M. Nitzberg D. Mumford T. Shiota

Filtering, Segmentation and Depth

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Series Editors

Gerhard Goos
Universität Karlsruhe
Postfach 69 80
Vincenz-Priessnitz-Straße 1
W-7500 Karlsruhe, FRG

Juris Hartmanis
Cornell University
Department of Computer Science
4130 Upson Hall
Ithaca, NY 14853, USA

Authors

Mark Nitzberg
David Mumford
Department of Mathematics, Harvard University
Cambridge, MA 02183, USA

Takahiro Shiota
Department of Mathematics, Kyoto University
Kyoto, Japan

CR Subject Classification (1991): I.4.3, I.4.6, I.4.8, I.4.10, I.2.10

ISBN 3-540-56484-5 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-56484-5 Springer-Verlag New York Berlin Heidelberg

Printed in Germany

Typesetting: Camera ready by author/editor
45/3140-543210 - Printed on acid-free paper

Preface

Computer vision seeks a process that starts with a noisy, ambiguous signal from a TV camera and ends with a high-level description of discrete objects located in 3-dimensional space and identified in a human classification. In this book we address this process at several levels. We first treat the low-level image-processing issues of noise removal and smoothing while preserving important lines and singularities in an image. At a slightly higher level, we describe a robust contour tracing algorithm that produces a cartoon of the important lines in the image. Finally, we begin the high-level task of reconstructing the geometry of objects in the scene.

The problems in computer vision are so interrelated that to solve one we must solve them all. This book manages to make serious progress at several levels of visual reconstruction by working in a restricted world of simple pictures of simple objects.

We use a model that represents a scene as a group of overlapping shapes corresponding to the projections of objects. In constructing this representation from an image, the algorithm must imitate the process in human perception of inferring contours and surfaces that are occluded or are not present in the luminance function. Consequently, the work depends strongly on what we know about the psychology of perception, especially from the Gestalt school and its heirs.

We define the problem in this way: to find a decomposition of the domain D of an image that has the fewest disrupted edges–junctions of edges, crack tips, corners, and cusps–by creating suitable continuations for the disrupted edges. The result is a decomposition of D into overlapping regions $R_1 \cup ... \cup R_n$ ordered by occlusion, called the 2.1-D Sketch.

Chapters 2 through 5 describe algorithms that have been implemented in the C language for a SUN workstation running Unix[1] and X-Windows, using a library of computer vision functions called HVision. The programs from this book are available via anonymous FTP from internet host `math.harvard.edu`, in the directory `vision`.

If computer vision is to have a hope of serious long-term progress in the engineering sense, we must adopt the tradition from the numerical analysis community of sharing computer code. In this way, the next generation of results is built upon the combined best parts of the previous generation.

[1] Unix is a trademark of AT&T Bell Laboratories

The authors gratefully acknowledge the support of the National Science and Technology Research Center for Computation and Visualization of Geometric Structures.

October 1992

Mark Nitzberg
David Mumford
Takahiro Shiota

(Left to right: David Mumford, Takahiro Shiota, Mark Nitzberg)

Contents

1 Overview 1
- 1.1 Segmenting with depth . . . 2
- 1.2 Edge terminations and continuations . . . 4
- 1.3 A variational model . . . 7
- 1.4 A computer algorithm . . . 11

2 Filtering for Occlusion Detection 13
- 2.1 Nonlinear diffusion . . . 18
- 2.2 Experimental results . . . 23
- 2.3 Using Q for corner detection . . . 27

3 Finding Contours and Junctions 33
- 3.1 Finding edges . . . 34
- 3.2 Finding corners and T-junctions . . . 40
- 3.3 Curve smoothing . . . 42

4 Continuations 51
- 4.1 The psychology of continuations . . . 51
- 4.2 Elastica . . . 55
- 4.3 Matching ends and computing continuations . . . 58
- 4.4 Results . . . 61

