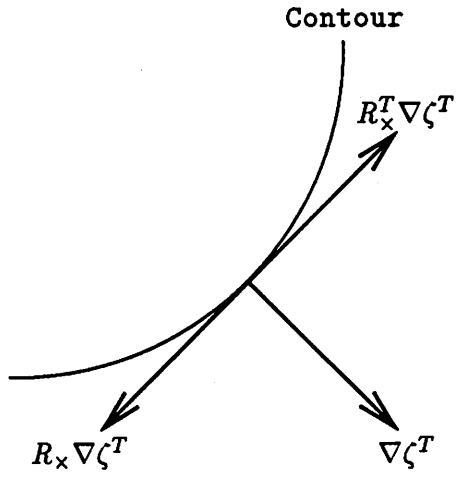


Figure C.5: Contour directions.

choose from, the jump mechanism therefore selects the jump candidate whose direction best matches the contour direction. When first tracing a contour in the original direction, the contour direction is defined to be that of the *left normal vector* $R_\times^T \nabla\zeta^T$ of the brightness gradient. The contours will, as we soon shall see, often be traced in the reverse direction as well. The contour direction is then given by the (opposite) *right normal vector* $R_\times \nabla\zeta^T$. The orientations of the contour directions relative to the brightness gradient are shown in figure C.5.

A Jump Selection Example. While the general description of the jump mechanism above is quite complete, an example might still be clarifying. Figure C.6 (a) illustrates a situation similar to that in figure C.4. As before the possible destination pixels are indicated by wide (filled or empty) circles. The wide filled circles, however, have a different meaning. They here highlight the destination candidates.

Suppose $d_m = 4$. The jump mechanism first finds and examines the possible destination pixels on the square $S_{d_m} = S_4$. Since there are no destination candidates on S_4, the search continues on S_3, where two destination candidates are found. The jump mechanism now selects the jump candidate whose direction is closest to the contour direction $R_\times^T \nabla\zeta^T$. The resulting extension pixels are indicated by the wide circles in figure C.6 (b). The circle at the destination pixel site is filled. Had there not been any destination candidates on S_3, the search would have continued on S_2 and so on.

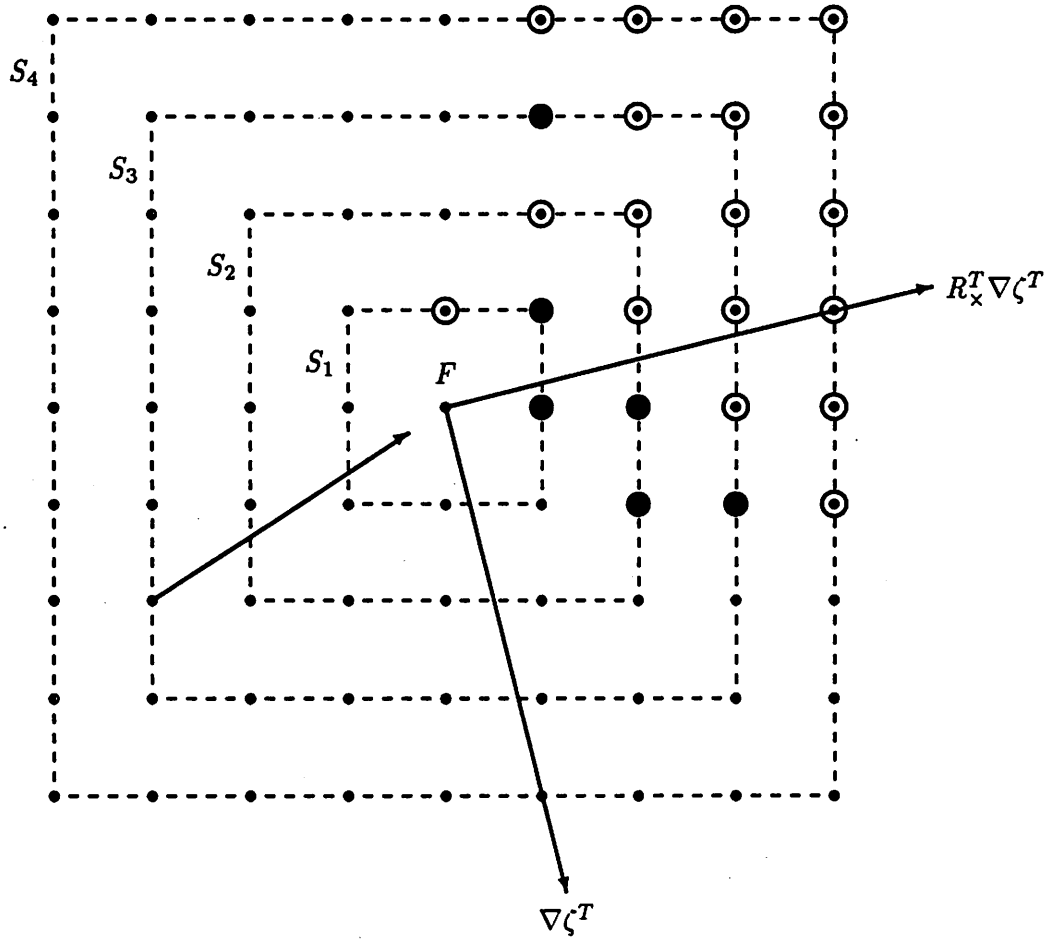


Fig. C.6: (a)

