An Extended Mumford-Shah Model and an Improved Region Merging Algorithm for Image Segmentation

THE UNIVERSITY OF ADELAIDE

Department of Applied Mathematics

Trevor Tao

A thesis submitted for the degree of

Doctor of Philosophy.

Primary Supervisor: Dr John van der Hoek (University of Adelaide)

Secondary Supervisor: Dr David James Crisp (Defence Science and Technology Organization)

October 25, 2005

Contents

List of Tables

List of Figures

Abstract

In this thesis we extend the Mumford-Shah model and propose a new region merging algorithm for image segmentation. The segmentation problem is to determine an optimal partition of an image into constituent regions such that individual regions are homogenous within and adjacent regions have contrasting properties. By optimal, we mean one that minimizes a particular energy functional. In region merging, the image is initially divided into a very fine grid, with each pixel being a separate region. Regions are then recursively merged until it is no longer possible to decrease the energy functional.

In 1994, Koepfler, Lopez and Morel developed a region merging algorithm for segmentating an image. They consider the piecewise constant Mumford-Shah model, where the energy functional consists of two terms, accuracy versus complexity, with the trade-off controlled by a scale parameter. They show that one can efficiently generate a hierarchy of segmentations from coarse to fine. This algorithm is complemented by a sound theoretical analysis of the piecewise constant model, due to Morel and Solimini.

The primary motivation for extending the Mumford-Shah model stems from the fact that this model is only suitable for "cartoon" images, where each region is uncomtaminated by any form of noise. Other shortcomings also need to be addressed. In the algorithm of Koepfler et al., it is difficult to determine the order in which the regions are merged and a "schedule" is required in order to determine the number

and fine-ness of segmentations in the hierarchy. Both of these difficulties mitigate the theoretical analysis of Koepfler's algorithm. There is no definite method for selecting the "optimal" value of the scale parameter itself. Furthermore, the mathematical analysis is not well understood for more complex models. None of these issues are convincingly answered in the literature.

This thesis aims to provide some answers to the above shortcomings by introducing new techniques for region merging algorithms and a better understanding of the theoretical analysis of both the mathematics and the algorithm's performance. A review of general segmentation techniques is provided early in this thesis. Also discussed is the development of an "extended" model to account for white noise contamination of images, and an improvement of Koepfler's original algorithm which eliminates the need for a schedule. The work of Morel and Solimini is generalized to the extended model. Also considered is an application to textured images and the issue of selecting the value of the scale parameter.

Signed Statement

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

SIGNED: ..... DATE: .....

Acknowledgements

I would like to acknowledge my supervisors John van der Hoek and David Crisp for their valuable guidance and much good advice.

I also wish to thank my family and friends for support and encouragement during my PhD program, and my mum for the endless proofreading.

The support of the Adelaide University Scholarship and the CSSIP Supplementary Scholarships is also gratefully appreciated, as are the friendly staff and excellent facilities at Adelaide University.

Chapter 1

Introduction

Image segmentation is a standard problem in the field of image processing. The goal is to partition an image into a set of regions that satisfy desirable properties, such as homogeneity within regions and contrast between adjacent regions, according to some predetermined criteria, enabling further processing such as classification to be performed using higher level structures of the image rather than image pixels. The image is usually represented as a function g on a rectangular domain Ω. g can be either a scalar or vector function. The latter is useful for characterizing multi-channel data such as colour, texture features, spline coefficients and so on.

In image segmentation algorithms it is common to specify some a-priori assumptions on image datum. For example, we might assume the image to consist of a "round-shaped" object plus background, or regions which are reasonably close to being constant and so on. It is a well-accepted notion that a good choice of segmentation algorithm is highly dependent on a-priori knowledge about an image. For example if we assume an image consists of a number of bright objects on a dark background and we have prior knowledge of distributions of the pixel intensities that comprise the object/background then the well-known technique of histogramming and thresholding is applicable. But such techniques hardly come into consideration for segmenting complex images such as those with a significant amount of noise.

1.1 Low-Level Methods

In the most general case, a segmentation can be viewed in two ways: either the partition of an image into regions can be determined, or its boundary set can be computed. These are "dual" in the sense that given one, we can determine the other. Indeed, there exist both region-based and edge-based algorithms for segmentation.

The basic concept of region-based methods is as follows: starting from a very fine segmentation, one gradually merges small regions or individual pixels into larger regions until a desired segmentation output is obtained. There are many ways to achieve this. For instance, an early paper by Muerle and Allen [60] proposes to begin merging with a seed, or single-pixel region and accumulate individual pixels according to some "acceptance criteria". When no more pixels are accepted, the region is complete and a new cell is selected. This is known as region aggregation (Rosenfeld and Kak, [73], Zucker, [97]). An alternative concept is region growing (Leonardis et al., [46]), where a set of small regions is first determined by some "pre-processing" segmentation stage such as determining connected components of constant gray value pixels. Neighbouring regions are merged together according to some acceptance criteria. Thus region growing differs from aggregation in that regions are treated as the basic "unit" instead of individual pixels. An example of an acceptance criteria is that pixels be homogenous, for example if we were segmenting a gray-level image with intensities between integers $0, \ldots, 255$ then we can define a region as being "legal" if the difference of gray values of any two pixels is less than 20. Note that this does not prevent overlapping regions in the final segmentation. A segmentation without overlapping regions is known as a partition (Rosenfeld and Kak, [73]).

In edge-linking algorithms (Bajcsy and Tavakoli, [7], Rosenfeld and Kak, [74], Weiss and Boldt [90]), a set of edges is determined by local properties such as image gradient. Contours are then formed by searching for pairs of egdes that can be conveniently linked into a "stronger" edge until some stopping criterion is satisfied. Edges that cannot link with other edges are discarded. Adjacent edges with large gradients and similar orientations make good candidates for merging, whereas those with smaller gradients or conflicting orientations would be discarded as noise. An example of an edge-linking algorithm is due to Perkins [65]. The algorithm consists of the following steps:

- Compute edge regions by a threshold on gradient.
- Thin the gradient regions.
- Expand edges to close the small gaps between them.
- Find different regions via connectivity algorithm.
- Eliminate small regions.
- Shrink again the edge regions.
- Add new edge pixels to close newly created gaps.
- Determine active edge regions.
- Eliminate small regions again.
- Calculate properties of uniform intensity regions.

It is clear that there are many parameters and thresholds to set. For example, a threshold must be defined in order to determine whether a region is "sufficiently" small, and whether a gap between two edges is sufficiently small. Moreover, one might argue that the above steps might be performed in a different order, or the

number of steps be changed. An advantage of low-level methods is that initialization problems are rarely significant, compared to high-level methods, which are discussed next. Also, for edge-based algorithms the usual edge detectors are generally applicable to almost any imagery. However there are also obvious disadvantages. For example, low-level methods only consider local information, and it is a cumbersome process to synthesize the data into something more meaningful. A diametrically opposite approach consists of starting with large-scale objects. The basic idea is to eliminate the use of low-level structures altogether.

1.2 High-Level Methods

Some region-based algorithms attempt to allow greater flexibility by admitting both splitting and merging operations (Rosenfeld and Kak, [73]). Typically, an image has dimensions $2^m \times 2^n$ for integers m, n (if this is not true, the algorithm can be easily modified by "padding out" the image with zeros). The algorithm starts by considering the entire image as a single region. If the region is not "homogenous enough" it is split into four congruent squares and each square is tested separately in the same manner. Once splitting is finished, adjacent regions may then be merged if desired. The most obvious disadvantage of this approach is the bias towards "blocky" regions with side-lengths likely to be powers of two. It is possible to start at an intermediate level with blocks of size $2^l \times 2^l, l < \min(m, n)$ but this causes initialization problems and is unlikely to affect the final result anyway.

An interesting approach to edge detection is to formulate the problem of finding the best edge using dynamic programming principles. An early paper by Martelli [49] analyzed a relatively simple problem of segmenting an image on a rectangular domain assumed to consist of two regions, "left" and "right" separated by an edge. An edge is defined as a connected path of edge "elements" starting from anywhere at the top to anywhere at the bottom. The edge elements can be considered as forming

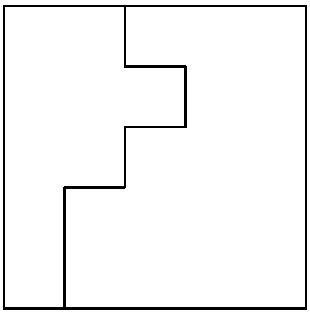


Figure 1.2.1: An edge separating two regions.

a completely connected graph (see Figure 1.2.1), two elements being neighbours if and only if they share a pixel, and Martelli argued that the problem of finding the best edge could be interpreted as solving a well-known dynamic programming problem: finding the cheapest path from a "start" node to "goal" node with specified costs for each link connecting two nodes. The best edge is defined as the one which minimizes the cost function $Cost = |\sum_e c(e)|$, where

$$c(e) = M - g(x) + g(y), \tag{1.2.1}$$

$$M = \sup_z g(z). \tag{1.2.2}$$

Here $c(e)$ is the cost for a single edge element e and x, y are the pixels on opposite sides of the edge element and $g(x) \geq g(y)$. Note that the cost is defined as a sum of positive contributions from each edge element, so the edge "wants" to be as short as possible. Also $M - g(x) + g(y)$ is small only when the difference in gray values $g(x) - g(y)$ is large.

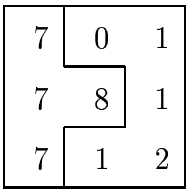


Figure 1.2.2: The optimal edge computed for a simple image for Martelli's problem.

An example of finding the optimal edge is shown in Figure 1.2.2 and the cost of this edge is 5. Obviously Martelli's ideas [49] are only of theoretical interest since the assumption that an image consists only of two regions is unrealistic for real applications. However, it does highlight an important idea: the quality of a segmentation can be quantified by a function, or more precisely, a functional which associates that segmentation with a real number. This brings us to the concept of the Variational Formulation, discussed in the next subsection.

1.3 The Variational Formulation

The idea of the Variational Formulation is simple: starting with an initial segmentation, we define a set of operations that vary the segmentation slightly (hence the word "variational"). We then alter the segmentation using these operations until it is no longer possible to improve it. The quality of a segmentation is measured by an energy functional E and by convention, lower numbers indicate better quality. Thus K' is an "improvement" over K if $E(K') < E(K)$. Typically E will consist of a number of terms, each seeking to promote a desirable property of a segmentation, such as a smooth boundary set or homogeneity within regions. Note that it is rarely possible to recover a globally optimal segmentation since the search for the best segmentation stops if a local optimum is reached. Generally, the search space of all possible segmentations is large enough to render a brute force search infeasable. Thus, it is usual to accept a local minimum instead of global if it "looks reasonable".

1.3.1 The Mumford-Shah Model and Region Merging

Pavlidis [64] made the important observation that the segmentation obtained by a set of regions or edges can be complemented using a simple approximation u to the original image data g. Thus a metric (such as L^2-norm) can be defined, allowing us to measure the difference between a given image and its approximation. A very well-known example of this is the celebrated Mumford-Shah functional [61]:

$$E(u, K) = \mu^2 \int (u - g)^2 dx + \int |\nabla u|^2 dx + \nu \cdot \ell(K), \tag{1.3.1}$$

where μ^2, ν are scale parameters, u is a piecewise smooth approximation of g and K is the boundary set, assumed to be piecewise smooth with finite total length of curves equal to $\ell(K)$. The first term is just the L^2-norm difference between the image data and its approximation. The second term measures the smoothness within each region and the final term measures the complexity of the segmentation. A useful and popular simplification is to assume u is piecewise constant on each region of the segmentation. Thus the term $\int |\nabla u|^2 dx$ is dropped. Setting $\lambda = \nu/\mu^2$ and scaling equation (1.3.1) yields the piecewise-constant Mumford-Shah model:

$$E(u, K) = \int (u - g)^2 dx + \lambda \cdot \ell(K). \tag{1.3.2}$$

In this thesis, we will only be concerned with equation (1.3.2) and not (1.3.1). For the sake of brevity we will refer to equation (1.3.2) as the "Mumford-Shah model" instead of the "piecewise-constant Mumford-Shah model" unless stated otherwise. As is well known, finding the global minimizer of (1.3.2) is computationally infeasible and it is necessary to settle for a local minimizer. A popular method for computing such a local minimizer is region merging (Koepfler et al., [42]), although other approaches are known (Nordstrom, [62], Richardson, [69]). It is assumed that the image domain Ω is divided into pixels, where g is constant within each pixel, and K is assumed to consist of vertical and horizontal "edge elements" between pixels.

4	2	1	9	6	7	8	5	3
6	7	5	3	1	8	4	9	2
3	8	9	2	4	5	6	1	7
1	9	8	7	3	4	5	2	6
7	4	2	8	5	6	1	3	9
5	6	3	1	2	9	7	4	8
2	1	6	5	7	3	9	8	4
8	3	7	4	9	1	2	6	5
9	5	4	6	8	2	3	7	1

Figure 1.3.1: An example segmentation of an image in the discrete domain.

Both terms of (1.3.2) are computed accordingly. For example, in Figure 1.3.1, if $u(\mathbf{x}) = 5$ for all $\mathbf{x}$ and K is the boundary set shown then $E(u, K) = 540 + 36\lambda$. In these algorithms, all pixels must belong to one region during the merging process and no two regions can overlap. Two regions can be merged if and only if they are adjacent, that is, they share a common boundary. Koepfler's algorithm (Koepfler et al., [42])[1] is one of the earliest examples of region merging and its scheme can be summarized as follows:

- Initialization: set $u_0 = g$ and K_0 is the union of all boundaries of all pixels. Set $\lambda_0 = 0$.
- Recursive merging: merge all pairs of regions whose merging decreases the energy.
- Change of scale: increase λ and return to the previous step.

[1] Although this paper was written by three authors, we will always write "Koepfler's algorithm" for reasons of brevity.

This is a simplified statement of Koepfler's algorithm from Morel and Solimini [58]. We study Koepfler's algorithm in more detail in Chapter 2.

Region merging methods are attractive due to their ease of implementation. One drawback is that the final segmentations tend to have jagged boundaries even for smooth images. The main reason is as follows: during the early stages of region merging, it is difficult to anticipate if merging two small regions will break an "edge" that is only obvious at a large scale, but once pixels have been merged into a single region at an early stage, they belong forever to the same region, making it impossible to undo these mistakes. An example of this phenomenon can be found in the Brodatz experiment in (Morel and Solimini, [58]). A possible solution lies in the Snake and Active Contour models which is discussed next. We present only a brief summary of the fundamental ideas and refer to (Aubert and Kornprobst, [6], McInerney, [54]) for a more detailed discussion.

1.3.2 The Snake Model, Active Contour Model and Level Set Methods

A natural formulation for the edge detection problem is to consider a moving curve, starting at some initial position and evolving over an artificial parameter of time, until it fits our "expected notion" of an edge. This is the idea behind the so-called "deformable models". Deformable models are also useful for computer vision/graphics applications: indeed it is the latter in which these methodologies were originally developed. An edge is usually represented by some structure in an image. For instance if an image is assumed *a-priori* to be noise-free then a high gradient probably indicates an edge.

One of the first efforts in analyzing this formulation was made by Kass, Witkin and Terzopoulos (Kass et al., [40]). Due to the complex motion of the curves, the authors called them *snakes* and the model itself is referred to as the Snake Model. Also well-established in the literature is the Geodesic Active Contour Model (Caselles et al., [14], Caselles et al., [15], Kichenassamy et al., [41]). The basic idea is that, starting from an initial position, a curve tries to move in such a way that it approaches image boundaries while also trying to stay smooth. The latter constraint ensures the snake does not respond to noisy "false" edges. In the Snake Model, a typical energy J is defined by (Aubert and Kornprobst, [6], Kass et al., [40])

$$J(c) = \int |c'(q)|^2 dq + \beta \int |c''(q)| dq + \lambda \int g^2(|\nabla I(c(q))|) dq, \tag{1.3.3}$$

where c is the curve, I is the image data, $\beta, \lambda \in \mathbf{R}$ are parameters and $g : \mathbf{R}^+ \to \mathbf{R}^+$ is the "edge detector function": g only depends on the magnitude of the gradient and is monotonically decreasing. A typical choice is

$$g(|\nabla I|) = \frac{1}{1 + |\nabla G_\sigma * I|^2},$$

where G_σ is a Gaussian kernel. Thus g is strictly positive in homogenous regions and near zero on the edges[2]. The first two terms measure the smoothness of the curve and the last measures how well the curve corresponds to sharp discontinuities in the image. The first two terms are called an internal energy since it is independent of the given data and the last is called an external energy. It is usual to omit the middle term ($\beta = 0$) since practice indicates that the curvature still decreases regardless of the condition $\beta = 0$. Note that (1.3.3) is not intrinsic, that is, by changing the parametrization of the curve one obtains a different solution. An alternative is to use the following functional, which does not depend on curve parametrization (Caselles et al., [14], Caselles et al., [15]):

[2]We abuse notation and use g both as a function of a real number (image gradient magnitude) and a point in the spatial domain ($x \in \Omega$).

$$J(c) = 2\sqrt{\lambda}\int g(|\nabla I(c(q))|)|c'(q)|dq \quad (1.3.4)$$

$$= 2\sqrt{\lambda}\Big\langle g(|\nabla I(c(\cdot))|), |c'(\cdot)|\Big\rangle, \quad (1.3.5)$$

where $\lambda \in \mathbf{R}$ is a parameter and $\langle \cdot, \cdot \rangle$ is the inner product defined in (1.3.4). This can be interpreted as a weighted Euclidean arc length. Equation (1.3.5) is known as the Geodesic Active Contours Model. In both the Snake and Active Contour models, the curve is usually moved according to steepest descent of the energy functional (in the sense of the variational calculus). A well known example of steepest descent is that of minimizing the length of a curve c in 2-D space, obtained when $g = C$ is constant in equation (1.3.5). If E is defined by

$$E(c(t,\cdot)) = \int |c'(t,s)|ds = \int \sqrt{x'(t,s)^2 + y'(t,s)^2}ds,$$

where the derivatives are with respect to s then the energy is minimized when

$$\left|\frac{dc}{dt}\right| \propto -\kappa,$$

where κ is the curvature, assuming the front moves normal to itself.