5 Finding the 2.1D Sketch 73
- 5.1 Recursive search . . . 73
- 5.2 Results on pictures . . . 75

6 Conclusion 83

A Derivations for the Nonlinear Filter 87

B Program Code 91
- B.1 Nonlinear filter . . . 91
- B.2 Edge tracing . . . 96

Bibliography 141

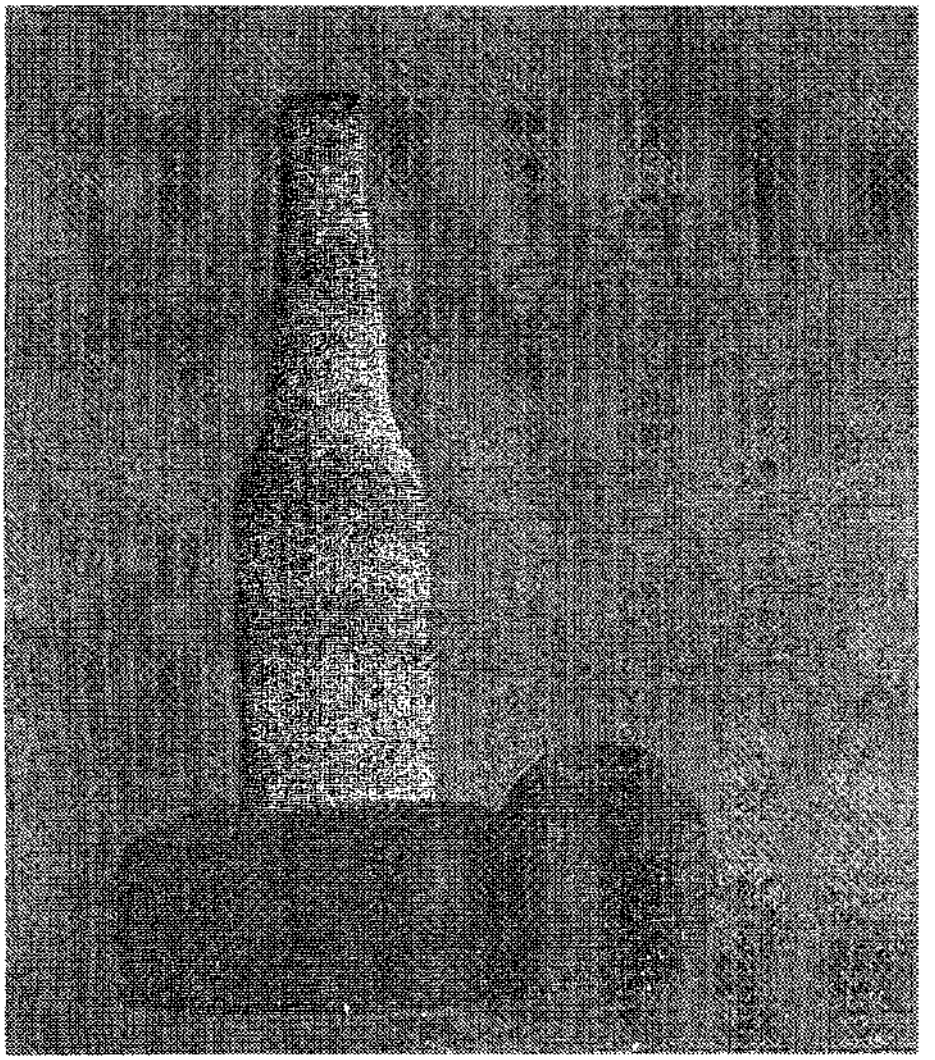


Figure 0.1: Still life with subliminal ad for local grocery store

Chapter 1

Overview

In the picture on the facing page, the potato partly covers the nectarine and also hides part of the bottle. This partial overlap, or *occlusion*, of farther objects by those nearer, is one of the most fundamental obstacles that the visual system must overcome to achieve its goal of recognizing and locating objects in our three-dimensional world.

Most of the structurally important lines in what we see are the boundaries that separate objects from the view of the things behind them. For the past 25 years, researchers in computer vision have been trying to find lines and edges in images in order to recognize objects automatically. What has been missing in this endeavor is that most of what we see is partially covered by something nearer. In this book we have tried to lay out a practical way of incorporating occlusion into the task of finding object outlines.