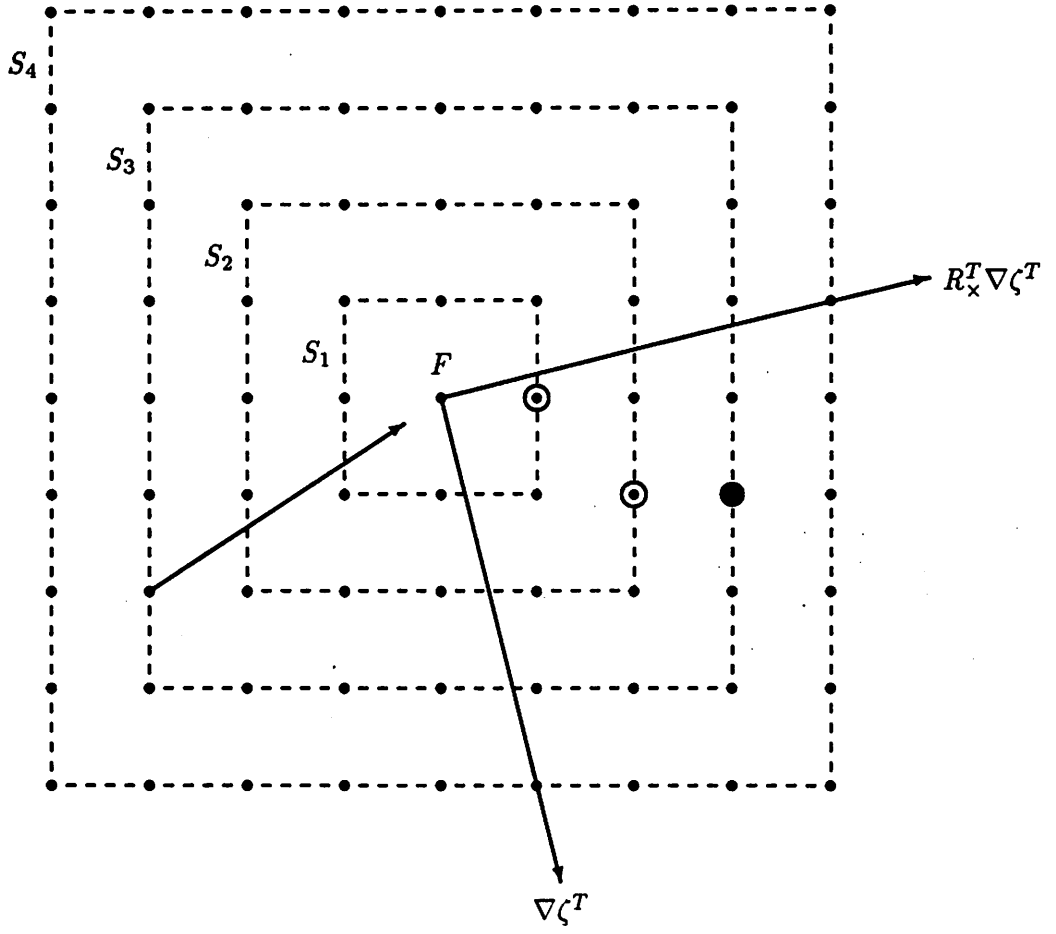


Fig. C.6: (b)

Figure C.6: Jump selection example. (a) Sites of possible destination pixels (wide circles) and destination candidates (wide filled circles). (b) Sites of extension pixels (wide circles) and destination pixel (wide filled circle).

Edge Termination Conditions

The edge extension proceeds as described above until one of the following edge termination conditions is satisfied:

1. The new preliminary edge is extended into an area where edge zone has previously been marked on the edge status array.

2. The brightness gradient magnitude drops below the edge termination threshold t_t so that no destination candidates can be found.

In the former case the preliminary edge detector will extend the new preliminary edge into eight-connected contact with (one of) the preexisting edge pixels inside the edge zone. In the latter case the contour tracing is terminated immediately at the present frontier pixel.

Edge Extension Inside the Edge Zone

The extension of a preliminary edge inside the edge zone begins with continued tracing of the contour. The contour tracing procedure is the same as before with the exception of the two following modifications of the jump mechanism:

1. Only jumps of unit distance are considered, (exactly as the edge extension outside the edge zone would be with $d_m = 1$.)

2. The contour direction is not updated, but remains normal to the brightness gradient at the last frontier pixel prior to the entry into the edge zone.

The first modification is meant to improve the localization of the preliminary edges in the neighborhoods of junctions. The second modification prevents the contour direction from getting distorted in the vicinity of another contour.

The contour tracing inside the edge zone terminates when one of the following four conditions is satisfied:

1. The new preliminary edge is extended into eight-connected contact with a *preexisting edge pixel*, that is a pixel that is already marked as an edge pixel on the edge status array.

2. The brightness gradient magnitude drops below the edge termination threshold t_t so that no destination candidates can be found.