The Snake and Active Contour models have proved useful in the medical community. One possible reason is that medical applications can cover a vast variety of image structures, such as membranes, tumours, blood vessels (long and thinny). Also, these models are intuitively easy to understand, which is suitable for user interaction. However, there are many serious drawbacks with both models: a contour cannot detect more than one object unless topological changes (breaking and merging) are allowed; but keeping track of a variable number of closed curves is an extremely difficult problem. Secondly, even in the case of detecting a single object, the initial curve must be reasonably close to the desired curve, otherwise only

a local minimum will be computed. Indeed, many applications of the Snake and Active Contour models require user-interaction to specify the initial curve and it is clearly advantageous to be able to automate this choice. Another difficulty is that the number of markers is fixed, causing numerical difficulties when markers cluster near each other (concentration) and/or leave "empty" regions (voiding).

Finally, we note that it is possible to combine region-based and edge-based methods to exploit the advantages of both. This was attempted by Zhu and Yuille [95, 96] and their experiments have yielded promising experimental results on a wide variety of images.

1.3.3 The Level Set Method

To address the above issues, Osher and Sethian [63, 77] proposed an effective representation of evolving curves known as the Level Set. Note that, unlike Snakes or Active Contours, the concept of Level Sets is not really a "model" but is rather a method for avoiding the above-mentioned difficulties in implementing the Snake and Active Contour models. However it is natural to discuss the latter in conjunction with the former. The basic idea is simple: a curve, or more precisely, a set of curves, can be implicitly represented as the set of points (x, y) satisfying $F(x, y) = 0$ for a certain continuous function F. The curve naturally divides the region of interest into where F is positive or negative. Level Sets can be thought of as a generalization of Snakes or Active Contours by allowing topological changes to be easily handled. A number of Level Set variational frameworks have been proposed by various researchers (Chan and Vese, [21], Samson et al., [76], Yezzi et al., [92]). Using the Level Set formulation, a moving boundary can be represented as a time-dependent function $\phi = \phi(x, y, t)$ where the zero-level of ϕ corresponds to the position of the curve for each time t. For example if a curve c evolves according to $dc/dt = RN$ where R is scalar curvature and N is inward normal then the Level Set formulation is given by

$$\frac{d\phi}{dt} = |\nabla\phi| div \frac{\nabla\phi}{|\nabla\phi|}. \tag{1.3.6}$$

More generally, if a front moves normal to itself with speed[3] F, that is, $dc/dt = F\vec{N}$, where $\vec{N} = \phi/|\phi|$ then the Level Set formulation is given by

$$\phi_t + F|\nabla\phi| = 0. \tag{1.3.7}$$

Here we assumed that each point on the front $c(s,t)$ moves along the front normal, otherwise we can reparametrize $t := t'$ so that it does.

In practice F is represented as values on a lattice. That is, $F(x,y) \in R$ where $(x,y) = (ih, jh)$ for integers i, j. It is immediately clear that the Level Set formulation avoids many of the difficulties associated with Snakes and Active Contours. For instance a Level Set automatically handles topological changes and also avoids clustering of markers and "void" regions. Moreover, the formulation can easily be extended to more than two dimensions. The most obvious disadvantage of Level Sets is computational complexity since an "extra dimension" is being used to model an image. For example, if we consider the problem of implementing (1.3.6) numerically on a grid of N^2 points and T time iterations then the complexity of the algorithm would be $\mathcal{O}(N^2T)$. The generally accepted solution is the so-called narrow band method (Adalsteinsson and Sethian, [1], Chopp, [23]). The values of ϕ are updated only near the zero Level Set. A common assumption is that the zero Level Set is approximated as the set of pixels where adjacent pixels differ in sign. We apply (1.3.7) only at points "sufficiently nearby" the zero Level Set, until the curve stops near the image boundary or reaches the edge of the "narrow band". In the latter case, a new narrow band is constructed and the process repeated.

[3]The speed F is not necessarily constant and can depend on, for example, the curvature of the front.

1.3.4 The Topological Snake Model

As is apparent from above discussion, implementing Level Set algorithms are far from trivial. Moreover, Level Sets are less convenient for user interaction and mathematical analysis. It is natural to ask if one can overcome the difficulties of traditional Snake methods without the disadvantages of Level Sets. One idea is topologically adaptable snakes, or T-snakes, due to Terzopoulos and Mcinerney [52, 53]. As the name implies, T-snakes attempt to overcome the disadvantage of the fixed topology by allowing the number of snakes to change, either by splitting one snake in two pieces or vice versa. The basic idea is if the motion of a snake causes itself to overlap with either itself or another snake then a topological change takes place. Assuming a snake is represented with N markers, it requires $\mathcal{O}(N)$ operations to check if an overlap occurs around one marker which is clearly inconvenient. Therefore it is necessary to impose a grid, where each discrete cell records whether it is in contact with a snake, and if so, which snake marker it touches. In effect the Topological Snake model attempts to combine the advantages of Lagrangian motion (representing a snake as a linked list of markers) and Eulerian motion (using a discrete grid). Furthermore the Eulerian grid allows a natural method of avoiding the problems associated with concentration and voiding. If adjacent markers are too far apart, then additional markers can be added. Conversely, if too many markers are concentrated at an area, then they can be removed (Figure 1.3.2). This means the number of markers is not constant. Although this implies a slight complication in the algorithmic implementation it is much less cumbersome than the difficulties associated with Level Sets. Each curve of a topological snake can move according to the same laws of motion derived for the Snake and Active Contour models.

Note that if starting from the outside of the boundary, it is impossible to detect strong edges *within* an object. A possible solution is the use of so-called "dual" T-snakes, (Giraldi et al., [35]), where snakes can move within an object as well as without. Thus dual T-snakes allow for greater flexibility in initializing the snakes.

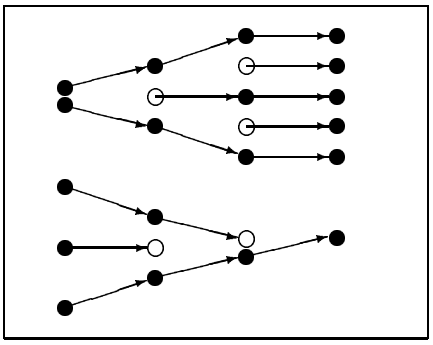


Figure 1.3.2: Adding and deleting markers for the Topological Snake model.

1.3.5 Stochastic Models

This section describes a recent approach to image segmentation borrowing ideas from Monte Carlo methods. The basic idea is that a Markov Field can be defined on the discrete field or lattice. A segmentation is "good" if it has a (relatively) high probability of occurring given the image data. The calculation of the optimal segmentation is based on Bayes law. In other words we seek to maximise

$$P(\theta|g) = \frac{P(g|\theta)P(\theta)}{P(g)}, \tag{1.3.8}$$

where θ is the segmentation model and g the data. The first term of the numerator measures how well the data fits the model. The second term of the numerator measures the simplicity of the model, which will usually be defined a-priori. The denominator is constant for given g and is almost always dropped in applications. Note that this differs from the variational method in that we seek to maximize a probability distribution instead of minimizing an energy functional. But by taking negative logarithms of (1.3.8) we see that maximizing the probability and minimizing an energy functional are equivalent in some sense. Indeed, this idea is used in Chapter 3.

A typical stochastic model is hierarchical with two levels. On the higher level a Gibbs distribution characterizes how pixels are clustered into regions that are homogenous in some sense, for example they have same textures or other features. The lower level also uses a Gibbs distribution which describes the individual texture or feature itself. We will not discuss the elementary theory of stochastic models in detail and instead refer the reader to (Derin and Elliott, [30], Dubes and Jain, [31], Geman and Geman, [33]). Dubes and Jain [31] also provides a sound review of stochastic models used in Image Processing.

Stochastic models have proved useful for texture segmentation. Among the most well-known are the Ising and Potts models (Chandler, [22]). The celebrated Geman and Geman model [33] is a forerunner of the Mumford-Shah model since it attempts to model a line process as well as pixel intensities. On the other hand, texture segmentation can be achieved without employing stochastic methods, through the use of texture features (Haralick et al., [38]).

1.4 Advantages of the Variational Formulation

It is almost universally accepted that the Variational Formulation represents a significant improvement in the development of segmentation algorithms. Morel and Solimini [58] mention a number of arguments supporting the superiority of algorithms developed in the Variational Formulation over other older "heuristic" methods. Firstly, heuristic methods are necessarily complex while the Variational Formulation is usually much simpler. Secondly, the energy functional allows one to compare two segmentations and specify that one is better than the other. Thirdly, the Variational Formulation can be seen as a general framework for many heuristic algorithms, in the sense that one can translate such algorithms via a suitable functional. In the case of algorithms such as (Perkins, [65]), it can be argued that many segmentation or edge detection algorithms can be regarded as equivalent in some

sense to a corresponding algorithm developed in the Variational Formulation. The latter algorithm will often result in segmentations satisfying the properties that are sought by the former. Thus it is completely unnecessary to specify a complex algorithm such as that in (Perkins, [65]) when a much simpler algorithm yields similar results. This observation is examined in detail in (Morel and Solimini, [58]).

1.5 Outline of the Thesis

The thesis is organized as follows: In the second chapter, we describe the Mumford-Shah functional and region merging algorithms. We describe a simple algorithm from Koepfler et al. and an improvement of this algorithm. We show the new algorithm has important advantages: The critical values of scale parameter (where the optimal segmentation changes) can be determined directly from the image data instead of being "guessed". Secondly, the globally "best" merge at each stage of the region merging process can be determined in reasonable time. Also, the new algorithm prefers to merge small regions "evenly" across the entire image instead of favouring one large region accumulating single pixels. In Chapter 3[4], we describe an important theorem for the Mumford-Shah model and how it generalizes to an extended model. This chapter is an extended version of [83]. In Chapter 4, we compute an explicit minimizer of a simple image for the extended model. In Chapter 5[5], we describe the "small sample problem" and its solution. The small sample problem is a consequence of extending the model and it arises because we require an extra "regularization" parameter to handle numerical instabilities when dealing with small-size regions. We describe an alternative approach to eliminate this problem. The alternative approach eliminates the use of the extra scale parameter and also succeeds in producing better segmentations in less time, albeit with a

[4]The contents of this chapter have been accepted for publication in the Journal of Mathematical Imaging and Vision.

[5]The contents of this chapter have been submitted to the Journal of Pattern Recognition.

few trade-offs: the mathematical analysis is less elegant, and it is difficult to generalize this to vector-valued images, hence this improvement is only of theoretical importance. Chapter 6 describes the application of the extended model to textured images. Chapter 7 discusses an important issue: that of automatic parameter-value selection. Equation (1.3.2) contains one scale parameter and it is common practice to find a good value by trial and error either by running an algorithm multiple times, or computing a multiscale hierarchy of segmentations first and letting the user choose. There are advantages in being able to automate the selection of a scale parameter value. We believe this issue is neglected in the literature. We propose two methods of determining an optimal trade-off: measuring merge significance and modelling the merge cost. In the former, we define a simple measure of the "significance" of each merge and choose the best segmentation as that corresponding to the most significant merge. We show that this measure is unsatisfactory. In the latter, we propose an "ideal model" of merge cost versus merge number and choose the best segmentation as that corresponding to the maximum deviation from this graph. We show that the latter method yields better results.

Chapter 2

The Basic Mumford-Shah Functional and Region Merging Algorithms

2.1 Introduction

In Chapter 1, we remarked that the idea of a Variational Formulation is to define an energy functional so that a lower energy indicates a better segmentation. We assume the segmentation represents both the boundary set K and an approximating function u of an image g. We seek a segmentation that satisfies two properties: (i) the segmentation must somehow be simple and (ii) the segmentation must also be meaningful with respect to the given image. The latter requirement simply states that u must be a reasonable approximation to g. Thus it is necessary to define some measure of how simple a segmentation actually is and the quality of the approximation u with respect to g. There are several alternatives to consider: we can measure the simplicity of a segmentation by the number of regions, number of edges, boundary length and so on. Similar considerations hold for the quality of the approximation.

In 1989, Mumford and Shah (Mumford and Shah, [61]) proposed to minimize

$$E(u,K) = \mu^2 \int_{\Omega/K} (u(x) - g(x))^2 dx + \int_{\Omega/K} |\nabla u(x)|^2 dx + \nu \cdot \ell(K), \qquad (2.1.1)$$

where μ^2, ν are scale parameters, u is an approximation of g, ∇u is the gradient of u, and K is assumed a-priori to be a piecewise smooth boundary set. More specifically, K is a finite set of singular points joined by a finite set of C^1 arcs, with total length $\ell(K)$. Equation (2.1.1) gives a relatively simple functional consisting of three terms. The idea of the model u is to specify what is meant by "homogenous" and "different". The various terms in the functional can be explained as follows:

- the first term measures how well the image model fits the data g.
- the second term measures the smoothness of the image model.
- the final term measures the complexity of the image model.

Mumford and Shah conjectured the existence of minimizers. More specifically, they conjectured that for all continuous functions g, E has a minimum in the set of all pairs (f, K) with f differentiable on each region R_i and K being a finite set of singular points joined by a finite set of C^1 arcs.

This conjecture remains open and only a number of "partial" results are known (Chambolle, [16]). Despite being weaker than the original conjecture they are still considered meaningful in practice (Morel and Solimini, [58, 59]). As stated in the first chapter, we will be mainly concerned with the simplest model, the piecewise constant model

$$E(u,K) = \int (u - g)^2 dx + \lambda \cdot \ell(K). \qquad (2.1.2)$$

This is sometimes called the "cartoon limit" of the original functional (2.1.1) (Cremers et al., [25]). If we fix $\nu/\mu^2 = \lambda$ to be constant and take the limit $\mu \to 0, \nu \to 0$ then minimizing equations (2.1.1) or (2.1.2) are equivalent.

The mathematical analysis of this model is considered complete (Koepfler, [42]). For instance the conjecture is known to be true for the piecewise constant model, as was proved by Mumford and Shah [61]. An elementary constructive proof was given by Morel and Solimini in [56, 57] and Chapter 5 of [58]. In this chapter, we discuss the Mumford-Shah functional and a region merging algorithm due to Koepfler [42]. We then discuss an improvement of this algorithm and give a theoretical comparison between the two.

2.2 Discussion of the Mumford-Shah Functional

In equation (2.1.2) the parameter λ controls the trade-off between the complexity and accuracy of the segmentation. If λ is small, a fine segmentation will result. When λ is increased, we would expect a coarser segmentation with a greater error in approximating g with u.

From Chapter 1 we recall the properties of region merging algorithms. At each iteration of the region merging process, the only operation allowed is the merging of two adjacent regions, provided it decreases the energy functional. We assume that each region's area and mean gray value are recorded, and that these quantities are always updated during the merging process. We now list some fundamental properties of region merging algorithms. Firstly, given a boundary set K the optimal u is obtained by taking the mean value of g on each region. That is, $u(x) = u_i = \frac{1}{|R_i|}\int_{R_i} g dx$ when $x \in R_i$. We therefore assume in the following that after merging two regions, the mean values are updated automatically. In other words E is a function of K alone and hence we will write $E(K)$ instead of $E(u, K)$.

Secondly, suppose we merge two regions R_i, R_j into R_{ij}. Then

1. The areas and mean values are updated via

$$|R_{ij}| = |R_i| + |R_j|, \tag{2.2.1}$$
$$u_{ij} = \frac{|R_i|u_i + |R_j|u_j}{|R_{ij}|}. \tag{2.2.2}$$

2. The corresponding change of energy is calculated via

$$\begin{aligned} E(K') - E(K) &= |R_{ij}|\sigma_{ij}^{*2} - |R_j|\sigma_j^{*2} \\ &\quad -|R_i|\sigma_i^{*2} - \lambda \cdot \ell(\partial(R_i, R_j)) \\ &= \frac{|R_i||R_j|}{|R_{ij}|}(u_i - u_j)^2 - \lambda \cdot \ell(\partial(R_i, R_j)), \end{aligned} \tag{2.2.3}$$

   where $\sigma_i^{*2} = \frac{1}{|R_i|}\int_{R_i}(g-u)^2$ is the "sample variance" and similarly for $\sigma_j^{*2}, \sigma_{ij}^{*2}$.

In Chapter 1 we gave a brief outline of a simplified version of Koepfler's region merging algorithm:

1. Initialization: set $u_0 = g$ and K_0 is the union of all boundaries of all pixels. Set $\lambda_0 = 0$.

2. Recursive merging: merge all pairs of regions whose merging decreases the energy.

3. Change of scale: increase λ and return to the previous step.

To demonstrate the utility of the algorithm, Koepfler defined two basic requirements that must be satisfied for the algorithm to be useful in practice:

- Correctedness: if g is piecewise constant then there exists a value $\lambda_0 > 0$ such that for every $0 < \lambda < \lambda_0$ the segmentation (u, K) obtained by the algorithm satisfies $u = g$ and K is the union of the boundaries of the regions where g is constant.

- Strong causality: if $\lambda_2 > \lambda_1$ then the boundaries provided by the algorithm for λ_2 are contained in those obtained for λ_1, and the regions of the segmentation associated to λ_2 are the unions of some of the regions obtained for λ_1.

We assume that the image domain Ω is a rectangle, divided into equi-sized square "pixels" and g is constant on each pixel.