Can we really believe that occlusion is detected at a low level? Even without stereopsis and motion cues, we can experience a striking impression of depth from a single, stationary retinal image such as from a photograph or painting. The main cue to occlusion in this setting comes from the points where objects overlap in a scene. These are *edge terminations*, which occur most often where one object outline stops, abruptly abuts against the outline of a nearer object, and forms a junction in the shape of the capital letter 'T'. Look again at the still life, and note how nearly all the occlusion relations can be readily computed based on the T-junctions.

One theory has it that the impression that one fruit is in front of another comes from first recognizing the objects, or at least familiar shapes, and then noticing that they are incomplete examples of those shapes. This implies that we have already performed a difficult task–to recognize a shape from an incomplete example of it–in order then to infer that it is indeed incomplete, and therefore partly occluded. A more plausible model for early visual organization, and one which psychologists support increasingly, is a process driven by locating disrupted boundaries and building continuations for them behind occluding surfaces, with recognition coming later. By reconstructing just this first "bit" of the third dimension, the visual system has simpler shape data from which to find objects.

Here are two illustrations of how higher-level cognition does not have a decisive effect on visual organization. Firstly, we readily perceive occlusion among unfamiliar shapes, as in Figure 1.1. This supports the idea that familiar shape is not necessarily what tells us which vegetable is in front of which in the still life. Secondly, the contemporary psychologist Gaetano Kanizsa, notable as much for his art as for his psychology, has

Figure 1.1: One unfamiliar shape occludes another.

sketched pictures for which our perception directly contradicts what we know about the world. In figure 1.2, for example, the man and woman are entangled in the fence. Knowing that they are behind the fence does not change this unusual perception. Thus perceptual organization follows its own set of rules that depend decisively on occlusion but not on higher-level object recognition. Simply put, we navigate in the world successfully by seeing what's in front of what independently of knowing what's what.

In this book we make explicit some of the these hidden rules of perceptual organization, and then cast them in a model that lets us compute the relative depth of objects from occlusion cues–in other words, to find depth from overlap.

1.1 Segmenting with depth

We propose a model that lets us reconstruct object shapes from a picture of several interposed objects, including parts that are occluded by nearer objects. The model also determines the nearness relations of the objects.

This is a novel approach to one of the principal goals of computer vision, that of *segmenting* an image. Roughly speaking, to segment means to find regions of interest in a picture, so that these regions can be parceled out for further analysis. The ultimate aim of our algorithm will be similar; however, regions of interest will now be allowed to overlap and occlude one another, and in addition, the hidden parts of incomplete regions will be restored by hypothetical completions.

Image segmentation has come to mean the process of cutting up a picture into the simplest shaped pieces possible while keeping the color or luminance of each piece as uniform as possible. An image is given as a function $g(x, y), (x, y) \epsilon D$ representing the light intensity or color vector produced by a 3D world and striking a lens from direction (x, y). The aim is to segment the domain D, i.e. partition D into regions $R_1, \ldots, R_k$ such that R_i is the part of the image in which the nearest object is some object O_i and on the boundary between any two regions R_i and R_j, object O_i occludes object O_j or vice-versa.

This assumes that we have some decisive way to find the regions that correspond to distinct objects. In general, the variety of lighting situations, surface characteristics and textures in the world make it necessary to integrate visual input of various types: depth cues from stereo and motion, texture boundaries, shading and shadows. To build a practical system, we restrict ourselves to single, stationary images of objects with uniform nontextured surfaces where shadows and shading do not hide important object features.