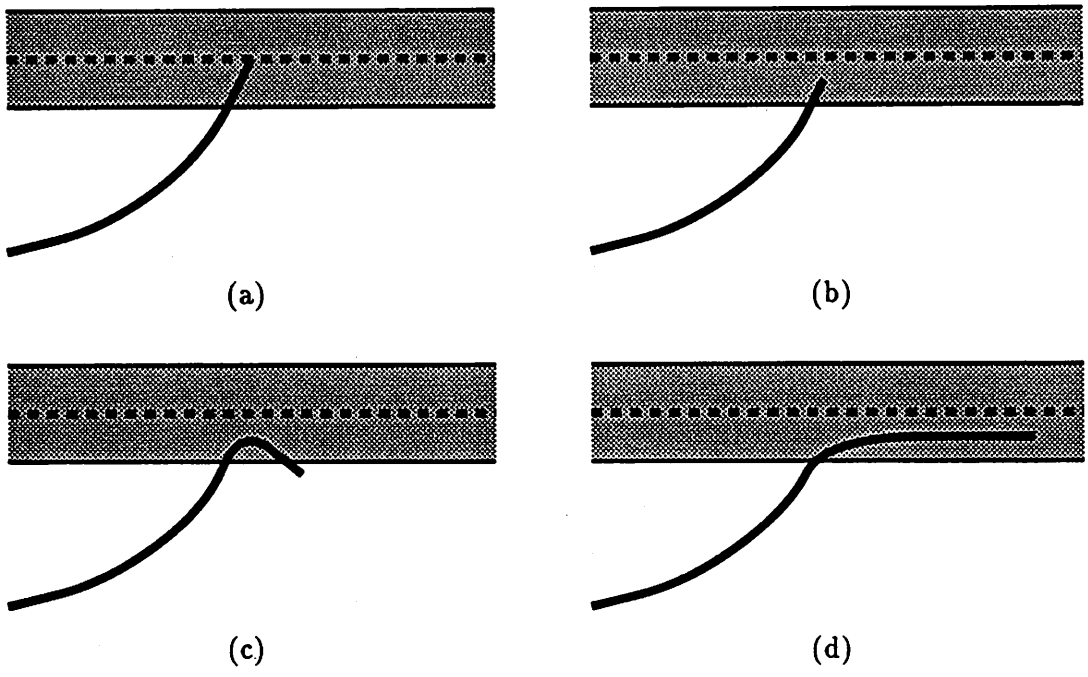


Figure C.7: Extent of new preliminary edge (solid) inside edge zone (shaded) surrounding preexisting edge (dashed) at contour tracing termination according to the conditions 1 (a), 2 (b), 3 (c) and 4 (d) listed in the text.

3. The new preliminary edge is extended back outside the edge zone.

4. The extension of the new preliminary edge inside the edge zone exceeds a certain limit.

These four situations are illustrated by the schematic diagrams in figure C.7. If the first termination condition is satisfied, the goal of bringing the new preliminary edge in eight-connected contact with the preexisting edge (pixels) inside the edge zone is already achieved. In each of the other three cases the short portion of the new preliminary edge that is inside the edge zone, is retraced, and the pixel closest to the preexisting edge pixels identified as the *last contour pixel.* Any pixels that might have been appended to the new preliminary edge after the last contour pixel, are then replaced by a sequence of eight-connected pixels along the line segment joining the last contour pixel with the closest of the preexisting edge pixels. Such a sequence is of course necessarily shorter than the edge zone radius r_e, which is typically set to some small integer value.

Backward Tracing, Edge Closure and Preliminary Junctions

When the contour tracing (followed by possible linear edge extension inside the edge zone) has come to an end, the preliminary edge detector checks whether the contour is open or closed. A contour is considered to be open if the first and last pixels of the corresponding preliminary edge are within a distance of one junction radius r_j from each other, or if this preliminary edge is so short that such a criterion does not make much sense. A contour that is not open is naturally said to be closed.

If the contour is open, the new preliminary edge is also declared to be open, and its last pixel is recorded as a preliminary junction. As always whenever a preliminary junction is recorded, all the pixels in the $(2r_j + 1) \times (2r_j + 1)$ window centered at the preliminary junction are simultaneously marked as junction zone pixels on the edge status array. The contour tracing then resumes where it started, but in the reverse direction. Meanwhile the new preliminary edge is being extended backwards from its (changing) first pixel. The backward tracing process is identical to the forward tracing process described earlier. The termination conditions for the two processes are also identical. When the backward tracing terminates, the first pixel of the new preliminary edge is recorded as a preliminary junction, and the edge status array updated accordingly.

If the contour instead is closed, the preliminary edge detector checks whether it closes smoothly on itself. If it does not, the new preliminary edge is again declared to be open, and its last pixel is recorded as a preliminary junction. The junction zone of this preliminary junction necessarily covers both the first and the last pixels of the new preliminary edge. No preliminary junction is therefore recorded for the first pixel. If the contour on the other hand does close smoothly on itself, the new preliminary edge is declared to be closed, and is then extended (by a few pixels along a short line segment) as necessary so that its first and last pixels become eight-connected. No preliminary junction is recorded in this case. A closed contour, whether smoothly closed or not, is for obvious reasons never traced backwards.