The property of correctedness ensures that it is at least possible to recover a simple segmentation: where two adjacent pixels belong to the same region if and only if their gray values are equal. The property of strong causality is clearly desirable since it allows one to easily compute a fast multiscale hierarchy of segmentations. Note that we are speaking about a sub-optimal segmentation: the segmentation generated by the algorithm and not the theoretical optimal segmentation defined by the energy functional. Koepfler's algorithm satisfies both correctedness and strong causality. We present our proofs of these facts.

Proof:

Let g be defined on a rectangle and set

$$\epsilon = \inf_{x,y\in\Omega, g(x)\neq g(y)} |g(x) - g(y)|. \tag{2.2.4}$$

Clearly $\epsilon > 0$ since there is a finite number of pixels and hence $g(x) - g(y)$ can only take a finite number of different values. Set $\lambda < \lambda_0 = \epsilon^2/3N$ where N is the boundary length of K_0. When two regions are merged the change in energy is

$$E(K') - E(K) = \frac{|R_i||R_j|}{|R_{ij}|}(u_i - u_j)^2 - \lambda \cdot \ell(\partial(R_i, R_j)), \tag{2.2.5}$$

where $K(K')$ is the segmentation before(after) the region merge and $\partial(R_i, R_j)$ is the common boundary of the two regions. Note that $\partial(R_i, R_j)$ is not necessarily a single curve. Let us call a region uniform if all its pixels have equal gray value. If two uniform regions have the same gray value, then they can be merged since the first term vanishes and the second is negative. If two uniform regions have different gray values then

$$\begin{aligned} E(K') - E(K) &\geq \frac{|R_i||R_j|}{|R_{ij}|}\epsilon^2 - \frac{\epsilon^2}{3N}N \\ &\geq \frac{1}{2}\epsilon^2 - \frac{1}{3}\epsilon^2 > 0, \end{aligned}$$

so they cannot be merged. It follows that during the segmentation process it is impossible to merge two uniform regions with different gray values and thus all regions remain uniform. Also it is easily seen that if two adjacent pixels have equal gray values they must eventually be merged. Thus correctedness holds.

The property of causality immediately follows from the definition of the algorithm. If $\lambda_1 < \lambda_2$ then the segmentation corresponding to λ_2 is obtained by performing region merging operations starting with the segmentation corresponding to λ_1. ∎

2.3 Koepfler's Algorithm

We now discuss Koepfler's algorithm in some detail. The essential algorithm runs as follows:

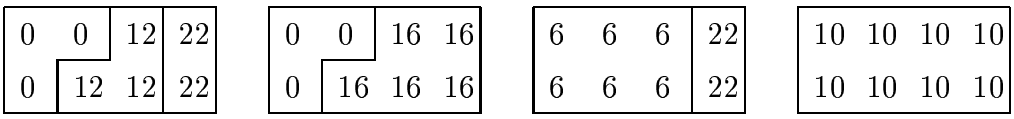
We are given an image g defined on a domain Ω. Ω is a rectangle divided into pixels with g constant on each pixel. A set of values of λ must be defined a-priori: $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ with $\lambda_i < \lambda_{i+1}$ for each i. All regions must be sorted in a "region list" (Koepfler et al. do not specify how to order the regions in this list).

1. For the initial segmentation, assume all pixels are separate regions. We will call this the **trivial segmentation**. Set $\lambda = \lambda_1$.

2. Take the first region in the list and determine which of its adjacent regions yield the maximal decrease of energy (Koepfler et al. do not specify how to break "ties" when two or more neighbours yield the same decrease of energy). If such a neighbouring region exists, merge the two and proceed to check the next region in the list. Continue merging until no further decrease in the energy functional is possible.

3. For every λ_i calculate a segmentation by iterating step 2 above. The algorithm stops if there is just one region left or after computing a segmentation using λ_n.

There are two difficulties with this algorithm. The first problem lies with the selection of the set of values $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$. The most obvious choice of Λ is a set of values λ_i increasing linearly. That is, $\Lambda = \{\delta, 2\delta \ldots N\delta\}$ for some integer N and real δ [66]. However, we still have the problem of selecting values for N and δ. In any case Λ must be chosen a-priori, implying a trade-off between the accuracy of the segmentation and computation speed. If the values of λ_i are too sparse, then it is possible to miss the correct segmentation altogether. If the values of λ_i are too dense, the algorithm will be too slow.



S1: $E = 5\lambda$ S2: $E = 120 + 3\lambda$ S3: $E = 216 + 2\lambda$ S4: $E = 600$

Figure 2.3.1: Four different segmentations obtainable by region merging.

The second problem is determining which order the regions are to be merged. This is ambiguous in Koepfler's algorithm since the order will depend on the state of the region list at each iteration. This causes serious difficulties in the theoretical analysis of the algorithm. It is natural to ask if, when selecting which pair of regions to merge, why not choose the globally best merge out of all possible pairs? Koepfler's algorithm deliberately avoids the globally best merge simply for reasons of speed. The search for the globally best merge can slow down the algorithm considerably, since it implies the sorting of a list of all possible merges, and during the initial stages, there are order $\mathcal{O}(4N^2)$ possible pairs of neighbouring regions from an initial datum of N^2 pixels.

Note that in Koepfler's algorithm it only makes sense to apply the definition of strong causality for values of $\lambda_1, \lambda_2 \in \Lambda$, and not $\lambda_1, \lambda_2 \in \mathbf{R}$. Furthermore the segmentation corresponding to any λ depends not only on g and λ but also Λ itself. A demonstration follows:

In Figure 2.3.1 let us assume that S1 is the initial segmentation and $g(\mathbf{x}) = u_1(\mathbf{x}) = 0, 12$ or 22 as shown in the corresponding diagram. Note that S1 is not the trivial segmentation but we have simplified matters by assuming all adjacent pixels with equal gray values have already been merged and it only remains to consider segmentations that can be obtained by applying region merging operations to S1. There are only four such segmentations, shown above. Recalling the formula (2.2.2) for updating the mean u at each region, we easily check $u_2(\mathbf{x}), u_3(\mathbf{x}), u_4(\mathbf{x})$ are given

Range of λ	Best segmentation
$0 < \lambda < 60$	S1
$60 < \lambda < 96$	S2
$96 < \lambda < 192$	S3
$192 < \lambda < \infty$	S4

Table 2.3.1: Best segmentation for different values of λ.

by the second, third and fourth diagrams of Figure 2.3.1 respectively. For any λ it is not hard to calculate which of the above four segmentations is optimal according to the definition of E. The results are summarized in Table 2.3.1. From this table it will be observed that any algorithm that always returns the globally optimal segmentation would not satisfy the property of causality since S3 is not a subset of S2.

Assume that $\Lambda = \{66, 100\}$. We start with S1. When $\lambda = 66$ we find only one possible pair of regions, which must be merged: namely, the middle and right regions, yielding S2. We then increase λ to 100 and find it is impossible to merge the remaining two regions. Hence the segmentation corresponding to $\lambda = 100$ is S2.

Now consider $\Lambda = \{100\}$. We start with S1. When $\lambda = 100$, merging two regions will produce either S2 or S3. It will be noted that both lead to a decrease in energy, but merging all three regions actually increases the energy. This shows an ambiguity in the algorithm: we do not know which pair of regions to merge since the choice depends on the region list in step 2. Let us assume we choose S3 since $E(\text{S3}) < E(\text{S2})$, which is equivalent to choosing a globally best merge. Thus the segmentation corresponding to $\lambda = 100$ is S3, not S2. Therefore, even if Koepfler's algorithm always selected the globally best merge at each iteration, the segmentation output is not only a function of λ but also Λ.

2.4 An Improvement of Koepfler's Algorithm

Redding et al. [67] show it is possible to address the above difficulties. They refer to Λ as a "λ-schedule" and show that by careful consideration of the energy functional, it is in fact possible to avoid the unpleasant trade-off between the λ-schedule being too sparse or too dense. The essential idea is that instead of guessing a λ-schedule, it is possible to determine the ideal schedule straight from the image data. As a result, it is possible to compute all segmentations efficiently, as if using an extremely dense λ-schedule without significant time-complexity. For this reason, Redding et al. call their algorithm the Full Lambda-Schedule Algorithm (FLSA).

They also demonstrate that at the cost of reasonably complex data structures, it is possible to find the globally best merge in reasonable time. We refer the reader to (Robinson et al., [72]) for a detailed implementation of the algorithm. We now discuss the FLSA.

Recall that when two regions are merged the change in energy is

$$\begin{aligned} E(K') - E(K) &= \frac{|R_i||R_j|}{|R_{ij}|}(u_1 - u_2)^2 - \lambda \cdot \ell(\partial(R_i, R_j)) && (2.4.1) \\ &= \Delta M - \lambda \cdot \Delta L, && (2.4.2) \end{aligned}$$

where ΔM and ΔL represent the change in model accuracy and boundary length terms respectively. The criteria for whether merging two regions is possible is if it (i) decreases the energy and (ii) is the best among all possible merges. From (i) it is seen that the critical value of λ is

$$\lambda_{crit} = \frac{|R_i||R_j|}{|R_{ij}|} \cdot \frac{(u_1 - u_2)^2}{\ell(\partial(R_i, R_j))}. \qquad (2.4.3)$$

We call this value of λ the "merge cost". If $\lambda < \lambda_{crit}$, merging will not occur, since it increases the energy. If $\lambda > \lambda_{crit}$ then merging will occur, unless another merge produces a lower value of λ.

In the formulation of the FLSA the criterion for which merge is the best among all possible merges no longer equates to the greatest decrease in E. Instead we choose whichever pair yields the smallest merge cost. Thus we are able to dispense with the λ-schedule altogether. However, we do need to maintain one parameter, namely the "stopping λ", denoted by λ_{stop}. Roughly speaking, this corresponds to the "scale" of the desired segmentation in which we seek and λ_{stop} is, in some sense, equivalent to the greatest value in the λ-schedule in Koepfler's algorithm. The modified algorithm is now as follows:

1. Take the pixels of the image as the initial segmentation.

2. Determine which pair of adjacent regions yields the smallest merge cost. If such a pair exists and the merge cost is less than λ_{stop} then merge the two.

3. Iterate step 2. The algorithm stops if there is just one region left or if the merge cost exceeds λ_{stop} for all possible merges.

In the case of "ties" (two or more merging operations with the same energy cost) we require a deterministic procedure to choose one merging operation over another. We label all pixels with coordinates (i, j) and sort the pixels in an arbitrary order, e.g. $(i, j) < (i', j')$ if either $i < i'$ or $i = i'$ and $j < j'$. A region R_1 takes precedence over R_2 if the "earliest" pixel (i_1, j_1) in R_1 is less than that of R_2. This approach is used in Redding et al. [67]. Note that this method of breaking ties can also be applied to Koepfler's algorithm. Thus at each step, the correct region to merge is uniquely determined. In this way, we can obtain for any value of λ_{stop} a unique segmentation given only the image g. It is easily seen that this satisfies the property of correctedness. Starting with the trivial segmentation, merging two uniform regions yields a merge cost of zero if and only if their gray values are equal. Thus, by always choosing the region merges with smallest merge cost, all such merges yielding a merge cost of zero will be performed first. By setting λ positive but sufficiently

Range of λ	Segmentation obtained from FLSA
$0 < \lambda < 60$	S1
$60 < \lambda < 160$	S2
$160 < \lambda < \infty$	S4

Table 2.4.1: Segmentation obtained by the FLSA for different values of λ.

small we can also prevent merging two uniform regions with unequal gray values. For example we can set $\lambda < \lambda_0 = \epsilon^2/3N$, where

$$\epsilon = \inf_{x,y\in\Omega, g(x)\neq g(y)} |g(x) - g(y)|,$$

and N is the boundary length of the trivial segmentation. The rest follows as in the proof of correctedness for Koepfler's algorithm. It is also easy to check that causality holds: if $\lambda_2 > \lambda_1$ then the boundaries provided by the algorithm for $\lambda_{stop} = \lambda_2$ are contained in those obtained for $\lambda_{stop} = \lambda_1$, and the regions of the segmentation associated to $\lambda_{stop} = \lambda_2$ are the unions of some of the regions obtained for $\lambda_{stop} = \lambda_1$.

Note that it is possible for the merge cost to temporarily decrease in the short-term. For example a region-merge may yield a merge cost of λ_1, say, and the next region merge may yield a merge cost of λ_2 with $\lambda_2 < \lambda_1$. But we would generally expect that during region merging, the merge costs will increase in the long run. Note that this does not violate the property of causality.

Using the FLSA we obtain, for any λ, the segmentation indicated in Table 2.4.1. The property of causality is clear. Note that the segmentation S3 is never recovered for any value of λ.

A reasonable question we can ask is whether the FLSA is indeed consistent with Koepfler's algorithm, in the sense they yield the same segmentations. More specifically, we ask: suppose that we are given an image g and a λ_{stop}. Then running the

FLSA must necessarily give a unique segmentation. Does there exist a sufficiently dense λ-schedule that will guarantee Koepfler's algorithm produces the same output as the FLSA? We assume that for purposes of running Koepfler's algorithm, the globally best merge is always sought and there is no issue with computation-time.

The answer is yes, if we assume the following: the merge costs increase monotonically over time. We prove this as follows: since Ω has a finite number of pixels there are only finitely many ways to form a region using any number of pixels. Similarly, there must be only finitely many ways to form two adjacent regions. It follows that whenever two regions are merged, the merge cost must be an element of a finite set $\Lambda = \{\lambda_0, \lambda_1, \ldots, \lambda_{stop}\}$. Choose Λ to be the λ-schedule. It is not hard to see that both Koepfler's algorithm and the FLSA must perform the same region merges in the same order and the end results must therefore be identical.

We should point out that the assumption of monotonically increasing merge costs is almost always false, hence we cannot expect our output to exactly match that produced by Koepfler's algorithm. But the merge costs always increase in the long run. Our experiments yield visually comparable results with that of Koepfler's. We conclude this chapter by demonstrating an interesting comparison between the FLSA and Koepfler's algorithm.

Lemma 2.4.1 *Suppose that the FLSA and Koepfler's algorithm are run on the same image. Assume that Koepfler's algorithm uses the singleton λ-schedule $\Lambda = \{\lambda\}$ and that the FLSA uses $\lambda_{stop} = \lambda$. We also assume that Koepfler's algorithm is "allowed" to always select the globally best merge at each iteration (since it is clear that Koepfler et al. "wish" to be able to do this). Suppose that during the region merging process there exist two pairs of adjacent regions AB and CD for both the FLSA and Koepfler's algorithm. Suppose that both the FLSA and Koepfler's algorithm merge AB and CD at some point during the process, but the FLSA merges AB before CD and Koepfler's algorithm merges CD before AB. Then AB has a smaller change in ΔL and ΔM.*

This means the FLSA has a stronger preference for merging smaller regions before larger regions (since the lower values of ΔL and ΔM imply the region sizes are more likely to be small also). In other words, region growing takes place "uniformly" across the entire image. We now prove this lemma.

Proof:

Recall that when two regions are merged the change in energy is

$$E(K) - E(K') = \Delta M - \lambda \Delta L,$$

where ΔM and ΔL are given by (2.4.1-2.4.2). For the Mumford-Shah model (2.1.2), this means that

$$\begin{aligned} \Delta M &= \frac{|R_i||R_j|}{|R_{ij}|}(u_1 - u_2)^2, \\ \Delta L &= \ell(\partial(R_i, R_j)), \end{aligned}$$

for any adjacent pair of regions R_i, R_j. In the FLSA, the merge cost is given by

$$\lambda = \frac{\Delta M}{\Delta L}.$$

Let $\Delta M_{AB}, \Delta M_{CD}, \Delta L_{AB}, \Delta L_{CD}$ be the corresponding changes in length and model accuracy terms. We need to show that $\Delta L_{AB} < \Delta L_{CD}$ and $\Delta M_{AB} < \Delta M_{CD}$.

Recall that for a merge to take place, the energy must decrease in Koepfler's algorithm or the merge cost must be less than λ_{stop} in the FLSA. Furthermore, for both algorithms the merge must be better than all other possible merges. In other words, for Koepfler's algorithm the decrease in energy must exceed that of any other merge,

and for the FLSA, the merge cost must be less than the merge cost corresponding to any other merge. Thus

$$\begin{aligned}
\lambda\Delta L_{CD} - \Delta M_{CD} &> 0, \\
\frac{\Delta M_{AB}}{\Delta L_{AB}} &< \lambda, \\
\lambda\Delta L_{CD} - \Delta M_{CD} &> \lambda\Delta L_{AB} - \Delta M_{AB}, \\
\frac{\Delta M_{AB}}{\Delta L_{AB}} &< \frac{\Delta M_{CD}}{\Delta L_{CD}}.
\end{aligned}$$

We have, subtracting λ from both sides of the last equation,

$$\begin{aligned}
\frac{\Delta M_{AB} - \lambda\Delta L_{AB}}{\Delta L_{AB}} &< \frac{\Delta M_{CD} - \lambda\Delta L_{CD}}{\Delta L_{CD}}, \\
\frac{\lambda\Delta L_{AB} - \Delta M_{AB}}{\Delta L_{AB}} &> \frac{\lambda\Delta L_{CD} - \Delta M_{CD}}{\Delta L_{CD}}.
\end{aligned}$$

Multiplying this by the third equation gives

$$\frac{1}{\Delta L_{AB}} > \frac{1}{\Delta L_{CD}},$$

since both sides are positive. In other words $\Delta L_{AB} < \Delta L_{CD}$. Similarly we can show that $\Delta M_{AB} < \Delta M_{CD}$. ∎

2.5 Conclusions

In this chapter we discussed Koepfler's region merging algorithm and an improved variant, called the Full Lambda Schedule Algorithm (FLSA). In both algorithms, each pixel is initially considered a separate region and the decision of which regions to merge is based on an energy functional. For each region-merge, Koepfler's algorithm seeks a locally maximal decrease of the energy functional whereas the FLSA seeks the smallest critical value of scale parameter λ, where the change of energy functional becomes zero. In both algorithms, the approximating function u is simply the mean value of the image data on each region and this can be easily updated when two regions are merged. Both algorithms satisfy the properties of (i) correctedness and (ii) causality. That is, (i) given a piecewise constant image, there exist sufficiently small values of λ that allow the algorithm to generate the "correct" segmentation and (ii) for any image increasing λ results in a subset of the previous segmentation. Koepfler's algorithm suffers two drawbacks: firstly, it is necessary to select not just a stopping value of scale parameter but also a "schedule" to compute a hierarchy of segmentations. Moreover, given any value of λ and a λ-schedule $\Lambda \ni \lambda$ the segmentation corresponding to λ still depends on Λ. Secondly, it is difficult to decide which regions will be merged next, since Koepfler's algorithm does not always select the globally best merge. The FLSA addresses both drawbacks. By defining the merge cost as the value of scale parameter required for the energy change to be zero, the need of a schedule is avoided. Also, by using moderately complex data structures the globally best region merge can be found in reasonable time. We also noted that the FLSA has a "preference" for merging smaller regions before larger ones. This implies the FLSA is more likely to grow regions evenly during region merging.