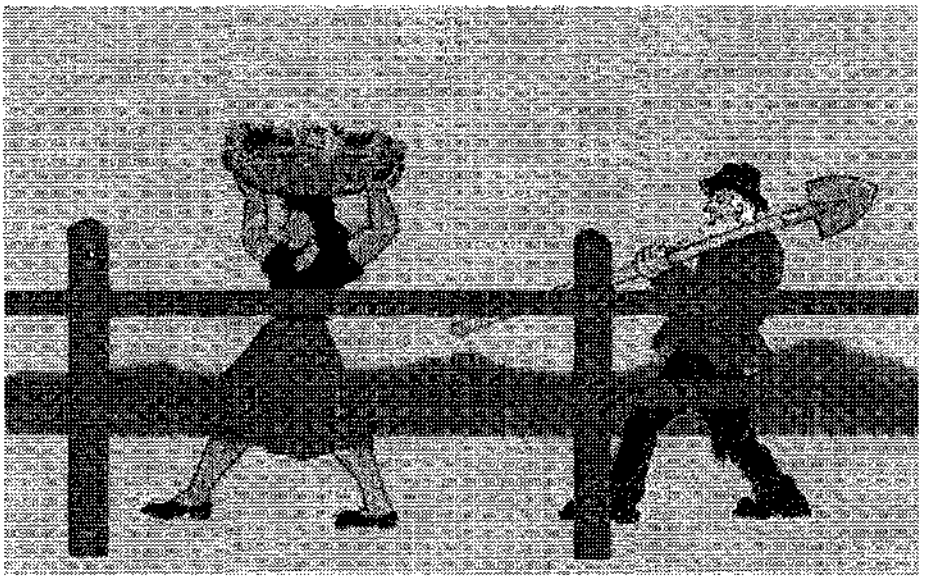


Figure 1.2: The man and woman are entangled in the fence.

Our model also addresses another goal of vision, that of computing or estimating what David Marr called the $2\frac{1}{2}$D sketch associated to an image [24]; i.e., the depth image $z(x, y)$ recording the distance from the lens to the nearest object in the direction (x, y) and its normalized partial derivatives:

$$\begin{aligned} p(x,y) &= z_x/\sqrt{1+z_x^2+z_y^2} \\ q(x,y) &= z_y/\sqrt{1+z_x^2+z_y^2} \end{aligned}$$

Marr proposed multiple sources of information contained in the intensity image $g(x, y)$ from which one could hope to estimate the $2\frac{1}{2}$D sketch (z, p, q). This too has proved hard to implement except under strong constraints, for example, where very accurate stereo or motion data is available, or where the lighting and surface reflectances are heavily constrained.

Our model achieves a synthesis of these two goals, segmentation and the $2\frac{1}{2}$D sketch, while avoiding the numerical burden of the $2\frac{1}{2}$D sketch and at the same time simplifying 2D segmentation by incorporating occlusion explicitly. Hence in the language of computer vision, we might call our model the 2.1D sketch.

Consider figure 1.3(a), an image of several blades of grass against a light background. Figure 1.3(b) shows the 12 disjoint regions that result from cutting (a) along visible object boundaries. However, the 12 regions do not correspond to 12 distinct objects in the world: there are only 4 objects reflecting light—the 3 blades of grass and the background "object". Although each of the original objects lies at varying depth, there is a simple

ordering of the objects that indicates which objects occlude which. We can describe the scene as a stage set with 4 "wings", transparent except where they contain an object. The background is last in the set and is everywhere opaque. This is shown in figure 1.3(c).

This is what we mean by a 2.1D sketch: it is a set of regions R_i in the domain D of the image which fill up D but which may overlap, plus a partial ordering $<$ on the regions indicating which are in front of which others. Often there will be a background object R_0 behind all others for which $R_0 = D$. Our contention is that this type of segmentation is more natural than the kind with disjoint, unordered R_i and that it captures the most accessible part of the $2\frac{1}{2}$D sketch.