C.2 Initial Junctions

At Y-junctions and arrow-junctions such as those depicted in figure C.8, the preliminary edge detector will ideally form two preliminary junctions at the same pixel. For

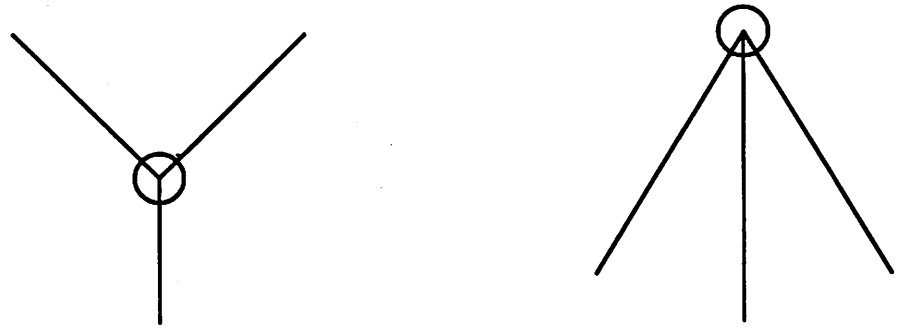


Figure C.8: (a) Y-junction. (b) Arrow-junction.

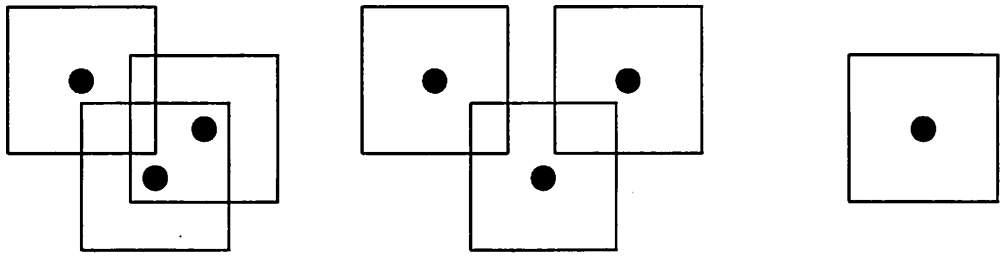


Figure C.9: Three clusters of preliminary junctions with overlapping junction zones.

more complicated junctions there might be even more preliminary junctions with junction zones lying on top of each other. In practice, however, the multiple preliminary junctions will not coincide perfectly. In order to recognize such collections of close preliminary junctions as single junctions, every cluster of preliminary junctions whose junction zones overlap is designated exactly one initial junction. Figure C.9 depicts three examples of such clusters. The preliminary junctions are shown as filled circles. Their junction zones are indicated by the surrounding squares. Each initial junction is given the location of the arithmetic mean—or equivalently the geometric center—of its constituent preliminary junctions. For the later purpose of forming splines each initial junction is also given a junction zone equal to the union of the junction zones of its constituents. The end result is a list of initial junctions whose associated junction zones are mutually disjoint. This list forms the beginning of the vertex list, which will be completed by the subsequent stages of the initial edge finder.

C.3 Splines and Initial Intermediate Vertices

C.3.1 Preliminary Edge Segment Formation

Although preliminary junctions are only formed at the end pixels of the open edges, most preliminary edges—both closed and open—will pass through many zones of preliminary junctions formed at the end pixels of other preliminary edges. In order to form splines that define an image segmentation configuration as described in section 4.2, the preliminary edges are retraced and broken up at each passage through any of the initial junction zones, so that a collection of *preliminary edge segments* is obtained.

As we recall, the open edges do always begin and terminate inside some initial junction zone(s). They are all thus partitioned into one or more preliminary edge segments, which also have this property. Such preliminary edge segments will be referred to as *open edge segments.*

The closed edges on the other hand do not necessarily intersect any junction zones at all, and even if they do, their first and last pixels may not be inside any of the junction zones. The retracing of a closed edge therefore begins with a search for an intersection with an initial junction zone. From the moment a pixel inside such an intersection is found, the closed edge is processed exactly as an open edge, which both begins and terminates at that pixel. The closed edge is in this case thus partitioned into one or more open edge segments. If the closed edge does not intersect any of the initial junction zones, it is "partitioned" into exactly one preliminary edge segment. Such a preliminary edge segment will be referred to as a *closed edge segment.*

C.3.2 Sampling

Every time the retracing procedure isolates a new preliminary edge segment, a new spline is formed and appended to the spline list. A closed edge segment generates a closed spline. An open edge segment generates an open spline. End vertices of open splines are always among the junctions, for which initial values have already been selected and recorded in the vertex list. Intermediate vertices on the other hand constitute new independent control vertices, which must be appended to the vertex list. Initial values for these control vertices are gathered from samples of the preliminary edge segments. The sampling is governed by an integer parameter i_m, known as the *maximum sampling interval,*