Chapter 3

Mathematical Analysis of an Extended Mumford-Shah Model for Image Segmentation

3.1 Introduction

We recall the well-known Mumford-Shah functional [61][1]:

$$E(u,K) = \mu^2 \int_{\Omega/K} (u(x) - g(x))^2 dx + \int_{\Omega/K} |\nabla u(x)|^2 dx + \nu \cdot \ell(K), \qquad (3.1.1)$$

where μ^2, ν are scale parameters, u is a piecewise smooth approximation of g and K is the boundary set, assumed to be piecewise smooth with finite total length of curves $\ell(K)$. The idea of the model u is to specify what is meant by "homogenous" and "different". Here ∇u is the gradient of u. The various terms in the functional can be explained as follows:

[1]We temporarily use μ^2, ν to be consistent with the notation in [61]. In the following, μ will take a different meaning.

- the first term measures how well the image model fits the data g.
- the second term measures the smoothness of the image model.
- the final term measures the complexity of the image model.

Thus segmentations with low energy will fit the data well with smooth simple image models. The parameters $\mu^2 > 0$ and $\nu > 0$ are preset and are chosen to control the balance between these competing entities. If μ^2 is large, then a poor fitting model is expensive, and so an optimal segmentation will have poor smoothness and high complexity. Conversely, with a small value of μ^2 and large ν we would expect a simpler model with small boundary set but large error in the fitting model. In [61], Mumford and Shah justify the use of piecewise smooth image models. They refer to such models as cartoons of the original image.

They state the segmentation problem as one of computing *optimal approximation(s)* of g by piecewise smooth function(s) u. Thus they view segmentation as computing the optimal pair (u, K) for which:

- u varies smoothly within the regions R_i defined by K.
- u varies discontinuously/rapidly across the boundary K.

In order to make the segmentation problem more tractable, Mumford and Shah then propose to restrict the class of image models to the piecewise constant functions. In this case, the energy functional simplifies to

$$E(u, K) = \mu^2 \int_{\Omega/K} (u(x) - g(x))^2 dx + \nu \cdot \ell(K). \tag{3.1.2}$$

Now fix K and let $u(x) = u_i$ on each open set R_i. It is immediately clear that the functional is minimized in the variables u_i when

$$u_i := \bar{u}_i = \text{mean}_{R_i}(g) = \frac{1}{|R_i|} \int_{R_i} g(x)\, dx,$$

where $|R_i|$ denotes the area of R_i. Considering the energy function as a functional of K alone, (3.1.2) reduces to

$$E_0(K) = \sum_i \int_{R_i} (\bar{u}_i - g(x))^2 dx + \lambda \cdot \ell(K), \tag{3.1.3}$$

where $\lambda = \nu/\mu^2$. Morel and Solimini [58] provide a detailed analysis of the piecewise constant model (3.1.3). They state Mumford and Shah's result as follows:

Theorem 3.1.1 (Morel–Solimini) *Let g be a measurable bounded function in Ω. Then the minimum of (3.1.2) or equivalently (3.1.3) is attained at some segmentation K. Moreover, the minimal boundary sets have the following geometric property: either the points of K are regular, C^1 and with curvature bounded[2] from above by $4(\sup(g) - \inf(g))^2/\lambda$, or singular points are of two types, namely, triple points where three branches meet at 120 degree angles and boundary points where K meets the boundary of Ω at right angles.*

It should be pointed out that Mumford and Shah prove an additional fact in [61]. They state that for the functional (3.1.1) it is also possible for "cracktips" to occur where a tip of a curve meets no other curve. Note that for the piecewise constant model, it does not make sense for cracktips to appear in a segmentation since the mean values on both sides of the crack must be equal, so we can decrease the energy by removing the curve. The main strength of this theorem lies in the fact that it places considerable constraints in the possible shapes of the boundary set K. In this chapter, we only generalize the theorem of (Morel and Solimini, [58]), not (Mumford and Shah, [61]).

[2]Morel and Solimini give the bound $8(\sup(g) - \inf(g))^2/\lambda$ but their proof implies that 8 can be replaced by 4.

3.2 Limitations of the Mumford-Shah Model

The Mumford-Shah model is limited by its inability to model noise or texture effects in images. In this chapter, we show how an extension of the original model can deal with such images [26]. We then provide a mathematical analysis of the new model. Our main result is a generalization of Mumford and Shah's original theorem to the extended model. Indeed, we generalize the theorem of Morel and Solimini stated above.

We extend the Mumford-Shah model by considering the image data g to be a random field over the image domain Ω. The random nature of our model captures the notion of noisy data in images. Thus we consider the problem of fitting a stochastic process/model that best explains the given data g. We use a Bayesian setting and define the optimal segmentation as that which satisfies a maximum a posteriori (MAP) criterion. In Section 2 we discuss the Bayesian model in more detail and Section 3 the proof of the theorem in the extended model. We conclude with some experiments and a short summary.

3.3 Image Segmentation in a Bayesian Setting

In this section, we use a Bayesian setting to introduce a stochastic version of the original Mumford-Shah model (Crisp and Newsam, [26]). The analysis in (Crisp and Newsam, [26]) is in the discrete domain but we will adapt the model to the continuous domain, since the proof of the theorem in (Morel and Solimini, [58]) is also established in the latter. As noted by Aubert and Kornprobst [6], imaging problems are often best understood in a continuous setting, which allows for easier interpretation of the physical constraints. Once established, results and algorithms can then be transferred to the discrete domain.

We begin with the discrete domain. In the original Mumford-Shah model, it was assumed that the pixel intensities in each region were approximated as a constant function. For the new model, we think of the pixel intensities in each region as realizations of a random field.

Formally, the two basic ingredients of a MAP likelihood approach to image processing problems are: a class of functions g which are to be considered as observed images, together with a set of image descriptors or models M which are to be fitted to the images. The MAP likelihood estimate of the best model for a given image is then

$$\hat{M} = \arg\max_{M} p(M|g),$$

where $p(M|g)$ is the likelihood of the image model M given the image data g. Using Bayes theorem and taking negative logarithms we have

$$\hat{M} = \arg\min_{M} E(M) = \arg\min_{M} \left[-\log p(g|M) - \log p(M) \right]. \tag{3.3.1}$$

In other words, the problem is reformulated as that of minimizing an energy functional. Several issues now need to be addressed:

- choosing the class of images $\mathcal{G}$ and image models $\mathcal{M}$.
- defining the prior probability distribution $p(M)$, and likelihood function $p(g|M)$.
- adaptation from the discrete to the continuous model.

In the discrete domain, a model $M \in \mathcal{M}$ consists of

$$M = (K, \theta),$$

where K is the segmentation boundary and $\theta = (\mu, \sigma^2)$ are now piecewise constant functions representing the mean and variance of each region. We assume a-priori

there exists constants A, B such that $A \leq g(\mathbf{x}) \leq B$ for all pixels $\mathbf{x} \in \Omega$. In other words $g \in \mathcal{G}$ is bounded. We say that a model is admissable if $A \leq \mu_i \leq B$ and $0 < \sigma_0^2 \leq \sigma_i^2 \leq \text{osc}^2(g)$ for each i, where $\mu_i = \mu(\mathbf{x})$ and $\sigma_i^2 = \sigma^2(\mathbf{x})$ for $\mathbf{x}$ in region R_i and $\text{osc}(g) = \sup_\Omega g - \inf_\Omega g \leq B - A$ denotes the *oscillation* of g. This is no strong assumption since g was also assumed bounded in the analysis of (Morel and Solimini, [58]). Roughly speaking, σ_0 plays the role of regularization parameter and is needed to prevent "blow up" of the energy functional E (Crisp and Newsam, [26], Tao and Crisp, [82]). The essential idea is that we replace $\sigma(\mathbf{x})$ with σ_0^2 whenever $\sigma(\mathbf{x}) < \sigma_0^2$. In what follows, we will frequently refer to σ_0^2 as an offset.

For the class of images $\mathcal{G}$ in the continuous domain, it is common to assume $g \in \mathcal{G}$ is a function of bounded variation (Ambrosio et al., [2], Evans and Gariepy, [32]). However we instead assume that g be (Borel) measurable and bounded with $A \leq g(x) \leq B$ for all x to be consistent with (Morel and Solimini, [58]) and (Crisp and Newsam, [26]). Indeed, the fact g be bounded was also essential to the proof in (Morel and Solimini, [58]).[3]

In the discrete formulation it is relatively straightforward to define the prior probability distribution and likelihood function. Following (Crisp and Newsam, [26], Tao and Crisp, [82]) we assume that g is constant on each pixel and that pixel intensities in the same region (determined by K) are realizations of i.i.d. normal distributions $N(\mu_i, \sigma_i^2)$. Thus taking the product over pixel values, we have

$$p(g|M) = \prod_{i \in \mathcal{I}} \prod_{\mathbf{x} \in R_i} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\Big(- \frac{(g(\mathbf{x}) - \mu_i)^2}{2\sigma_i^2} \Big), \tag{3.3.2}$$

where $\mathcal{I}$ is an indexing set for all regions and $\mathbf{x}$ runs over all pixels in R_i. We choose to define the prior as follows:

$$p(M) = p(K, \theta) \propto \exp(-\lambda \cdot \ell(K)). \tag{3.3.3}$$

[3]Note that $g \in BV(\mathbf{R}^n)$ does not imply g is bounded. This only holds when $n = 1$.

This says a segmentation is simpler and more likely to be selected if it has shorter boundary. Also it is independent of μ, σ, in other words it is uniformly distributed in μ, σ. Note that we avoid the use of improper priors since there are only a finite number of possible segmentations K and μ_i, σ_i^2 lie in finite intervals. A short discussion on the use of this prior is in order. There are many ways of deriving a prior. The well-known Minimum Description Length (MDL) principle (Rissanen, [71]) attempts to measure both complexity and fidelity terms in common measure, namely number of bits (Kanungo et al., [39], Leclerc, [44], Lee [45]). Therefore it eliminates the use of scale parameter. Our model is not MDL based: for instance we do not penalize the number of regions directly. Although Jeffrey's prior (Box and Tiao, [13]) is standard for the model parameters μ, σ^2, the uniform distribution is also common in the literature (Zhu and Yuille, [96]). Moreover, our choice allows one to generalize the Mumford-Shah functional, as seen below. From equations (3.3.1), (3.3.2), (3.3.3) the energy is given by (ignoring an additive constant):

$$\begin{aligned} E(K,\theta) &= \sum_{i\in\mathcal{I}} \#(R_i)\frac{\log\sigma_i^2}{2} + \sum_{i\in\mathcal{I}}\sum_{\mathbf{x}\in R_i}\frac{(g(\mathbf{x})-\mu_i)^2}{2\sigma_i^2} + \lambda\cdot\ell(K) \\ &= \sum_{\mathbf{x}\in\Omega}\frac{\log\sigma^2(\mathbf{x})}{2} + \sum_{\mathbf{x}\in\Omega}\frac{(g(\mathbf{x})-\mu(\mathbf{x}))^2}{2\sigma^2(\mathbf{x})} + \lambda\cdot\ell(K), \end{aligned} \tag{3.3.4}$$

where $\#(R_i)$ denotes the number of pixels in region R_i. If K is fixed and $E(K,\theta)$ is minimized with respect to $\theta = (\mu, \sigma^2)$ subject to the condition that $M = (K, \theta)$ be admissable, then one can show that the optimal $(\bar{\mu}_i, \bar{\sigma}_i^2)$ are given by

$$\bar{\mu}(\mathbf{x}) = \bar{\mu}_i = \frac{1}{\#(R_i)}\sum_{\mathbf{x}\in R_i} g(\mathbf{x}), \tag{3.3.5}$$

$$\bar{\sigma}^2(\mathbf{x}) = \bar{\sigma}_i^2 = \max(\sigma_0^2, \sigma_i^{*2}) = \max\left(\sigma_0^2, \frac{1}{\#(R_i)}\sum_{\mathbf{x}\in R_i}(g(\mathbf{x})-\bar{\mu}_i)^2\right), \tag{3.3.6}$$

for all $\mathbf{x} \in R_i$. Thus we simply consider E as a functional of K alone. Notice that if σ_0 is large, we recover the Mumford-Shah functional since $\bar{\sigma}_i = \sigma_0$ is constant.

We now discuss the adaptation of the model to the continuous setting. The correspondence between the discrete and continuous versions of the Mumford-Shah energy has long been studied in literature. For instance, the well known method of Ambrosio and Tortorelli [3, 4] uses elliptic functionals to approximate (3.1.1) in the sense of Gamma convergence (De Giorgi,[34], Dal Maso, [51]) and are easily implementable numerically (Richardson, [70]). Finite element methods based on adaptive methods are also used, (Chambolle and Dal Maso, [17]). The Mumford-Shah energy functional is itself derived from the discrete energies of Blake and Zisserman [12] and particularly Geman and Geman [33]. These energies are simpler, however in two dimensions they only Gamma-converge to an anisotropic version of the Mumford-Shah functional (Chambolle, [16]). Mumford and Shah [61] do not justify (3.1.1) but merely consider the discrete/continuous equations as "equivalent". The Geman and Geman energy is

$$E(X,l) = \sum_i (X_i - g_i)^2 + \alpha \sum_{ij} (X_i - X_j)^2 (1 - l_{ij}) + \beta \sum_{ij} l_{ij}, \tag{3.3.7}$$

where α, β are parameters, X represents the approximation of the image data g, and l represents the unobserved "edge" data. The first summation is taken over all pixels i and the second and third are taken over all adjacent pixel pairs i, j. An edge between two pixels is either on or off and $l_{ij} = 1$ or 0 respectively. When considering only piecewise constant images X, (this is equivalent to setting $\alpha = \infty$) the second term disappears which yields

$$E(X,l) = \sum_i (X_i - g_i)^2 + \beta \sum_{ij} l_{ij}. \tag{3.3.8}$$

In accordance with (3.1.1) and (3.3.8) we therefore propose to minimize in the continuous domain:

$$\begin{aligned} E(K) &= \sum_i \int_{R_i} \frac{\log \bar{\sigma}_i^2}{2} dx + \sum_i \int_{R_i} \frac{(g(x) - \bar{\mu}_i)^2}{2\bar{\sigma}_i^2} dx + \lambda \cdot \ell(K) \\ &= \int_\Omega \frac{\log \bar{\sigma}^2(x)}{2} dx + \int_\Omega \frac{(g(x) - \bar{\mu}(x))^2}{2\bar{\sigma}^2(x)} dx + \lambda \cdot \ell(K) \\ &= \sum_i \frac{|R_i| \log \bar{\sigma}_i^2}{2} + \sum_i \frac{|R_i| \sigma_i^{*2}}{2} + \lambda \cdot \ell(K), \end{aligned} \tag{3.3.9}$$

where $\bar{\mu}_i, \bar{\sigma}_i^2$ are defined by the continuous versions of (3.3.5), (3.3.6), that is

$$\bar{\mu}(x) = \bar{\mu}_i = \frac{1}{|R_i|} \int_{R_i} g(\xi) d\xi \quad \text{for all} \quad x \in R_i, \tag{3.3.10}$$

$$\begin{aligned} \bar{\sigma}^2(x) = \bar{\sigma}_i^2 &= \max(\sigma_0^2, \sigma_i^{*2}) \\ &= \max\left(\sigma_0^2, \frac{1}{|R_i|} \int_{R_i} (g(\xi) - \bar{\mu}_i)^2 d\xi\right) \text{ for all } x \in R_i. \end{aligned} \tag{3.3.11}$$

3.4 Our Main Result for the Extended Model

We now present and prove our main result:

Theorem 3.4.1 *Let g be a bounded measurable function in Ω. Let $\sigma_0^2 > 0$ be constant. Then the minimum of (3.3.9) is obtained at some K. Moreover the minimal boundary sets have the following properties: any point x of K either belongs to a curve c which is regular, C^1, and the curvature at x is bounded by $8\text{osc}^2/\lambda\sigma_0^2$ or is an intersection point where three branches meet at 120^o angles or one curve of K meets $\partial\Omega$ at a right angle.*

We will adapt arguments of the proof of the main theorem in Morel and Solimini [58]. We start by recalling some definitions from (Morel and Solimini, [58]) but with appropriate changes for our new model:

- A *Rectifiable Curve* is a map $c(t)$ from $[0,1]$ to $\mathbf{R}^2$ such that $\ell(c) := \sup(|c(a_2) - c(a_1)| + \ldots + |c(a_n) - c(a_{n-1})|$ is finite, where the supremum is taken among all finite increasing sequences of real numbers in $[0,1]$. The real number $\ell(c)$ is called the *length* of c. Let c be a rectifiable curve and let $\sigma(t) = \frac{\ell(c_t)}{\ell(c)}$ where c_t is the restriction of the curve c to the interval $[0,t]$[4]. Then we can reparametrize c by setting, for any s in the interval $[0,1]$, $c^1(s) = c(\sigma^{-1}(s))$. One deduces easily from the definition of ℓ and the triangular inequality that $|c^1(s) - c^1(s')| \leq \ell(c)|(s - s')|$. Conversely, if $c(t)$ is a curve with Lipschitz constant L on the interval $[0,1]$, then c is rectifiable and its length is less than L. In the following, we therefore always assume that the considered rectifiable curves have been parametrized with length. We shall always identify a rectifiable curve c and its range so that "c" means both the curve and its range in the plane.