1.2 Edge terminations and continuations

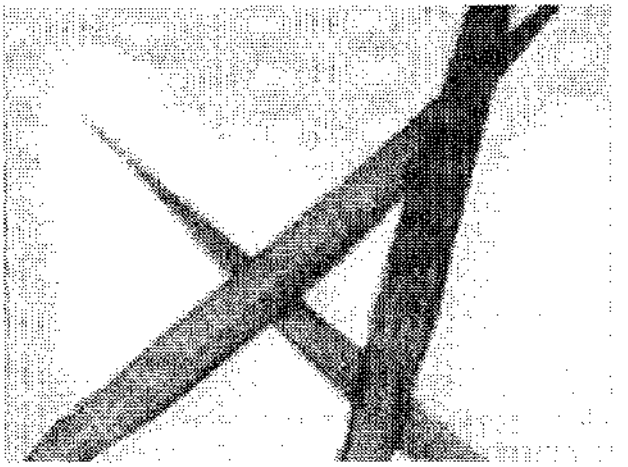
The chief reason we expect the 2.1D sketch to be readily computable is the presence of edge terminations, and in particular T-junctions. T-junctions are points where the edges in the image form a "T", with one edge Γ_1 ending abruptly in the middle of a second edge Γ_2. Such points often arise because Γ_2 is an occlusion edge and Γ_1 is any kind of edge—occlusion, shadow, surface-marking—of a more distant object whose continuation disappears behind Γ_2.

The importance of T-junctions in the human visual system has been known for a long time, but their role and power have been greatly clarified by recent work. In particular, it has become increasingly clear that T-junctions are computed early in the visual process and are not merely part of an object recognition paradigm as in the early blocks world algorithms of Guzman, Roberts, Waltz, etc (cf. [39]). The gestalt school of psychology and, particularly, the contemporary psychologist Gaetano Kanizsa have made a thorough and deep analysis of T-junctions [17]. Consider figure 1.4 from Kanizsa. 1.4(a) and 1.4(b) differ only in the addition of diagonal lines which change the corners in 1.4(a) to T-junctions in 1.4(b); 1.4(b) is unmistakably 3-dimensional. More importantly, we infer that something is being occluded and fill in the hidden parts. 1.4(a) and 1.4(c) differ only in the subtraction of short connecting lines which change corners in 1.4(a) to terminators in 1.4(c).

A terminating line is a weak form of T-junction in that it signals occlusion approximately perpendicular to the line at its end. Likewise corners can be thought of as degenerate forms of T-junctions, especially when pairs of their edges are aligned, as in Kanizsa's triangle illusion (see figure 1.5). In general, when several edge terminations are aligned, we tend to perceive a contour "connecting" the terminations along which one surface occludes another. The alignment of terminations seems to cause the hypothetical T-junctions to mutually confirm one another.

A striking confirmation of the reality of these so-called illusions and the illusory contours that we see was found by R. von der Heydt and his colleagues [15] The responses of single cells in visual areas are codified by describing their visual field: the area within which moving or stationary bars and edges produce activity. They found, however, that many cells in visual area V2[1] responded when no actual stimulus was

[1]Known as Brodmann area 18 in man. this area is adjacent to the primary visual area V1 (= area 17 = striate cortex) and is a recipient of a high proportion of its axonal output.



(a)

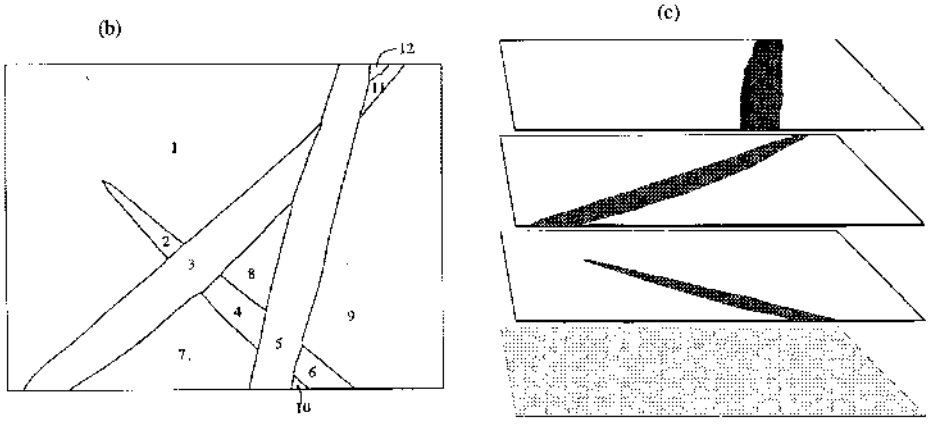


Figure 1.3: (a) blades of grass image; (b) a disjoint segmentation; (c) a segmentation with overlaps.