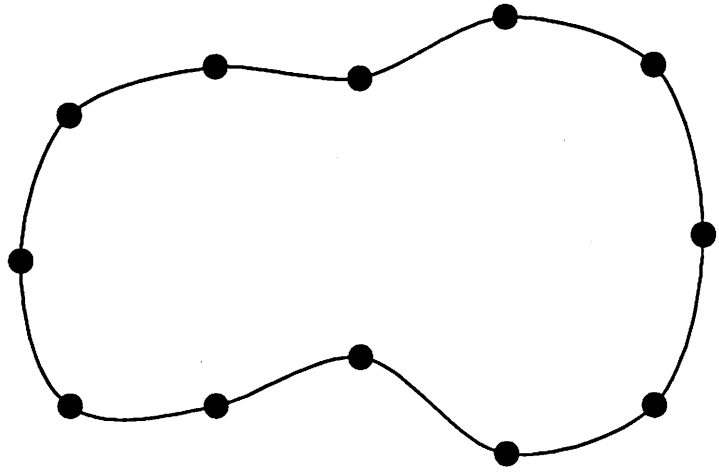


Figure C.10: Samples (indicated by filled circles) of closed edge segment.

which dictates the maximum number of preliminary edge (segment) pixels that are allowed between successive samples. Besides assisting the selection of the initial control vertices, the sampling therefore also determines the length of the control vertex sequence associated with each spline.

Closed Edge Segments

A closed edge segment is simply sampled evenly at every i_cth pixel site as indicated in figure C.10. Where on the closed edge segment the sampling begins, does not matter. Normally $i_c = i_m$, but for very short closed edge segments the parameter i_c is reduced as necessary so that at least three samples are obtained. The sampling thus yields a sequence $\langle u_m \rangle_{m=0}^{M-1}$ of pixel sites for some $M \geq 3$. This sequence is appended to the vertex list, and a new closed spline with control vertex sequence

$$\langle v_m \rangle_{m=0}^{M+2} = \langle u_0, \ldots, u_{M-1}, u_0, u_1, u_2 \rangle$$

is appended to the spline list.

Open Edge Segments

An open edge segment begins and terminates, as we recall, always inside the zones W_{j_b} and W_{j_t} of some initial junction(s) w_{j_b} and w_{j_t} respectively. The samples of such a

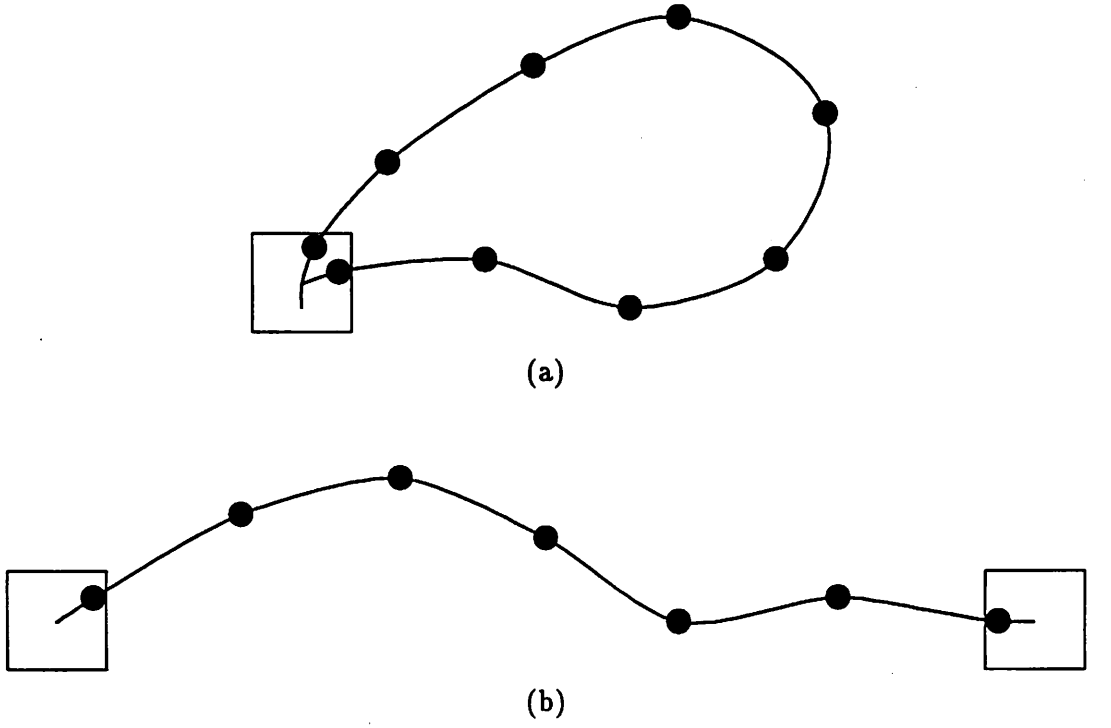


Figure C.11: Samples (indicated by filled circles) of open edge segments beginning and terminating inside initial junction zone(s) (indicated by squares).