- *Ascoli-Arzela Theorem.* Let c^k be a sequence of functions which are uniformly Lipschitz on the interval $[0,1]$ and such that the set $c^k(0)$ are bounded. Then one can extract a subsequence of c^k which converges uniformly to a function c with at most the same Lipschitz constant. Applied to a sequence of curves c^k with $\ell(c^k) \leq L$ and $c^k(0)$ bounded, this theorem asserts that a subsequence of the c^k converges uniformly to some rectifiable curve c with length less than or equal to L.

- Let $c(t)$ be a C^1 rectifiable curve parametrized with length. Then one checks easily that $|c'(t)| = 1$. If $c(t)$ is twice differentiable at t we define the *curvature* at t, $curv(t)$, as the real positive number $|c''(t)|$.

[4] Although we previously used σ^2 for the sample variance of a region we decided to keep our notation consistent with (Morel and Solimini, [58]) in these definitions.

- A *Jordan curve* is a continuous curve such that for all $\sigma, \sigma' \in [0,1] : c(\sigma) \neq c(\sigma')$ unless $\{\sigma, \sigma'\} = \{0,1\}$; if $c(0) = c(1)$ then the Jordan curve is said to be *closed*. If $c(0)$ and $c(1)$ are different, we call them *tips* of the Jordan curve. All other points are called *interior points* of the Jordan curve.

- A *segmentation* is a union of a finite set of rectifiable curves.

- *Length of a segmentation* $\ell(K)$: we define $\ell(K)$ as the infimum of the lengths of all countable sets of rectifiable curves whose union is K. If, for instance, K is the union of a set of rectifiable curves meeting only at a countable set of points, it is easily seen that $\ell(K)$ is exactly the sum of the lengths of the curves.

- The *Regions of a segmentation* are the connected components of $\Omega \setminus K$. We shall denote them by $(R_i)_i$. The *two-dimensional Lebesgue measure* of R_i is denoted by $|R_i|$.

- *Convergence of a segmentation*: we shall say that a sequence of segmentations K_n *converges to a segmentation* K if each K_n is a union of Jordan curves (c_i^n) for $1 \leq i \leq k$, if each (c_i^n) tends uniformly to some curve c_i and if K is the union of the ranges of the c_i.

- The *Energy* $E(K)$: given K, θ is determined from (3.3.5, 3.3.6). Therefore we write $E(K)$ instead of $E(K, \theta)$.

- The *Common boundary* of two regions R_i and R_j is denoted by $\partial(R_i, R_j)$. It is contained in K. If $i = j$, ∂R_i denotes the boundary of R_i.

- The *Isoperimetric inequality* in $\mathbf{R}^2$ and Ω states that for any region R whose area does not exceed $|\Omega|/2$ and whose boundary is a countable union of finite curves, $\ell(\partial R \cap \Omega) \geq C_{iso}\sqrt{|R|}$ for some constant C_{iso}, where

$$0 < C_{iso} = \inf\left\{ \frac{\ell(\partial R \cap \Omega)}{\sqrt{|R|}} \quad \middle| \quad 0 < |R| \leq \frac{|\Omega|}{2}, \quad R \subset \Omega \right\} \leq 2\sqrt{\pi}$$

  with equality if and only if $\Omega = \mathbf{R}^2$. If $\Omega \subset \mathbf{R}^2$ is a rectangle of finite area then C_{iso} will depend on the length and breadth of Ω.

- *1-normal segmentations*: a segmentation is *1-normal* if it is made of a finite number of rectifiable Jordan curves, meeting each other and $\partial\Omega$ only at their tips and if *each Jordan curve separates two different regions.* This last property ensures that the number of regions in the segmentation is finite. Notice that if a Jordan curve does not separate two different regions, then we can decrease the energy by simply removing the curve. Thus this property is no restriction. We finally impose that each tip is a common tip of at least three Jordan curves (if only two such curves meet at a tip we can "merge" them into a single curve without altering the segmentation).

- The *crossing points of a 1-normal segmentation* are all the points of K where at least three Jordan curves have a common tip, or where a Jordan curve meets $\partial\Omega$ at a tip.

- The *edges of a 1-normal segmentation* are each one of the Jordan curves defining a 1-normal segmentation. The edges can be equivalently defined as all the connected components of the common boundaries $\partial(R_i, R_j)$.

- *2-normal segmentations*: a segmentation K will be called 2-normal if for every pair of adjacent regions, the new segmentation K' obtained by merging these regions satisfies $E(K') > E(K)$. (Compare to the definition of 1-normality.)

The following lemmas from (Morel and Solimini, [58]) are easily seen to hold for the MAP model. The proofs are omitted since they are just the same as those in (Morel and Solimini, [58]).

Lemma 3.4.2 (Jordan Curve Lemma) *Every closed Jordan curve c divides the plane into exactly two connected components, one bounded, "enclosed by c", and the other one unbounded.*

Lemma 3.4.3 *Let K be a 1-normal segmentation with α regions. Then K can be decomposed into a union of $\alpha - 1$ Jordan curves meeting only at a finite set of points.*

Lemma 3.4.4 *Let α be the number of regions of a 1-normal segmentation K, β the number of edges and γ the number of crossing points. If $\alpha > 1$ then*

$$\gamma \leq 2\alpha - 2 \quad \textit{and} \quad \beta \leq 3(\alpha - 1) - 2. \tag{3.4.1}$$

Remark: While Morel and Solimini do not state $\alpha > 1$, its necessity is implied in their proof, so we state it explicitly. Before continuing the proof, we need some extra preliminary results[5], concerning the effects of merging two regions of a segmentation.

[5]We deliberately state this as an unnumbered lemma to keep the numbering consistent with Morel and Solimini's original proof.

Lemma (Preliminary results): Suppose we merge two regions R_i, R_j into R_{ij}. Then

1. The model parameters $\bar{\mu}_i, \bar{\sigma}_i^2$ are updated via

$$\bar{\mu}_{ij} = \frac{|R_i|\bar{\mu}_i + |R_j|\bar{\mu}_j}{|R_{ij}|}, \tag{3.4.2}$$

$$\sigma_{ij}^{*2} = \frac{|R_i|\sigma_i^{*2} + |R_j|\sigma_j^{*2}}{|R_{ij}|} + \frac{|R_i||R_j|}{|R_{ij}|^2}(\bar{\mu}_i - \bar{\mu}_j)^2, \tag{3.4.3}$$

$$\bar{\sigma}_{ij}^2 = \max(\sigma_0^2, \sigma_{ij}^{*2}), \tag{3.4.4}$$

   where σ_i^{*2} is the true variance of region R_i, defined in (3.3.11).

2. The corresponding change of energy is calculated via

$$E(K') - E(K) = \Delta E_1 + \Delta E_2 - \Delta E_3, \tag{3.4.5}$$

   where

$$\begin{aligned}
\Delta E_1 &= \frac{|R_{ij}|}{2}\log\bar{\sigma}_{ij}^2 - \frac{|R_j|}{2}\log\bar{\sigma}_j^2 - \frac{|R_i|}{2}\log\bar{\sigma}_i^2, \\
\Delta E_2 &= \frac{|R_{ij}|\sigma_{ij}^{*2}}{2\bar{\sigma}_{ij}^2} - \frac{|R_j|\sigma_j^{*2}}{2\bar{\sigma}_j^2} - \frac{|R_i|\sigma_i^{*2}}{2\bar{\sigma}_i^2}, \\
\Delta E_3 &= \lambda \cdot l(\partial(R_i, R_j)),
\end{aligned}$$

   where $\partial(R_i, R_j)$ is the common boundary between the two regions and K, K' are the segmentations before and after merging the two regions respectively.

3. The following bounds hold:

$$|\Delta E_1| \leq \min(|R_i|, |R_j|)\frac{\mathrm{osc}^2(g)}{\sigma_0^2}, \tag{3.4.6}$$

$$|\Delta E_2| \leq \min(|R_i|, |R_j|)\frac{\mathrm{osc}^2(g)}{\sigma_0^2}. \tag{3.4.7}$$

Proof: The first item is an elementary calculation, so its proof is omitted. The second item follows immediately from (3.3.9). To prove the third item it is sufficient to show that

$$\begin{aligned} \Delta E_1 &\leq \min(|R_i|,|R_j|)\frac{\mathrm{osc}^2(g)}{\sigma_0^2}, && (3.4.8)\\ -\Delta E_1 &\leq \min(|R_i|,|R_j|)\frac{\mathrm{osc}^2(g)}{\sigma_0^2}, && (3.4.9)\\ 0 \leq \Delta E_2 &\leq \min(|R_i|,|R_j|)\frac{\mathrm{osc}^2(g)}{\sigma_0^2}. && (3.4.10) \end{aligned}$$

To show (3.4.8) we have

$$\Delta E_1 = \frac{|R_{ij}|}{2}\log\bar{\sigma}_{ij}^2 - \frac{|R_j|}{2}\log\bar{\sigma}_j^2 - \frac{|R_i|}{2}\log\bar{\sigma}_i^2.$$

If $\sigma_{ij}^{*2} \leq \sigma_0^2$ then $\Delta E_1 \leq 0$ and (3.4.8) follows immediately. If $\sigma_{ij}^{*2} \geq \sigma_0^2$ then

$$\begin{aligned} \Delta E_1 &= \frac{|R_i|}{2}\log\frac{\bar{\sigma}_{ij}^2}{\bar{\sigma}_i^2} + \frac{|R_j|}{2}\log\frac{\bar{\sigma}_{ij}^2}{\bar{\sigma}_j^2}\\ &\leq \frac{|R_i|(\sigma_{ij}^{*2}-\sigma_i^{*2})}{2\bar{\sigma}_i^2} + \frac{|R_j|(\sigma_{ij}^{*2}-\sigma_j^{*2})}{2\bar{\sigma}_j^2}\\ &= \frac{|R_i|}{2\bar{\sigma}_i^{*2}}\Big[\frac{|R_j|}{|R_{ij}|}(\sigma_j^{*2}-\sigma_i^{*2}) + \frac{|R_i||R_j|}{|R_{ij}|^2}(\bar{\mu}_i-\bar{\mu}_j)^2\Big]\\ &\quad + \frac{|R_j|}{2\bar{\sigma}_j^{*2}}\Big[\frac{|R_i|}{|R_{ij}|}(\sigma_i^{*2}-\sigma_j^{*2}) + \frac{|R_i||R_j|}{|R_{ij}|^2}(\bar{\mu}_i-\bar{\mu}_j)^2\Big]\\ &= \frac{|R_i||R_j|}{2|R_{ij}|}(\sigma_j^{*2}-\sigma_i^{*2})\Big(\frac{1}{\bar{\sigma}_i^2}-\frac{1}{\bar{\sigma}_j^2}\Big) + \frac{|R_i||R_j|(\bar{\mu}_i-\bar{\mu}_j)^2}{2|R_{ij}|}\Big(\frac{|R_i|}{|R_{ij}|\bar{\sigma}_i^2}+\frac{|R_j|}{|R_{ij}|\bar{\sigma}_j^2}\Big)\\ &\leq \min(|R_i|,|R_j|)\frac{\mathrm{osc}^2(g)}{\sigma_0^2}, \end{aligned}$$

where we have used $\log(x) \leq x-1$ for all positive x, $\bar{\sigma}_i^2 \geq \sigma_i^{*2}$, $\bar{\sigma}_j^2 \geq \sigma_j^{*2}$, $\bar{\sigma}_{ij}^2 = \sigma_{ij}^{*2}$ and $|R_i||R_j|/|R_{ij}| < \min(|R_i|,|R_j|)$.

Now consider (3.4.9). Clearly $-\Delta E_1 \leq 0$ if $\sigma_{ij}^{*2} \geq \max(\sigma_i^{*2}, \sigma_j^{*2})$. Also from (3.4.3) the case $\sigma_{ij}^{*2} < \max(\sigma_i^{*2}, \sigma_j^{*2})$ is impossible. If $\sigma_0^2 \geq \max(\sigma_i^{*2}, \sigma_j^{*2}, \sigma_{ij}^{*2})$ then $-\Delta E_1 = 0$. If $\sigma_0^2 \leq \min(\sigma_i^{*2}, \sigma_j^{*2}, \sigma_{ij}^{*2})$ then $-\Delta E_1 \leq 0$ since the function $\log(x)$ is concave and $\sigma_{ij}^{*2} \geq (|R_i|\sigma_i^{*2} + |R_j|\sigma_j^{*2})/|R_{ij}|$. Thus we only need to consider (without loss of generality) the case
$\sigma_i^{*2} \leq \min(\sigma_0^2, \sigma_{ij}^{*2}) \leq \max(\sigma_0^2, \sigma_{ij}^{*2}) \leq \sigma_j^{*2}$. But then

$$\begin{aligned}
-\Delta E_1 &= \frac{|R_i|}{2}\log\frac{\bar{\sigma}_i^2}{\bar{\sigma}_{ij}^2} + \frac{|R_j|}{2}\log\frac{\bar{\sigma}_j^2}{\bar{\sigma}_{ij}^2} \\
&\leq \frac{|R_j|}{2}\log\frac{\bar{\sigma}_j^2}{\bar{\sigma}_{ij}^2} \\
&\leq \frac{|R_j|(\sigma_j^{*2} - \sigma_{ij}^{*2})}{2\bar{\sigma}_{ij}^2} \\
&= \frac{|R_j|}{2\bar{\sigma}_{ij}^2}\Big[\frac{|R_i|}{|R_{ij}|}(\sigma_j^{*2} - \sigma_i^{*2}) - \frac{|R_i||R_j|}{|R_{ij}|^2}(\bar{\mu}_i - \bar{\mu}_j)^2\Big] \\
&\leq \min(|R_i|, |R_j|)\frac{\mathrm{osc}^2(g)}{\sigma_0^2}.
\end{aligned}$$

To show (3.4.10) observe that if $\sigma_{ij}^{*2} \geq \sigma_0^2$ then $\Delta E_2 \geq |R_{ij}| - |R_i| - |R_j| = 0$. If $\sigma_{ij}^{*2} \leq \sigma_0^2$ then

$$\Delta E_2 = \frac{|R_i|\sigma_i^{*2}}{2}\Big(\frac{1}{\bar{\sigma}_{ij}^2} - \frac{1}{\bar{\sigma}_i^2}\Big) + \frac{|R_j|\sigma_j^{*2}}{2}\Big(\frac{1}{\bar{\sigma}_{ij}^2} - \frac{1}{\bar{\sigma}_j^2}\Big) + \frac{|R_i||R_j|(\bar{\mu}_i - \bar{\mu}_j)^2}{2|R_{ij}|\bar{\sigma}_{ij}^2} \geq 0.$$

Thus the left hand side inequality of (3.4.10) holds. Now consider the right hand side inequality. If $\sigma_0^2 \leq \min(\sigma_i^{*2}, \sigma_j^{*2}, \sigma_{ij}^{*2})$ then $\Delta E_3 = 0$. If $\sigma_0^2 \geq \max(\sigma_i^{*2}, \sigma_j^{*2}, \sigma_{ij}^{*2})$ then $\Delta E_3 = |R_i||R_j|(\bar{\mu}_i - \bar{\mu}_j)^2/2|R_{ij}|\sigma_0^2 < \min(|R_i|, |R_j|)\mathrm{osc}^2(g)/\sigma_0^2$. If $\max(\sigma_i^{*2}, \sigma_j^{*2}) \leq \sigma_{ij}^{*2}$ then

$$\begin{aligned}
\Delta E_2 &= \frac{|R_i|\sigma_i^{*2}}{2}\Big(\frac{1}{\bar{\sigma}_{ij}^2} - \frac{1}{\bar{\sigma}_i^2}\Big) + \frac{|R_j|\sigma_j^{*2}}{2}\Big(\frac{1}{\bar{\sigma}_{ij}^2} - \frac{1}{\bar{\sigma}_j^2}\Big) + \frac{|R_i||R_j|(\bar{\mu}_i - \bar{\mu}_j)^2}{2|R_{ij}|\bar{\sigma}_{ij}^2} \\
&\leq \min(|R_i|, |R_j|)\frac{\mathrm{osc}^2(g)}{\sigma_0^2},
\end{aligned}$$

since the first two terms are negative. Thus it only remains to consider the case $\sigma_i^{*2} \leq \min(\sigma_0^2, \sigma_{ij}^{*2}) \leq \max(\sigma_0^2, \sigma_{ij}^{*2}) \leq \sigma_j^{*2}$. But

$$\begin{aligned}\Delta E_2 &= \frac{|R_i|\sigma_i^{*2}}{2}\left(\frac{1}{\bar{\sigma}_{ij}^2} - \frac{1}{\bar{\sigma}_i^2}\right) + \frac{|R_j|\sigma_j^{*2}}{2}\left(\frac{1}{\bar{\sigma}_{ij}^2} - \frac{1}{\bar{\sigma}_j^2}\right) + \frac{|R_i||R_j|(\bar{\mu}_i - \bar{\mu}_j)^2}{2|R_{ij}|\bar{\sigma}_{ij}^2} \\ &\leq |R_j|\frac{\mathrm{osc}^2(g)}{\sigma_0^2},\end{aligned}$$

since the first term is negative. Also

$$\begin{aligned}\Delta E_2 &\leq \frac{|R_{ij}| - |R_j|}{2} - \frac{|R_i|\sigma_i^{*2}}{2\bar{\sigma}_i^2} \\ &= \frac{|R_i|}{2}\left(1 - \frac{\sigma_i^{*2}}{\bar{\sigma}_i^2}\right) \\ &\leq \frac{|R_i|\mathrm{osc}^2(g)}{\sigma_0^2}.\end{aligned}$$

Thus $\Delta E_2 \leq \min(|R_i|, |R_j|)\mathrm{osc}^2(g)/\sigma_0^2$. ∎

Lemma 3.4.5 *Let α be the number of regions of a 2-normal segmentation K. Then*[6]

$$\alpha \quad \leq \quad \max\left(4, \quad \frac{1152|\Omega|\mathrm{osc}(g)^4}{C^2\lambda^2{\sigma_0}^4}\right), \tag{3.4.11}$$

where C is the isoperimetric constant.