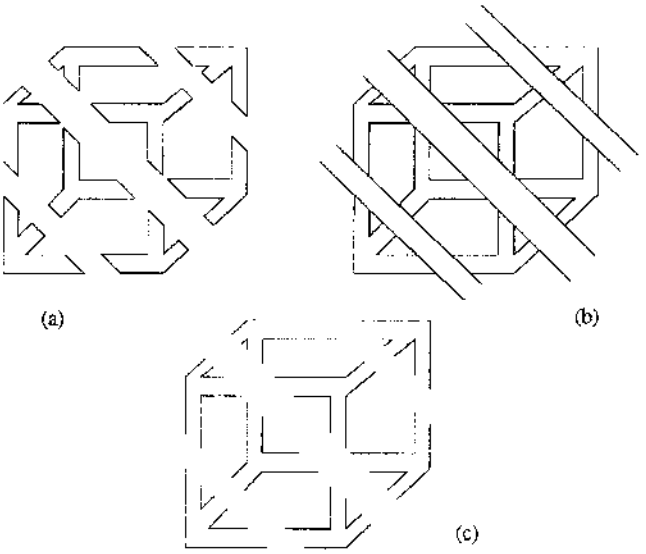


Figure 1.4: A demonstration from Kanizsa of the importance of T-junctions and terminators.

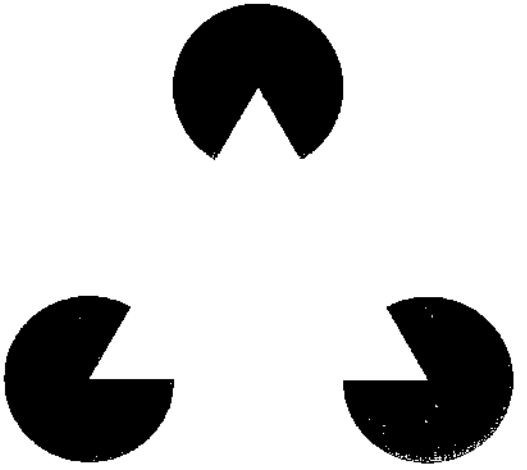


Figure 1.5: Kanizsa's triangle illusion: not three Pac-Man shapes, but three circles occluded by a nearer white triangle.

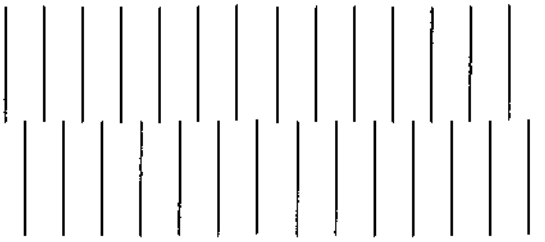


Figure 1.6: Aligned terminators elicit neuronal responses along the subjective contour.

present in their visual field, but rather when edges outside this field produced illusory contours that crossed the field. Thus, stimuli such as in figure 1.6 evoke responses from "horizontal line-detector" cells whose job is normally to find horizontal lines of high contrast.

In all these cases, the mind seems to create a 3D scene in which occluded parts of visible objects are reconstructed. In the case of the Kanizsa triangle, the mind goes further and creates missing outlines of the nearer occluding triangle, and compensates for their absence in the raw data by a perceptual impression that the white triangle is brighter than the more distant white background.

Contours such as as the sides of the Kanizsa triangle are known as subjective contours, because they are not present in the gray level image, yet they are particularly vivid under certain circumstances. Nakayama and Shimojo [31] have studied the mechanics of these subjective surfaces, and have produced numerous demonstrations that the pictorial cues to occlusion are used in an early processing stage of human vision that drives the grouping processes. Theirs and related results in psychology that bear directly on the 2.1D sketch model are discussed in Chapter 4.

1.3 A variational model

To set down the requirements of the 2.1D sketch more precisely, we define an energy functional that takes its minimum at an optimal 2.1D sketch of an input image. We begin by recalling the variational model used for image segmentation without overlaps.