preliminary edge segment always include the last pixel site before its exist from W_{j_b} and the first pixel site after its entry into W_{j_t}. Between these pixel sites the open edge segment is sampled evenly at every i_oth pixel site. The two junction indices $j_b, j_t \in \{1, \ldots, J\}$ may of course be identical. This situation is depicted in figure C.11 (a). The case when $j_b \neq j_t$, is illustrated in figure C.11 (b). Normally $i_o = i_m$, but if $j_b = j_t$ *and* the open edge segment is very short, i_o is reduced as necessary so that again at least three samples are obtained. The sampling thus yields a sequence $\langle u_m \rangle_{m=3}^{M-1}$ for some $M \geq 5$. (If $j_b = j_t$, $M \geq 6$.) This sequence is appended to the vertex list, and a new open spline with control vertex sequence

$$\langle v_m \rangle_{m=0}^{M+2} = \langle w_{j_b}, w_{j_b}, w_{j_b}, u_3, \ldots, u_{M-1}, w_{j_t}, w_{j_t}, w_{j_t} \rangle$$

is appended to the spline list.

C.3.3 Alternative Initial Intermediate Vertex Selection

Picking the initial values for the intermediate vertices from the samples of the preliminary edge segments, as in the method outlined above, is very simple minded. In fact, the method works only because the resulting spline curves roughly approximate their defining control polygons, whose vertices by definition are identical to the sampled pixel sites. It would make much more sense to select the initial control vertices so that the resulting spline curves interpolate the samples of the preliminary edge segments. This would most likely yield superior starting points and hence shorter convergence times for the steepest descent procedure. Our main interest, however, is to study the performance of the steepest descent procedure, and for this purpose the simple minded method is quite adequate. Indeed, the worse the starting points are, the more robustness (of the procedure) can be demonstrated.

If spline curves that interpolate the preliminary edge segment samples are desired, they are not very hard to achieve. Let q be a new spline with control vertex sequence $\langle v_m \rangle_{m=0}^{M+2}$. If q is closed, the new initial independent control vertices, which have to be selected, are $v_0, \ldots, v_{M-1}$. The remaining control vertices are determined by the constraints

$$v_{M+m} = v_m \qquad m = 0, 1, 2 \tag{C.1}$$

In order to determine $v_0, \ldots, v_{M-1}$ so that q interpolates $u_0, \ldots, u_{M-1}$, one can for example demand that

$$u_0 = q(M-1) \tag{C.2a}$$

$$u_m = q(m-1) \qquad m = 1, \ldots, M-1 \tag{C.2b}$$

Substituting (2.2), (2.3) and (C.1) in (C.2) one then easily obtains the $2M \times 2M$ system

$$\begin{bmatrix} 4 & 1 & & & & 1 \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ 1 & & & & 1 & 4 \end{bmatrix} \otimes I_2 \begin{bmatrix} v_0 \\ \vdots \\ v_{M-1} \end{bmatrix} = \begin{bmatrix} u_0 \\ \vdots \\ u_{M-1} \end{bmatrix} 6 \tag{C.3}$$

where $\otimes$ denotes the Kronecker product, and I_2 is the identity matrix in $\mathbf{R}^2$. (The zero elements in the leading are left blank for better readability.) If q is open, the new initial

independent control vertices, which have to be selected, are $v_3, \ldots, v_{M-1}$. The remaining control vertices are determined by the constraints

$$v_0 = v_1 = v_2 = w_{j_b} \tag{C.4a}$$

$$v_M = v_{M+1} = v_{M+2} = w_{j_t} \tag{C.4b}$$

In this case it is reasonable to demand that

$$u_m = q(m-1) \qquad m = 3, \ldots, M-1 \tag{C.5}$$

A substitution similar to the one above then yields the $2(M-3) \times 2(M-3)$ system

$$\begin{bmatrix} 4 & 1 & & & & 0 \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ 0 & & & & 1 & 4 \end{bmatrix} \otimes I_2 \begin{bmatrix} v_3 \\ \vdots \\ v_{M-1} \end{bmatrix} = \begin{bmatrix} u_3 \\ \vdots \\ u_{M-1} \end{bmatrix} 6 - \begin{bmatrix} w_{j_b} \\ 0 \\ \vdots \\ 0 \\ w_{j_t} \end{bmatrix} \tag{C.6}$$

The coefficient matrices of the systems (C.3) and (C.6) are obviously positive, symmetric and diagonally dominant. By Geršgorin's circle theorem [62, p371] they are consequently also strictly positive definite. Both the systems (C.3) and (C.6) thus have unique solutions, which are easily computed with some numerical method such as Gauss-Seidel. Since the coefficient matrices are real, symmetric and strictly positive definite, the Gauss-Seidel method is indeed guaranteed to converge to the unique solution from any starting point [63, p355]. The obvious choice for such a starting point is of course given by

$$v_m^{(0)} = u_m \qquad m = 3o, \ldots, M-1$$

where o is the openness of the new spline.

C.4 Type Variables

The computation of the type variables o_n, β_n and τ_n, $n = 1, \ldots N$, actually begins already during the spline formation procedure just described. To each control vertex in the vertex list there is for this purpose a designated counter, which is originally set to zero. Every time a control vertex sequence $\langle v_m \rangle_{m=0}^{M+2}$ of a new open spline is recorded in the

spline list, the counter(s) associated with v_0 ($= v_1 = v_2$) and v_M ($= v_{M+1} = v_{M+2}$) are incremented. (If the endpoints of the new open spline are constrained to coincide, that is $v_0 = v_M$, the same counter associated with both v_0 and v_M is incremented twice.) When the spline formation procedure terminates, each counter shows how many times its associated control vertex in the vertex list serves as an end vertex for an *open* spline. The counters associated with junctions thus contain strictly positive integers while those associated with intermediate vertices remain at zero.