Proof: This is provided in detail as there are significant differences from (Morel and Solimini, [58]) in several steps. We first observe from (3.4.6), (3.4.7) that for a 2-normal segmentation each pair of adjacent regions (R_i, R_j) in a segmentation satisfies

$$\lambda \cdot \ell(\partial(R_i, R_j)) \quad \leq \quad 2\frac{\mathrm{osc}^2(g)}{\sigma_0^2}\min(|R_i|, |R_j|). \tag{3.4.12}$$

otherwise $\Delta E_3 > \Delta E_1 + \Delta E_2$ and merging R_i, R_j will decrease the energy.

For each region R denote by $\#(\mathcal{N}(R))$ the number of neighbour regions. If $|R| \leq |\Omega|/2$ then

$$2\#(\mathcal{N}(R)) \quad \geq \quad \frac{C_{iso}\lambda\sigma_0^2}{\sqrt{|R|}\mathrm{osc}(g)^2}, \tag{3.4.13}$$

To prove this, call R_j a neighbouring region of R. The fact R, R_j cannot be merged without increasing E implies that $\lambda\cdot\ell(\partial(R, R_j)) \leq 2|R|(\mathrm{osc}(g)/\sigma_0)^2$. Thus by adding these inequalities for all neighbours of R we obtain

$$\lambda \cdot \ell(\partial R) \quad \leq \quad 2\#(\mathcal{N}(R))|R|\left(\frac{\mathrm{osc}(g)}{\sigma_0}\right)^2. \tag{3.4.14}$$

We conclude by applying to R the isoperimetric inequality in $|\Omega|$.

Without loss of generality, assume now that $\alpha \geq 4$. The union of all regions R_i is equal to the rectangle Ω and therefore $\sum_i |R_i| = |\Omega|$. Thus the number of R_i's satisfying $|R_i| \leq \min(\frac{2}{\alpha}|\Omega|, \frac{1}{2}|\Omega|) = \frac{2}{\alpha}|\Omega|$ is at least $\frac{\alpha}{2}$.

[6]The statement of the corresponding lemma in [58] is imprecise because its proof applies the isoperimetric inequality to regions without establishing the pre-condition the area is no greater than $|\Omega|/2$.

Let us now apply the previous result to all such regions R_i. Each region has at least

$$\frac{C_{iso}\lambda\sigma_0^2}{2\sqrt{|R_i|}\mathrm{osc}^2(g)} \geq C_{iso}\lambda\sqrt{\frac{\alpha}{2|\Omega|}}\cdot\frac{\sigma_0^2}{2\mathrm{osc}^2(g)}$$

neighbouring regions. Consequently the number of pairs of adjacent regions, and therefore the number β of edges satisfies

$$\beta \quad \geq \quad \frac{\alpha}{8}\cdot C_{iso}\lambda\sqrt{\frac{\alpha}{2|\Omega|}}\frac{\sigma_0^2}{\mathrm{osc}^2(g)} \quad = \quad 2^{-7/2}C_{iso}\lambda\frac{\alpha^{3/2}}{\sqrt{|\Omega|}}\frac{\sigma_0^2}{\mathrm{osc}^2(g)}.$$

Since $\beta \leq 3\alpha$ by (3.4.1) we obtain

$$\alpha \quad \leq \quad \frac{1152|\Omega|\mathrm{osc}^4}{C_{iso}^2\lambda^2{\sigma_0}^4}$$

thus proving the desired result. ∎

Remark (Elimination of small and thin regions): By analogy with Morel and Solimini's work, the last lemma can be used to show that 2-normal segmentations do not contain small or thin regions. The lower bound which equation (3.4.13) provides on the number of neighbours of a segmentation region R is inversely proportional to the "diameter" of R. This means that small regions must have many neighbours. But equation (3.4.11) bounds the total number of regions and hence provides a crude bound on the number of neighbours. Combining these bounds gives

$$\max\left(4, \quad \frac{1152|\Omega|\mathrm{osc}^4}{C_{iso}^2\lambda^2\sigma_0^4}\right) \geq \frac{C_{iso}\lambda\sigma_0^2}{2\sqrt{|R|}\mathrm{osc}^2(g)},$$

which rearranges to

$$\sqrt{|R|} \geq \min\left(\frac{C_{iso}\lambda\sigma_0^2}{4\mathrm{osc}^2(g)}, \quad \frac{C_{iso}^3\lambda^3\sigma_0^6}{1152|\Omega|\mathrm{osc}^6(g)}\right).$$

This result is significant in that it provides a lower bound on the size of a segmentation region and the bound only depends on g, λ, σ_0 and Ω. By working with the

bounds (3.4.11), (3.4.13) and (3.4.14) in a similar way we can also obtain an "inverse iosperimetric inequality" to prove that regions cannot be too "thin". In particular, we can show that

$$\sqrt{|R|} \geq C \cdot \ell(\partial R), \tag{3.4.15}$$

where C is a constant depending only on g, λ and Ω. More specifically,

$$\begin{aligned} \lambda \cdot \ell(\partial R) &\leq 2\#(\mathcal{N}(R))|R|\left(\frac{\mathrm{osc}(g)}{\sigma_0}\right)^2 \\ &\leq \alpha|R|^{1/2}|\Omega|^{1/2}\left(\frac{\mathrm{osc}(g)}{\sigma_0}\right)^2 \\ &\leq \max\left(4, \quad \frac{1152|\Omega|\mathrm{osc}^4}{C_{iso}^2\lambda^2{\sigma_0}^4}\right)|R|^{1/2}|\Omega|^{1/2}\left(\frac{\mathrm{osc}(g)}{\sigma_0}\right)^2. \end{aligned}$$

Morel and Solimini refer to the next result as a **compactness property** and comment that it provides a mathematical explanation of the efficiency of region growing methods. The reasoning in the lemma is that the set of all 2-normal segmentations is small in some sense. The efficiency follows since region growing methods produce 2-normal segmentations. Its proof uses Hausdorff distance (Evans and Gariepy, [32]) to measure the intrinsic distance between subsets of Ω. If K is a subset of Ω then K^ϵ denotes the set of all $x \in \Omega$ such that $d(x, K) = \inf_{y \in K} d(x, y) \leq \epsilon$. The *Hausdorff distance* $d(\cdot, \cdot)$ between two subsets $K, L \subset \Omega$ is then defined to be the infimum of all $\epsilon > 0$ such that $K \subset L^\epsilon$ and $L \subset K^\epsilon$.

Lemma 3.4.6 *For every sequence $\{K_n\}$ of 2-normal segmentations there exists a subsequence $\{K_{n_l}\}$ converging for the Hausdorff distance to a segmentation K such that*
$E(K) \leq \lim_{l\to\infty} \inf E(K_{n_l})$.

Proof: We use the fact, established by the preceding lemmas that the number of edges in a 2-normal segmentation is bounded from above. Hence there is a universal upper bound on the number of edges in the segmentations K_n.

By considering an appropriate subsequence we can assume each K_n has the same number of regions and edges. Using the inverse isoperimetric inequality (3.4.15) and the Lipschitz inequality for rectifiable curves implies, using the Ascoli-Arzela Theorem, that there exists a subsequence of $\{K_n\}$ which converges to a segmentation K, say. This segmentation need not be 1- or 2- normal in general. We can also show $d(K_n, K) \to 0$ as $n \to \infty$, that is, for sufficiently large n the Hausdorff distance between K, K_n satisfies $d(K_n, K) \leq \epsilon$. To prove this, we use the notation c_m^i for the i-th curve in the subsequence K_m converging to the i-th curve c^i in the limit segmentation K. Since the curves c_m^i converge uniformly to c^i, we have

$$\forall \epsilon > 0 \; \forall i \in \mathcal{I} \; \exists n_i : |c_m^i - c^i| \leq \epsilon, \forall m \geq n_i, t \in [0,1],$$

where $\mathcal{I}$ is an indexing set for the edges in any of the segmentations K_m (recall that each K_m has the same number of edges). But since there are only a finite number of curves we have

$$\forall \epsilon \; \exists n : |c_m^i - c^i| \leq \epsilon, \forall m \geq n, t \in [0,1], i \in \mathcal{I}.$$

Therefore,

$$\begin{aligned} K^\epsilon = \bigcup_{i\in\mathcal{I}} (c^i)^\epsilon &\equiv \bigcup_{i\in\mathcal{I}} \bigcup_{t\in[0,1]} B(c^i(t),\epsilon) \\ &\supset \bigcup_{i\in\mathcal{I}} \bigcup_{t\in[0,1]} c^i_n(t) \\ &= K_n \\ K^\epsilon_n = \bigcup_{i\in\mathcal{I}} (c^i_n)^\epsilon &\equiv \bigcup_{i\in\mathcal{I}} \bigcup_{t\in[0,1]} B(c^i_n(t),\epsilon) \\ &\supset \bigcup_{i\in\mathcal{I}} \bigcup_{t\in[0,1]} c^i(t) \\ &= K \end{aligned}$$

which establishes Hausdorff convergence.

For large enough n, $d(K_n, K) \leq \epsilon \Rightarrow K_n \subset K^\epsilon \Rightarrow \Omega \setminus K^\epsilon \subset \Omega \setminus K_n$ which has a finite number of components for each n so θ_n is constant on each component of $\Omega \setminus K^\epsilon$ as each component of $\Omega \setminus K^\epsilon$ is in a component of $\Omega \setminus K_n$.

Observe that $\Omega \setminus K^\epsilon$ has a countable number of connected components. To see this, observe that for any $\epsilon > 0$, there can only be a finite number of components with area greater than ϵ, since Ω is finite. Thus for any positive integer n there are only a finite number of regions R_i with

$$\frac{|\Omega|}{(n+1)} < |R| \leq \frac{|\Omega|}{n}$$

and therefore the set of connected components $\mathcal{R}$

$$\mathcal{R} = \bigcup_{n=1}^{\infty} R_n \equiv \bigcup_{n=1}^{\infty} \bigcup_{R\in\mathcal{R}} \left\{ R : \frac{|\Omega|}{(n+1)} < |R| \leq \frac{|\Omega|}{n} \right\}$$

must be countable. Thus we can extract a subsequence of $\{\theta_n\}$ which converges pointwise on $\Omega \setminus K^\epsilon$. To prove this note that the fact we have a countable number

of connected components means θ_n is isomorphic to some sequence $\{s_i\}_i$ where each s_i is the vector (μ_i, σ_i^2) corresponding to the estimated parameters for each region. Clearly each s_i is bounded since by hypothesis the image data g is bounded. We then construct a subsequence such that $\{s_i\}$ converges for the first component and then another subsequence for the second component and so on. By taking the "diagonal" we see θ_n converges pointwise on each component.

Therefore for each $\epsilon > 0$ we can construct a subsequence $\theta_n^\epsilon \to \theta^\epsilon$ converging on $\Omega \setminus K^\epsilon \subset \Omega \setminus K$. For $\epsilon = 1, 1/2, 1/3, \ldots$ we can repeat the above argument on θ^ϵ to obtain a subsequence $\{\theta_n\}$ converging pointwise everywhere on $\Omega \setminus K$. Let $\theta = \lim \theta_n$. Clearly θ is constant in every connected component of $\Omega \setminus K$. Indeed, since any component of $\Omega \setminus K$ is open, given any two points x, y in such a component we have x, y in the same component of $\Omega \setminus K^\epsilon$ for sufficiently small ϵ, and thus $\theta(x) = \lim \theta_n(x) = \lim \theta_n(y) = \theta(y)$.

Since the original data g is bounded we have

$$\begin{aligned} \inf_\Omega g \leq \bar{\mu}_i &= \frac{1}{|R_i|} \int_{R_i} g(x)dx \leq \sup_\Omega g, \\ 0 < \sigma_0^2 \leq \bar{\sigma}_i^2 &= \frac{1}{|R_i|} \int_{R_i} (g(x) - \bar{\mu}_i)^2 dx \leq \mathrm{osc}^2(g). \end{aligned}$$

Thus Fatou's Lemma implies

$$\int_\Omega \frac{\log(\bar{\sigma}_i^2)}{2} dx + \int_\Omega \frac{(g(x) - \bar{\mu})^2}{2\bar{\sigma}_i^2} dx \leq \liminf \int_\Omega \frac{\log \bar{\sigma}_n^2}{2} dx + \int_\Omega \frac{(g(x) - \bar{\mu}_n)^2}{2\bar{\sigma}_n^2} dx.$$

As with (Morel and Solimini, [58]) the total lengths of the curves $\ell(K)$ can only decrease when passing to the limit and thus $\lim E(K_i) \leq E(K_n)$. ∎

Lemma 3.4.7 *There exists a 2-normal segmentation realizing* $\inf_K E(K)$*, this minimum being taken among all 2-normal segmentations.*

Proof: From Lemma 3.4.6 we can construct a sequence K_n of 2-normal segmentations with the following properties:

- $K_n = \{c_n^k\}_{k\geq 1}$ is made of a constant number of Jordan curves $c_n^k : (0,1) \to \Omega$. Their tips are a finite set of points, $A = (a_n^i)_i$, the crossing points of K_n, which may belong to $\partial\Omega$.
- Each sequence $\{a_n^i\}_{n\geq 1}$ converges to a point a^i of the closure of Ω.
- Each sequence of curves $\{c_n^k\}_{n\geq 1}$ converges uniformly to a rectifiable curve c^k contained in the closure of Ω.
- The sequence $E(K_n)$ converges to $\inf_K E(K)$, this infimum being taken among all 2-normal segmentations.

To see this, let $I = \inf E(K)$, the infimum being taken among all 2-normal segmentations. Then there exists a sequence $\{K_n\}$ such that $I \leq E(K_n) < I + 1/n$. Since $\log \sigma(K)$ is uniformly bounded from below $I > -\infty$ and therefore $\lim_n E(K_n) = I$. Using a similar argument from the previous lemma, we can assume all segmentations in the subsequence have the same number of edges and regions and each sequence $\{c_n^k\}_{n\geq 1}$ of curves converges to a limit rectifiable curve c^k. Since the crossing points are tips of the curves, they belong to the curves and thus each sequence of crossing points a_n^i converges to some a^i. Let K be the union of all such $(c^k)_k$. From the previous lemma it is clear that $E(K) \leq \liminf E(K_n) = I = \inf E(K)$, the infimum being taken over all 2-normal segmentations. It is clear that $\inf E(K)$ is the same whether the infimum is taken over all 2-normal segmentations or simply all segmentations. Indeed, if a segmentation is not 2-normal then we can perform region merging until it becomes 2-normal. Thus a segmentation that is not 2-normal

cannot realize the infimum of $\inf E(K)$. It remains to show K is 2-normal. Clearly, merging two regions of K cannot decrease the energy functional. Thus it only remains to verify that the Jordan curves only meet each other and $\partial\Omega$ at their tips and each curve separates two different regions. This is equivalent to proving the only crossing points of K are the limit points a_i where each a_i is the limit of some tips of the Jordan curves in the segmentations K_n. Assume by contradiction that a new crossing point a appears as we pass to the limit. Consider a disc D with centre a and radius ϵ with $\epsilon \ll 1$. Assume that for all i, k,

$$E(K_n) \leq I + \epsilon^2, \quad |a_i^n - a_i| \leq \epsilon^2, \quad \sup_x |c_n^k(x) - c^k(x)| \leq \epsilon^2$$

for sufficiently large n, say $n \geq n_0$, and consider all maximal pieces of curves K contained in D. Let us now replace each of these pieces by affine curves. Since the curves c_n^k do not cross in D, these affine curves also do not cross. By this process the connected components of $\Omega \setminus K$ remain unchanged outside D and the length of the curves decreases. Call K' the new segmentation. We still have $E(K') \leq \inf_K E(K) + C\epsilon^2$. Now consider two affine curves $[u, v]$ and $[x, y]$ of K' which have been substituted for two parts of curves of K_n passing at a distance less than ϵ^2 from a. Since a is a crossing point of K, such curves exist for large n. The lengths of the corresponding pieces of curves are greater than $2\epsilon - 2\epsilon^2$ since they pass at a distance less than ϵ^2 from the centre of D. Moreover, when replaced by affine curves, their lengths cannot decrease by more than $C\epsilon^2$, otherwise the segmentation would not have been optimal in the first place. The contradiction comes by modifying K' as follows: add to K' the segments $[u, x], [v, y]$ and remove *one* of the segments $[u, v]$ or $[x, y]$. This does not modify the connected components outside D.

Thus we have subtracted from $E(K')$ a length of order 2ϵ, which is a contradiction. A similar argument shows we obtain a similar contradiction if a new crossing point appears in K on the boundary of Ω. ∎

Lemma 3.4.8 *Let K be an optimal 2-normal segmentation. Any Jordan curve in K is a.e. twice differentiable. At such regular points the curvature is defined and bounded by* $16\ \mathrm{osc}^2(g)/\lambda\sigma_0^2$.