The piecewise smooth model of the segmentation problem in computer vision asks how to clip a picture into as few and simple pieces as possible while keeping the color of each piece as smooth and/or slowly varying as possible. One approach to the problem, taken by Mumford and Shah [28], is to define a functional that takes its minimum at an optimal piecewise smooth approximation to a given image. The image is a function g defined on a domain D in the plane. It is approximated by a function f, which is smooth except at a finite set Γ of piecewise C^1 contours which meet ∂D and meet each other only at their endpoints. The functional defined below gives a measure of the match between an image g and a segmentation f, Γ:

$$E_{\text{M-S}}(f,\Gamma) = \mu^2 \int_D (f-g)^2 d\mathbf{x} + \int_{D\setminus\Gamma} \|\nabla f\|^2 d\mathbf{x} + \nu \int_\Gamma ds.$$

The first term asks that f approximate g, the second asks that f vary slowly except at boundaries, and the third asks that the set of contours be as short, and hence as simple and straight as possible. The contours of Γ cut D into a finite set of disjoint regions $R_1, \ldots, R_n$, the connected components of $D \backslash \Gamma$.

In this book, however, we seek a model that incorporates partially the way that g derives from a 2D projection of a 3D scene. Rather than base our 2.1D model on a set of curves Γ that cuts D into disjoint regions, we ask for a set of regions R_i whose union equals D, and with a partial ordering that represents relative depth. The overlapping of regions gives in a sense the most primitive depth information. The domain D is considered as a window that reveals the value of g only on a portion of the plane. As a result, contour integrals will exclude portions of a contour that coincide with the boundary of D.

We now seek a functional $E_{2.1}$ much like $E_{\text{M-S}}$ that achieves a minimum at the optimal overlapping segmentation of g. Let $\{R_1, \ldots, R_n\}$ be a set of regions such that $\bigcup_i R_i = D$, with a partial ordering $<$ that represents occlusion, e.g., $R_i < R_j$ means R_i occludes R_j.

$$R_i' = R_i \backslash \bigcup_{R_i < R_j} R_j$$

is the "visible" portion of R_i. Throughout the book, R_i denotes a closed subset of D with piecewise smooth boundary and connected interior. The expression $(\{R_i\}, <)$ denotes an ordered set of overlapping regions, which we will call a *segmentation*.

We then define the energy $E_{2.1}(\{R_i\}, <)$ as

$$\sum_{i=1}^{n} \left(\mu^2 \int_{R_i'} (g - m_i)^2 dx + \epsilon \int_{R_i} dx + \int_{\partial R_i \backslash \partial D} \phi(\kappa) ds \right).$$

In this formula, m_i is the mean of g on R_i', and κ is the curvature of ∂R_i, i.e. $\|\ddot{\gamma}\|$ where γ parameterizes ∂R_i by arc length. The function $\phi : \mathbb{R} \to \mathbb{R}$ is defined by

$$\phi(\kappa) = \begin{cases} \nu + \alpha\kappa^2 & \text{for } |\kappa| < \beta/\alpha \\ \nu + \beta|\kappa| & \text{for } |\kappa| \geq \beta/\alpha \end{cases}.$$

The scalar constants μ, ν, ϵ, α and β in the definition of ϕ, determine the characteristics of a segmentation which minimizes $E_{2.1}$. Their dimensions are:

$$\begin{aligned} \mu &\sim \text{intensity}^{-1}.\text{dist.}^{-1} \\ \nu &\sim \text{dist.}^{-1} \\ \alpha &\sim \text{dist.} \\ \beta &\sim \text{dimensionless} \\ \epsilon &\sim \text{dist.}^{-2} \end{aligned}$$

Before analyzing the functional, we should describe its relation to the 2.1D sketch *model*, and to the computer algorithm that finds the 2.1D sketch of an image. The model refers to the representation of an image by a set of possibly overlapping shapes, together with depth relations between them. Writing down a functional $E_{2.1}$ is a way of describing concisely what makes a good set of shapes for a given image. For example, the first term of $E_{2.1}$ asks that the various visible parts of a single region ought to be of nearly the same