A special type variable computation procedure, which follows after the spline formation procedure, examines the counters associated with the end vertices of each spline in the spline list. A value of zero indicates (an end vertex of) a closed spline, a value of one indicates a free end vertex (of an open spline), and a value greater than one indicates a constrained end vertex (of an open spline). With this information at hand the type of each spline is trivially determined and recorded in the spline list.

Bibliography

[1] D. Marr, *Vision.* W.H.Freeman & Co., 1982.

[2] J. Malik and T. O. Binford, "A theory of line drawing interpretation," in *Image Understanding Workshop* (L. S. Baumann, ed.), pp. 188–194, Defense Advanced Research Projects Agency, Oct. 1984.

[3] J. Malik, "Interpreting line drawings of curved objects," *International Journal of Computer Vision*, vol. 1, no. 1, pp. 73–103, 1987.

[4] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679–698, 1986.

[5] R. M. Haralick, "Edge and region analysis for digital image data," *Computer Graphics and Image Processing*, vol. 12, pp. 60–73, Jan. 1980.

[6] M. H. Hueckel, "An operator which locates edges in digital pictures," Memo AIM-105, Stanford Computer Science Department, Oct. 1969.

[7] M. H. Hueckel, "An operator which locates edges in digital pictures," *JACM*, vol. 18, pp. 113–125, Jan. 1971.

[8] M. H. Hueckel, "A local visual operator which recognizes edges and lines," *JACM*, vol. 20, pp. 634–647, Oct. 1973.

[9] V. S. Nalwa and T. O. Binford, "On detecting edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 699–714, Nov. 1986.

[10] F. O'Gorman, "Edge detection using Walsh functions," in *Proc AISB*, p. 195, July 1976.

[11] K. S. Shanmugam, F. M. Dickey, and J. A. Green, "An optimal frequency domain filter for edge detection in digital images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 39–47, Jan. 1979.

[12] R. M. Haralick, "The digital edge," in *Proceedings of IEEE Conference on Pattern Recognition and and Image Processing*, pp. 285–291, Aug. 1981.

[13] R. M. Haralick, "Digital step edges from zero crossing of second directional derivatives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 58–68, Jan. 1984.

[14] D. Marr and E. Hildreth, "Theory of edge detection," A. I. Memo 518, Massachusetts Institute of Technology Artificial Intelligence Laboratory, Cambridge, MA, Apr. 1979.

[15] A. P. Blicher, *Edge Detection and Geometric Methods in Computer Vision.* PhD thesis, Stanford, CA, Oct. 1984.

[16] D. Mumford and J. Shah, "Boundary detection by minimizing functionals." Unpublished, 1986.

[17] V. Torre and T. Poggio, "On edge detection," A.I. Memo 768, Massachusetts Institute of Technology Artificial Intelligence Laboratory, Cambridge, MA, Aug. 1984.

[18] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 371, pp. 314–319, Sept. 1985.

[19] A. N. Tikhonov and V. I. Arsenin, *Solutions of Ill-Posed Problems.* Washington, D.C.: Winston, 1977.

[20] D. Terzopoulos, "Multilevel computational processes for visual surface reconstruction," *Computer Vision, Graphics, and Image Processing*, vol. 24, pp. 52–95, 1983.

[21] D. Terzopoulos, "Computing visible-surface representations," A.I. Memo 800, Massachusetts Institute of Technology Artificial Intelligence Laboratory, Cambridge, MA, Mar. 1985.

[22] A. Blake and A. Zisserman, "Invariant surface reconstruction using weak continuity constraints," in *Conference on Computer Vision and Pattern Recognition*, pp. 62–67, IEEE, 1986.

[23] A. Blake and A. Zisserman, "Some properties of weak continuity constraints and the GNC algorithm," in *Conference on Computer Vision and Pattern Recognition*, pp. 656–660, IEEE, 1986.

[24] A. Blake and A. Zisserman, *Visual Reconstruction.* The MIT Press, 1987.

[25] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, Nov. 1984.

[26] J. L. Marroquin, *Probabilistic Solution of Inverse Problems.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, Sept. 1985.

[27] P. Perona and J. Malik, "Scale space and edge detection using anisotropic diffusion," in *Workshop on Computer Vision—Miami*, pp. 16–22, IEEE Computer Society, June 1987.

[28] P. Perona and J. Malik, "A network for edge detection and scale space," in *International Symposium on Circuits and Systems—Helsinki*, pp. 2565–2568, IEEE, June 1988.

[29] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," Report UCB/CSD 88/483, Computer Science Division University of California, Berkeley, CA, Dec. 1988.

[30] K. N. Nordström, "Biased anisotropic diffusion—a unified regularization and diffusion approach to edge detection," Report UCB/CSD 89/514, Computer Science Division University of California, Berkeley, CA, June 1989.