Proof:

From (3.4.6) and (3.4.7) it is easily seen if a segmentation is altered by perturbing a single curve within a rectangle of area A, the change of the "integral term" of the energy is bounded by some constant times its area, namely:

$$|\Delta E_1 + \Delta E_2| \leq 2A\frac{osc^2(g)}{\sigma_0^2}.$$

Let $c(s)$ be the arc-length parametrization of a Jordan curve of K defined on some interval containing $[-L, L]$. Thus we have $|c(s) - c(-s)| \leq 2s$, the equality being true if and only if c is an affine curve. Set $|c(s) - c(-s)| = 2(s - \epsilon)$. Notice that all points $c(r)$ with $r \in [-s, s]$ must be enclosed by an ellipse C with foci $c(-s), c(s)$ since c is parametrized with arc length. One easily sees that the ellipse is contained in a rectangle with sides of respective length $2s$ and $2\sqrt{(s^2 - (s-\epsilon)^2)}$. Thus the area of the rectangle is bounded from above by $4s\sqrt{2s\epsilon}$.

We now use the optimality of the segmentation with respect to the energy by stating that a certain rearrangement of the segmentation cannot decrease the energy. This rearrangement consists of replacing the curve c by its affine curve and altering the parameters (μ_i, σ_i^2) in the corresponding regions. This decreases the length term by $2\lambda\epsilon$ and the integral term of E cannot increase by more than $8s\sqrt{2s\epsilon}(\mathrm{osc}^2(g)/\sigma_0^2)$. Thus, the optimality of the segmentation implies that

$$2\lambda\epsilon \leq 8s(2s\epsilon)^{\frac{1}{2}}\frac{\mathrm{osc}^2(g)}{\sigma_0^2}$$

and therefore

$$\epsilon \leq \frac{32s^3\mathrm{osc}^4(g)}{\lambda^2\sigma_0^4}.$$

We now consider two successive increments of c, $v_1 = c(0) - c(-s), v_2 = c(s) - c(0)$. We want to estimate the difference of these increments since it will yield an estimate for the curvature at 0. Using the classical parallelogram identity, $|v_1 - v_2|^2 = 2(|v_1|^2 + |v_2|^2) - |v_1 + v_2|^2$ we obtain:

$$|v_1 - v_2|^2 \leq 4s^2 - (2s - 2\epsilon)^2 \leq 256 s^4 \frac{\mathrm{osc}^4(g)}{\lambda^2 \sigma_0^4}$$

and finally

$$|v_1 - v_2| \leq \frac{16 s^2 \mathrm{osc}^2(g)}{\lambda \sigma_0^2}.$$

This yields an upper bound for the discretized second-order derivative of c since if c'' exists then

$$c''(0) = \lim_{s \to 0} \frac{c(s) - 2c(0) + c(-s)}{s^2}$$

which implies

$$|c''(0)| \leq \frac{16 \mathrm{osc}^2(g)}{\lambda \sigma_0^2}. \tag{3.4.16}$$

As c is Lipschitz in s, $c'(s)$ exists a.e. Let a and b be two points at which c is differentiable and without loss of generality assume $a < b$. We divide the interval $[a, b]$ into n equal intervals with vertices $s_i = a + i(b - a)/n, i = 0 \ldots n$. Set $w_i = \frac{n}{b-a}(c(s_i) - c(s_{i-1}))$. Then, using the above estimate with $s = \frac{b-a}{n}$ gives

$$|w_i - w_{i-1}| \leq 16 \frac{(b-a)\mathrm{osc}^2(g)}{n \lambda \sigma_0^2}.$$

By adding these estimates for all i and using the triangular inequality,

$$|w_n - w_1| \leq 16 \frac{(b-a)\mathrm{osc}^2(g)(n-1)}{n \lambda \sigma_0^2} < 16 \frac{(b-a)\mathrm{osc}^2(g)}{\lambda \sigma_0^2}.$$

Letting $n \to \infty$ then $w_1 \to c'(a), w_n \to c'(b)$, and we obtain

$$|c'(b) - c'(a)| \leq 16(b-a) \frac{\mathrm{osc}^2(g)}{\lambda \sigma_0^2}. \tag{3.4.17}$$

We claim that c' actually exists everywhere (instead of just almost everywhere) and Lipschitz continuous with constant no greater than $16\text{osc}(g)/\lambda\sigma_0^2$. To prove this, consider an arbitrary value of s. Then there exist sequences a_n, b_n with $a_n < s < b_n$, $c'(a_n), c'(b_n)$ exist and $b_n - s = s - a_n = h_n \to 0^+$. For h_n, the discrete curvature estimate implies $|c(s+h_n) - 2c(s) + c(s-h_n)| < Ch_n^2$ for some constant C depending only on $\text{osc}(g), \sigma_0^2$ and λ. It is easy to show that

$$\frac{|c(s+h_n) - c(s)|}{h_n} - \frac{|c(b_n) - c(a_n)|}{2h_n} \to 0$$

as $n \to \infty$. Therefore

$$c'(s^+) = \lim_{n\to\infty} \frac{c(b_n) - c(a_n)}{b_n - a_n}.$$

However, it is also easy to verify

$$\begin{aligned} & \left| c(b_n) - c(a_n) - (b_n - a_n)\frac{c'(a_n) + c'(b_n)}{2} \right| \\ = & \int_{a_n}^{b_n} \left(\frac{|c'(s) - c'(a_n)|}{2} + \frac{|c'(s) - c'(b_n)|}{2} \right) ds \\ \leq & \int_{a_n}^{b_n} C(b_n - a_n) ds \\ = & C(b_n - a_n)^2. \end{aligned}$$

Therefore

$$c'(s^+) = \lim_{n\to\infty} \frac{c(b_n) - c(a_n)}{b_n - a_n} = \lim_{n\to\infty} \frac{c'(a_n) + c'(b_n)}{2}$$

exists. By symmetry $c'(s^-)$ also exists and equals $c'(s^+)$, thus proving the claim. From equation (3.4.17) we deduce that c'' is defined in the distributional sense and with modulus bounded from above by $16\text{osc}^2(g)/\lambda\sigma_0^2$ and its modulus is nothing but the curvature of c. ∎

Lemma 3.4.9 *The crossing points of K are as described in Theorem 3.4.1.*

Proof: Assume by contradiction that more than three curves arrive at a crossing point a or that exactly three arrive, but with angles different from 120 degrees. Then two of these curves form an angle strictly less than 120 degrees. Let ϵ be very small and $D = B(a, \epsilon)$ where $B(a, r)$ denotes the ball centred at a with radius r. We claim that for sufficiently small ϵ the parts of curves lying inside D are approximately affine curves, accurate to order $O(\epsilon^2)$. Let $c : [0, \epsilon_0] \to \Omega$ be such a curve with $c(0) = a$. For any $s \in [0, \epsilon_0]$ we consider sufficiently small ϵ such that $s = k\epsilon$ for some integer k. Then consider the polygonal curve $c(0), c(\epsilon), \ldots, c(s)$. From the discrete curvature bound (3.4.16) from the previous lemma we conclude that the maximum deviation from a straight line is obtained by tracing the discrete approximation of the circumference of a circle with radius equal to $\lambda\sigma_0^2/16\mathrm{osc}^2(g)$. Thus for sufficiently small ϵ, the curve must be well approximated by an affine curve, in the sense that the actual curve length is within $\mathcal{O}(\epsilon^2)$ of the straight line distance and c' is within $\mathcal{O}(\epsilon^2)$ of a constant in $\mathbf{R}^2$. In what follows we assume the curves inside D are affine.

Call c_u, c_v any pair of curves forming an angle less than 120 degrees, and u, v the intersections of c_u, c_v with D. Let w be the *Fermat point* of auv, that is, w is the unique point such that the lines wu, wv, wa have 120 degree angles. Since $|au| = |av|$ the triangle auv is isosceles and all angles of auv are less than 120 degrees. Thus w lies inside triangle auv. Let α be the angle determined by c_u, c_v. Let $|wa| = p, |wu| = |wv| = q, |au| = |av| = \epsilon$. Then the total lengths of all curves decreases by $2\epsilon - (p + 2q)$. Since auv is isosceles, an elementary computation shows that

$$2\epsilon - (p + 2q) = \frac{3r^2\epsilon}{2 + \sqrt{4 - 3r^2}} > 0,$$

where $r = \sin(60 - \alpha/2)/\sin(120) \in (0, 1]$.

Remove from the segmentation the pieces of curves au, av and add the segments wu, wv, wa. This modification does not alter the connected components outside D, changes the area terms with order ϵ^2 and reduces the length of K with order ϵ, which is impossible.

Note that in the case of a curve meeting $\partial\Omega$ a similar argument can be used to obtain the same contradiction. Let c be a curve meeting $\partial\Omega$ at any angle except 90 degrees. Then consider $D = B(a, \epsilon)$ where a is the intersection of c with $\partial\Omega$. Let u be the intersection of c with D and replace the curve au with the perpendicular from u to $\partial\Omega$. The rest follows as above. ∎

3.5 Experimental Results

We adapted the FLSA-algorithm (Chapter 2) by using the energy functional (3.3.4) instead of (3.1.3). This results in a new algorithm which we denote FLSA-MAP. Figure 3.5.1 shows the result of the FLSA-MAP segmentation algorithm with variance offset of $\sigma_0^2 = 1.3 \times 10^{-6}$. We displayed the results for both 50 (middle) and 200 (right) regions remaining, as was done in (Morel and Solimini, [58]). The image is a 256×256 subimage of the standard "boat" image and is the same subimage as the one appearing in (Morel and Solimini, [58]). The actual output was generated by Koepfler et al. in [42], as acknowledged by (Morel and Solimini, [58]). We believe our results are comparable with that of (Koepfler et al., [42]). Figure 3.5.2 shows the result of segmenting a synthetic image consisting of nine circles on a background. Each circle and the background had pixel intensities taken from normal distributions with the same mean but different variance. We kept the last ten regions and again used an offset of $\sigma_0^2 = 1.3 \times 10^{-6}$. The algorithm is able to recover all circles, despite the fact that the circles and background have the same mean gray-value.

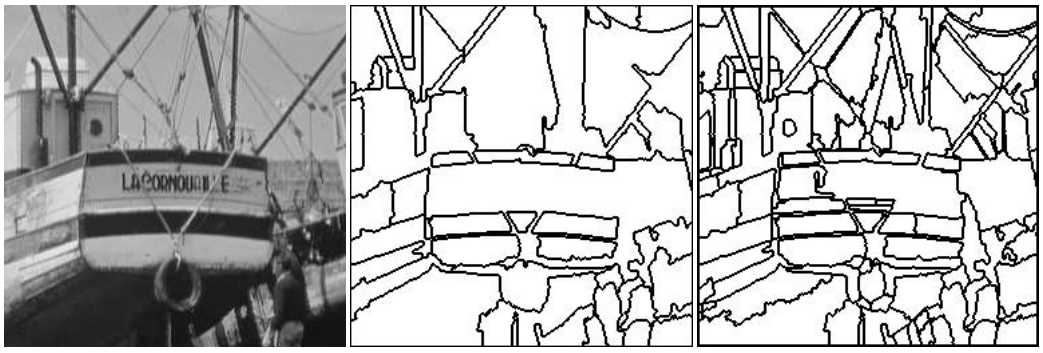


Figure 3.5.1: Segmentation of the Boat image.

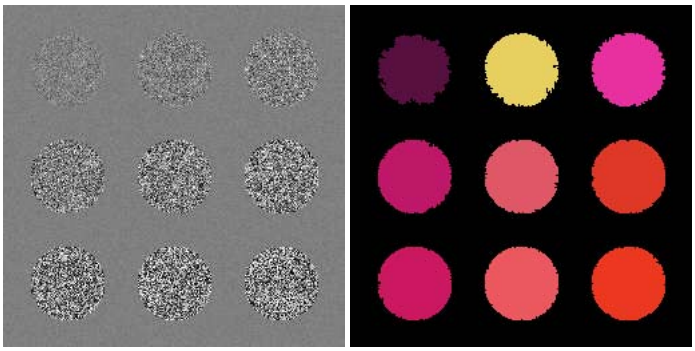


Figure 3.5.2: Segmentation of a synthetic image

3.6 Conclusions

We have proposed a Bayesian model for segmentation in the Mumford-Shah model. Our new model generalizes the Mumford-Shah model and handles noisy images at the cost of an extra regularization parameter. We have proved that optimal segmentations exist and have the same desirable properties as given in the standard Mumford-Shah model. We can adapt our region merging algorithm by using a new energy functional instead of the original Mumford-Shah functional.

Chapter 4

Computation of a Unique Minimizer of the Energy Functional for the Extended Mumford-Shah Model.

4.1 Introduction

Recall that for the standard Mumford-Shah model the segmentation problem is to minimize the following energy:

$$E(u, K) = \int_{\Omega} (g(x) - \mu(x))^2 dx + \lambda \cdot \ell(K), \tag{4.1.1}$$

where u is piecewise constant, $u(x) = \mu_i$ on the regions R_i and $\ell(K)$ is the length of K. The optimal choice of μ_i given K is

$$\bar{\mu}_i = \frac{1}{|R_i|} \int g(x) dx$$

so as before we consider the minimization of

$$E(K) = \int_\Omega (g(x) - \bar{\mu}_i)^2 dx + \lambda \cdot \ell(K), \tag{4.1.2}$$

where $\bar{\mu}_i$ is the mean gray value of region R_i.

In the previous chapter we derived the following energy functional for our extended Mumford-Shah model in the continuous domain:

$$E(K) = \sum_i \frac{|R_i| \log \bar{\sigma}_i^2}{2} + \sum_i \frac{|R_i| \sigma_i^{*2}}{2\bar{\sigma}_i^2} + \lambda \cdot \ell(K), \tag{4.1.3}$$

with the optimal parameters given by

$$\bar{\mu}(x) = \bar{\mu}_i = \frac{1}{|R_i|} \int_{R_i} g(z) dz \quad \text{for} \quad x \in R_i, \tag{4.1.4}$$

$$\begin{aligned} \bar{\sigma}^2(x) = \bar{\sigma}_i^2 &= \max(\sigma_0^2, \sigma_i^{*2}) \\ &= \max\left(\sigma_0^2, \frac{1}{|R_i|} \int_{R_i} (g(z) - \bar{\mu}_i)^2 dz\right) \quad \text{for} \quad x \in R_i, \end{aligned} \tag{4.1.5}$$

and σ_0 is a given constant. For both the original and the extended Mumford-Shah models, we assume that the optimal parameters are always selected, and it only remains to choose the boundary set K. For ease of reference we refer to $\lambda \cdot \ell(K)$ as the length term and $\sum_i |R_i| \log \bar{\sigma}_i^2/2 + \sum_i |R_i| \sigma_i^{*2}/2\bar{\sigma}_i^2$ as the integral term. We will also use the same names for a 1-dimensional version of (4.1.3), to be defined later.

In Chapter 2, we stated and proved an elementary result of the Mumford-Shah functional: if g is piecewise constant and $\lambda > 0$ is sufficiently small then the optimal segmentation is obtained by taking the union of the boundaries of the regions where g is constant. This property is referred to as "correctedness" (Koepfler et al., [42]). The result indicates that for a simple class of images, namely those which

are piecewise constant, it is possible to obtain the "correct" segmentation, given an appropriate choice of scale parameter(s) and this confirms that the Mumford-Shah model is theoretically sound. We note that in the literature it is common to compute an explicit minimizer for simple images for other models too. For example in (Strong and Chan, [80]) a minimizer is computed for the Rudin-Osher-Fatemi (ROF) model [75] where the image is a characteristic function of a disc. Chan and Esedoglu [19] do likewise, except they vary the original ROF model by taking the L^1 instead of L^2-norm of the fidelity term.

We recall that the main motivation for extending the original Mumford-Shah model is that the original model does not handle noisy images. Said differently, if two adjacent regions have the same mean but different variance (assuming that the normal distribution is a reasonable approximation to region statistics) then two regions will be classified as one. Therefore, we will not consider simple images such as the characteristic function of a disc. We will consider a more complex image and aim to show that the extended model outperforms the original model in the following sense: the original model does not recover the ideal segmentation for any choice of the set of scale parameters (in this case, only one, namely λ), but the extended model does for at least one choice of the set of scale parameters. The proof of this result is given in Section 4.2 and a short summary is provided in Section 4.3.

4.2 The Image g and its Unique Minimizer of the Energy Functional

Let L be a large integer, s large and $\Omega = [0, 2L] \times [0, 2L]$. L should also be significantly larger than s. Let $g : \Omega \to \mathbf{R}$ be defined by

$$\begin{aligned} g(x,y) &= +1, \text{ if } [x] < L, [x] \text{ even}, \\ g(x,y) &= -1, \text{ if } [x] < L, [x] \text{ odd}, \\ g(x,y) &= +s, \text{ if } [x] > L, [x] \text{ even}, \\ g(x,y) &= -s, \text{ if } [x] > L, [x] \text{ odd}, \end{aligned}$$

where $[x]$ is the "floor" function of x, i.e. the smallest integer greater than or equal to x. Set $\lambda = L \log(s/2)$. For example, if $L = 6, s = 5$ then $g : \Omega \to \mathbf{R}$ would be defined as shown in Figure 4.2.1. Note that we are considering the continuous domain so the segmentation K of g need not necessarily consist of horizontal and vertical lines only.

+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5
+1	-1	+1	-1	+1	-1	+5	-5	+5	-5	+5	-5

Figure 4.2.1: Definition of g with parameters $L = 6, s = 5$.

Obviously the "desired" segmentation K is the division of Ω into two regions with the vertical line $x = L$. In other words $K = \{(x, y) : x = L\}$. We aim to show (i) in the original Mumford-Shah model, the desired segmentation is not attained for any s, L, λ and (ii) in the new model, it is attained for some choice of $s, L, \lambda, \sigma_0^2$.