[31] A. Blake, "Reconstructing a visible surface," in *Proc. AAAI conf.*, pp. 23–26, 1984.

[32] P. V. C. Hough, "Method and means for recognizing complex patterns." U.S. Patent 3,069,654, 1962.

[33] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, pp. 11–15, Jan. 1972.

[34] R. O. Duda and P. E. Hart, *Pattern Recognition and Scene Analysis.* New York: John Wiley & Sons, 1973.

[35] C. Kimme, D. H. Ballard, and J. Sklansky, "Finding circles by an array of accumulators," *Commun. ACM*, vol. 18, no. 2, pp. 120–122, 1975.

[36] D. H. Ballard, "Generalizing th hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[37] C. K. Chow and T. Kaneko, "Automatic boundary detection of the left ventricle from cineangiograms," *Computers and Biomedical Research*, vol. 5, pp. 388–410, Aug. 1972.

[38] R. Ohlander, K. Price, and D. R. Reddy, "Picture segmentation using a recursive region splitting method," *Computer Graphics and Image Processing*, vol. 8, Dec. 1979.

[39] D. H. Ballard and C. M. Brown, *Computer Vision*. Englewood Cliffs, NJ: Prentice-Hall, 1982.

[40] D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 413–424, July 1986.

[41] D. Mumford and J. Shah, "Boundary detection by minimizing functionals," in *Conference on Computer Vision and Pattern Recognition*, IEEE, 1985.

[42] D. Lee and T. Pavlidis, "One-dimensional regularization with discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, pp. 822–829, Nov. 1988.

[43] A. Martelli, "Edge detection using heuristic search methods," *Computer Graphics and Image Processing*, vol. 1, pp. 169–182, Aug. 1972.

[44] A. Martelli, "An application of heuristic search methods to edge and contour detection," *Commun. ACM*, vol. 19, pp. 73–83, Feb. 1976.

[45] U. Montanari, "On the optimal detection of curves in noisy pictures," *Commun. ACM*, vol. 14, pp. 335–345, May 1971.

[46] N. J. Nilsson, *Problem-Solving Methods in Artificial Intelligence*. New York: McGraw-Hill, 1971.

[47] N. J. Nilsson, *Principles of Artificial Intelligence*. Palo Alto, CA: Tioga, 1980.

[48] B. A. Barsky, "A study of the parametric uniform b-spline curve and surface representations," Report UCB/CSD 83/118, Computer Science Division University of California, Berkeley, CA94720, May 1983.

[49] B. A. Barsky, *Computer Graphics and Geometric Modeling Using Beta-Splines.* Springer, 1988.

[50] A. N. Tikhonov, "Regularization of incorrectly posed problems," *Sov. Math. Dokl.*, vol. 4, pp. 1624–1627, 1963.

[51] M. Atteia, "Fonctions "spline" et noyaux reproduisants d'Aronszajn-Bergman," *Revue Française d'Informatique et de Rccherche Operationelle*, vol. 4, pp. 31–43, Oct. 1970. Série Rouge.

[52] J. Duchon, "Splines minimizing rotation-invariant semi-norms in sobolev spaces," in *Conference Held at Oberwolfach April 25–May 1* (W. Schempp and K. Zeller, eds.), (Berlin), pp. 85–100, Springer, 1976. Series: Constructive Theory of Functions of Several Variables, edited by A. Dold and B. Eckmann.

[53] J. Meinguet, "An intrinsic approach to multivariate spline interpolation at arbitrary points," in *Polynomial and Spline Approximation—Theory and Applications* (B. N. Sahney, ed.), (Dortrecht, Holland), pp. 163–190, NATO Advanced Study Institute, D. Reidel, 1978.

[54] J. Meinguet, "Multivariate interpolation at arbitrary points made simple," *Journal of Applied Mathematics and Physics (ZAMP)*, vol. 30, pp. 292–304, 1979.

[55] R. A. Adams, *Sobolev Spaces*, vol. 65 of *Pure and Applied Mathematics.* New York: Academic Press, 1975.

[56] R. E. Showalter, *Hilbert Space Mcthods for Partial Differential Equations.* London: Pitman, 1977.

[57] W. Rudin, *Real and Complex Analysis.* McGraw-Hill, 1974.

[58] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques.* Paris: Masson, 1967.

[59] M. Spivak, *Calculus on Manifolds.* Menlo Park, CA: Benjamin/Cummings, 1965.

[60] J. R. Munkres, *Topology—A First Course.* Englewood Cliffs, NJ: Prentice-Hall, 1975.

[61] H. L. Royden, *Real Analysis.* New York: McMillan, second ed., 1963.

[62] P. Lancaster and M. Tismenetsky, *The Theory of Matrices.* Academic Press, 1985.

[63] G. H. Golub and C. F. Van Loan, *Matrix Computations.* The Johns Hopkins University Press, 1983.

[64] D. G. Luenberger, *Linear and Nonlinear Programming.* Reading, MA: Addison-Wesley, second ed., 1984.

[65] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, vol. II. John Wiley & Sons, 1962.

[66] J. F. Canny, "Finding edges and lines in images," Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, June 1983.