We immediately note (i) is trivial. From equation (4.1.2) it follows that if two adjacent regions have the same mean, merging them will leave the penalty term unchanged and decrease the length term by the common boundary length of the two regions times the scale parameter. Therefore if the desired segmentation were obtained for the Mumford-Shah model, we can immediately decrease the energy by merging the two regions into one.

For (ii) we choose the following parameter values:

$$\sigma_0^2 = \frac{8}{9}, \tag{4.2.1}$$

$$\lambda = L \log \frac{s}{2}, \tag{4.2.2}$$

$$L = 10^{12}, \tag{4.2.3}$$

$$s = 10^4. \tag{4.2.4}$$

Since g does not depend on y it is natural to consider the corresponding 1-dimensional minimization problem

$$\begin{aligned} E'(K') &= \sum_{i=0}^{M} \frac{|R_i| \log \bar{\sigma}_i^2}{2} + \sum_{i=0}^{M} \int_{R_i} \frac{(g'(x) - \bar{\mu}_i)^2}{2\bar{\sigma}_i^2} dx + \lambda \#(K') \\ &= \int_{[0,2L]} \frac{\log \bar{\sigma}^2(x)}{2} dx + \int_{[0,2L]} \frac{(g'(x) - \bar{\mu}(x))^2}{2\sigma_i^2} dx + \lambda \#(K'), \end{aligned} \tag{4.2.5}$$

where $g' : [0, 2L] \to R$ is a "horizontal slice" of the original g, i.e. $g'(x) = g(x, 0)$. K' is a finite set of break points i.e. $K' = \{k_1, k_2, \ldots, k_M\}$ (if $M = 0$ then K' is the empty set), and R_i denotes the interval (k_i, k_{i+1}) where $k_0 = 0$ and $k_{M+1} = 2L$. $\#(K')$ is the number of break points, i.e. the cardinality of the set K'. For $x \in R_i$, the optimal values $\mu(x) = \bar{\mu}_i$ and $\sigma^2(x) = \bar{\sigma}_i^2$ are again given by equations (4.1.4) and (4.1.5) except that each R_i is an interval instead of a region.

Given a segmentation K and $0 \leq j \leq 2L$ we denote a "horizontal slice" of Ω and K by $\Omega_y = \{(x, y) : x \in \mathbf{R}\}$ and $K_y = K \cap \Omega_y$ respectively. We want to establish:

1. The unique minimizer of the 1-dimensional problem (4.2.5) is $K_0' = \{L\}$, in other words, the optimal segmentation is a single break point at L.

2. If K is a minimizer of the 2-dimensional problem (4.1.3), each horizontal slice K_j' of K must be a minimizer of (4.2.5) for all $0 \leq y_0 \leq 2L$.

Items 1. and 2. are obviously sufficient to establish that the desired segmentation is the unique minimizer of equation (4.1.3).

We now address item 1. From the previous chapter we deduced a lower bound on the size of any region in a 2-normal segmentation.

$$\lambda \cdot \ell(\partial(R_i, R_j)) \quad \leq \quad 2\frac{\mathrm{osc}^2(g)}{\sigma_0^2} \min(|R_i|, |R_j|).$$

which is equation (3.4.12) in Chapter 3. This rearranges to

$$\min(|R_i|, |R_j|) \geq \frac{\lambda \cdot \ell(\partial(R_i, R_j))\sigma_0^2}{2\mathrm{osc}^2(g)}. \tag{4.2.6}$$

The proof of (3.4.12) can easily be adapted to the 1-dimensional case to prove (4.2.6) with g' replacing g. In the 1-dimensional case, we have $\ell(\partial(R_i, R_j)) = 1$ and from the definition of g' and choice of parameters in equations (4.2.1) - (4.2.4) we have $\mathrm{osc}(g') = 2s$ and $\sigma_0^2 = 8/9$. Thus, any interval of an optimal partition must have length no less than $\lambda\sigma_0^2/9s^2 \approx 9.46 \times 10^3 \gg 2$ since the oscillation of g' is $2s$. We immediately observe this simplifies the 1-dimensional energy considerably. From the definition of g' it is easy to see no interval R_i can have $\bar{\sigma}_i^2 < 8/9$ otherwise its length must be smaller than 2, which we already know is impossible. (If the length of the interval R_i exceeds 2 then the minimum value $\bar{\sigma}_i^2 = 8/9$ occurs, for example, when $R_i = \{x : 1 < x < 4\}$). But then the term $\sum_i \int_{R_i}(g'(x) - \bar{\mu}_i)^2/2\bar{\sigma}_i^2 = \sum_i |R_i|\bar{\sigma}^{*2}/2\bar{\sigma}^{*2} = 2L/2 = L$ is constant so we need only consider the minimization of

$$\begin{aligned} E'(K') &= \frac{1}{2}\sum_i |R_i| \log \bar{\sigma}_i^2 + \lambda \#(K') \\ &= \frac{1}{2}\int_{\Omega'} \log \bar{\sigma}^2(x) dx + \lambda \#(K'). \end{aligned} \tag{4.2.7}$$

We claim the unique minimizer is a single break point at L i.e. $K_0' = \{L\}$. First observe that for this partition K_0' the energy (4.2.7) is

$$\begin{aligned} E'(K_0') &= \frac{L}{2}\log s^2 + \frac{L}{2}\log 1 + L\log\frac{s}{2}\cdot 1 \\ &= L\log\frac{s^2}{2} \end{aligned}$$

so we need to show no other partition can have lower energy. If $K' = \phi$ is the "empty" partition then

$$E'(K') = \frac{2L}{2}\log\frac{s^2+1}{2} > L\log\frac{s^2}{2}.$$

Next, we observe K' cannot consist of two or more intervals on the "left", i.e. contained in $[0, L]$. This would imply K' must be of the form $K' = \{k_1, \ldots, k_M\}$ where $k_1 < k_2 < L$ and $M \geq 2$ and we can decrease the energy by merging the two left-most intervals, i.e. replacing K' by $K' \setminus \{k_1\}$. We already established any interval of K' must have length at least $\lambda\sigma_0^2/8s^2$, so their length must be at least 2, say. It is easy to see for any interval contained in $[0, L]$, we have $8/9 \leq \bar{\sigma}_0^2 \leq 1$. Thus when K' is replaced with $K' \setminus \{k_1\}$, the error term changes by no more than $L\log 9/8 - L\log 1 = L\log 9/8$ but the complexity term decreases by $L\log s/2$.

Similarly we observe K' cannot consist of two or more intervals on the right, i.e. contained in $[L, 2L]$. This would imply K' must be of the form $K' = \{k_1, \ldots, k_M\}$ with $L < k_{M-1} < k_M$ and $M \geq 2$. We can decrease the energy by merging the two right-most intervals, i.e. replacing K' by $K' \setminus \{k_M\}$. Again we know that any interval of K' must be of length at least 2. Let I be any interval contained in $[L, 2L]$ and assume it has length of $A + B \geq 2$ where $A = |\{x \in I : g(x) = s\}|$ and $B = |\{x \in I : g(x) = -s\}|$ and $|.|$ denotes Lebesque measure. It is not hard to show $\max(A/B, B/A) \geq 2$ and thus $8/9s^2 \leq \bar{\sigma}_I^2 = 4AB/(A+B)^2 \leq s^2$.

When two such intervals are merged, the error term changes by no more than

$$L \log s^2 - L \log \frac{8s^2}{9} = L \log \frac{9}{8} < L \log \frac{s}{2} = \lambda \cdot 1.$$

Thus merging the two intervals decreases the energy, a contradiction.

Next, observe we cannot have three or more break points, since this would imply either two intervals on the left or two intervals on the right, which we know is impossible. We next establish K' cannot have even two break points. Suppose $K' = \{k_1, k_2\}$ and denote the three intervals $R_i = (k_i, k_{i+1})$ for $i = 0, 1, 2$ where $k_0 = 0$ and $k_3 = 2L$. To avoid two intervals on the left or right, we must have $k_1 \leq L \leq k_2$ The complexity term contributes $2L \log(s/2) = L \log s^2/4$. Hence the error term must be no greater than $L \log 1/2$. But consider k_2. If $k_2 \leq 3L/2$ then R_3 has error at least $L/2 \log 8s^2/9$, which is too high. Similarly if $k_2 \geq 3L/2$ then it is not hard to show that it is region R_2 which contributes at least $L/2 \log 8s^2/9$ to the error term, another contradiction.

It remains to consider segmentations with only a single break point. Let K' be the segmentation where the break point occurs at $0 < k_0 < 2L$. The energy (4.2.7) reduces to

$$E'(K') = \frac{|R_0|}{2} \log \bar{\sigma}_0^2 + \frac{|R_1|}{2} \log \bar{\sigma}_1^2 + \lambda,$$

where R_0, R_1 are the two regions in question and $\bar{\sigma}_1^2, \bar{\sigma}_2^2$ are the corresponding estimated variances. We consider the following cases: $x_i \leq k_0 \leq x_{i+1}$ for $i = 0, 1, \ldots, 6$ (but not in that order) where $\{x\}_i = \{0, 2, L-1, L, L+1, L+10^3, 2L\}$.

We first show the break point cannot be to the left of L. Suppose the break point occurs at $L - 1 < k_0 = L - r < L$. Then

$$\begin{aligned}\bar{\sigma}_0^2 &= 1 - \frac{\Delta k^2}{(L-\Delta k)^2},\\ \bar{\sigma}_1^2 &= \frac{Ls^2}{L+\Delta k} + \frac{L\Delta k}{(L+\Delta k)^2}.\end{aligned}$$

To deal with the energy (4.2.7) effectively we need to estimate the natural logarithm function in terms of rational functions of L and r. The usual trick would be to estimate the natural logarithm function via $\log(x) \le x-1$. Unfortunately it turns out the inequality is in the "wrong direction", so we instead use $-\log(1/x) \ge 1-1/x$. An elementary calculation shows that

$$\begin{aligned}E'(K') &= \frac{L-\Delta k}{2}\log\left(1-\frac{\Delta k^2}{(L-\Delta k)^2}\right)\\ &\quad + \frac{L+\Delta k}{2}\log\left(\frac{Ls^2}{L+\Delta k}+\frac{L\Delta k}{(L+\Delta k)^2}\right)+\lambda\\ &\ge \frac{L-\Delta k}{2}\log\left(1-\frac{\Delta k^2}{(L-\Delta k)^2}\right)+\frac{L+\Delta k}{2}\log\left(\frac{Ls^2}{L+\Delta k}\right)+L\log\frac{s}{2}\\ &= -\frac{L-\Delta k}{2}\log\left(1+\frac{\Delta k^2}{L^2-2\Delta kL}\right)\\ &\quad + \frac{L+\Delta k}{2}\Big(\log s^2-\log(1+\frac{\Delta k}{L})\Big)+L\log\frac{s}{2}\\ &\ge -\frac{L-\Delta k}{2}\cdot\frac{\Delta k^2}{L^2-2\Delta kL}\\ &\quad + \frac{L}{2}\log s^2+\frac{\Delta k}{2}\log s^2-\frac{\Delta k(L+\Delta k)}{2L}+L\log\frac{s}{2}\\ &> \frac{L}{2}\log s^2+L\log\frac{s}{2}=L\log\frac{s^2}{2}\end{aligned}$$

since $\Delta k/2\log s^2$ dominates the negative terms.

We already know $k_0 > 2$ since the length of any interval of K' must exceed 2. For $2 < k_0 < L-1$, the proof is similar. Although the exact calculation of $\bar{\sigma}_0^2, \bar{\sigma}_1^2$ is more involved it is sufficient to estimate them via

$$\begin{aligned}\bar{\sigma}_0^2 &\geq \frac{(L-\Delta k)^2-1}{(L-\Delta k)^2},\\ \bar{\sigma}_1^2 &\geq \Big(\frac{Ls^2}{L+\Delta k}\Big),\end{aligned}$$

and the rest follows as before.

We next demonstrate that the break point cannot be to the right of L. First, suppose that $L < k_0 = L + \Delta k < L + 1$. Then

$$\begin{aligned}\bar{\sigma}_0^2 &= \frac{L}{L+\Delta k} + \frac{L\Delta k s^2}{(L+\Delta k)^2},\\ \bar{\sigma}_1^2 &= \Big(1 - \frac{\Delta k^2}{(L-\Delta k)^2}\Big)s^2,\end{aligned}$$

and therefore

$$\begin{aligned}E'(K') &= \frac{L+\Delta k}{2}\log\bar{\sigma}_1^2 + \frac{L-\Delta k}{2}\log\bar{\sigma}_2^2 + L\log\frac{s}{2}\cdot 1\\ &= \frac{L+\Delta k}{2}\log\Big(\frac{L}{L+\Delta k}\\ &\quad + \frac{L\Delta k s^2}{(L+\Delta k)^2}\Big) + \frac{L-\Delta k}{2}\log\Big(s^2\Big(1-\frac{\Delta k^2}{(L-\Delta k)^2}\Big)\Big) + L\log\frac{s}{2}\\ &= -\frac{L+\Delta k}{2}\Big[\log\Big(1+\frac{\Delta k}{L}\Big) + \log\Big(1-\frac{\Delta k s^2}{L+\Delta k+\Delta k s^2}\Big)\Big]\\ &\quad + \frac{L-\Delta k}{2}\Big(\log s^2 - \log\Big(1+\frac{\Delta k^2}{L^2-2\Delta k L}\Big)\Big) + L\log\frac{s}{2}\\ &\geq -\frac{L+\Delta k}{2}\Big(\frac{\Delta k}{L} - \frac{\Delta k s^2}{L+\Delta k+\Delta k s^2}\Big)\\ &\quad + \frac{L-\Delta k}{2}\Big(\log s^2 - \frac{\Delta k^2}{L^2-2\Delta k L}\Big) + L\log\frac{s}{2}\\ &= \frac{L}{2}\log s^2 + \frac{\Delta k(L+\Delta k)}{2}\Big(\frac{s^2}{L+\Delta k+\Delta k s^2} - \frac{1}{L}\Big)\\ &\quad - \frac{\Delta k}{2}\log s^2 - \frac{L-\Delta k}{2}\cdot\frac{\Delta k^2}{L^2-2\Delta k L} + L\log\frac{s}{2}\\ &> \frac{L}{2}\log s^2 + L\log\frac{s}{2} = L\log\frac{s^2}{2}.\end{aligned}$$

The last inequality is valid since there is a positive term of order $\mathcal{O}(\Delta k s^2)$ which dominates all the negative terms.

If $L+1 < k_0 = L+\Delta k < L+10^3$ the proof is similar. Although the exact calculation of $\bar{\sigma}_0^2, \bar{\sigma}_1^2$ is more involved, it is sufficient to estimate them via

$$\begin{aligned} \bar{\sigma}_0^2 &= \frac{L}{L+\Delta k} + \mathcal{O}\Big(\frac{\Delta k s^2}{L+\Delta k}\Big), \\ \bar{\sigma}_1^2 &\geq \Big(\frac{(L-\Delta k)^2 - 1}{(L-\Delta k)^2}\Big) s^2, \end{aligned}$$

and the rest follows as before. Now consider the case when $L+10^3 < k_0 = L+\Delta k < 2L$ and Δk is an even integer. We have

$$\begin{aligned} E'(K') - E'(K_0) &= \frac{L+\Delta k}{2} \log \frac{L+\Delta k s^2}{L+\Delta k} + \frac{L-\Delta k}{2} \log s^2 - \frac{L}{2} \log 1 - \frac{L}{2} \log s^2 \\ &= -\frac{L}{2} \log \Big(1 - \frac{\Delta k (s^2-1)}{L+\Delta k s^2}\Big) - \frac{\Delta k}{2} \log \Big(1 + \frac{L(s^2-1)}{L+\Delta k s^2}\Big) \\ &\geq \frac{L \Delta k (s^2-1)}{2(L+\Delta k s^2)} - \frac{L \Delta k (s^2-1)}{2(L+\Delta k s^2)} \\ &= 0, \end{aligned}$$

where we used the fact $-\log(1+x) \geq -x$. Note that we could have immediately obtained this inequality simply by observing $\log(x)$ is concave, but this is not enough. We also need to establish $|E'(K') - E'(K_0')|$ is sufficiently large so we can estimate the difference when the break point k_0 is not an even integer by rounding k_0 to the nearest even integer. In this case we observe that $\Delta k(s^2-1)/L + \Delta k s^2 > 10^{11}/2 * 10^{12} = 1/20$ and that $-\log(1-1/20) - 1/20 > 10^{-3}$ so the error incurred by applying the inequality (4.2.8) is at least $L \times 10^{-3} = 10^9$ from the first term alone.

Now consider the case when $L + 10^3 < k_0 = L + \Delta k < 2L$ and Δk is not an even integer. By rounding the break point to the nearest even integer, the break point shifts by no more than one unit. Thus the energy of the segmentation changes by

no more than $2 \cdot \text{osc}^2/\sigma_0^2 = 2(2s)^2/(8/9) = 9 * 10^8$, a result proved in the previous chapter (the factor of 2 is necessary since shifting the break point is equivalent to "undoing a merge" by adding a new break point next to the original break point and then performing a different merge by deleting the original break point). But this is not sufficient to compensate for an error of 10^9 as shown above. Therefore the only correct segmentation can be a single break point at $K = \{L\}$.

We now address item 2. The general strategy is as follows: we decompose each region as an uncountable union of horizontal strips of zero width, (as in a Riemann integration) so we can estimate the energy of any segmentation by some expression involving the integration of (4.2.5) for every horizontal strip $y = y_0, 0 < y_0 < 2L$. We have observed from the previous chapter that when two regions are merged, the integral term changes by no more than $|R|\text{osc}^2/\sigma_0^2 = 9|R|s^2/2$ where $|R|$ is the size of the smaller region. Obviously the same inequality holds when a region is split into two (i.e. undoing the merge). This also applies to the 1-dimensional case.

We have already seen that for the variance of an interval to be below 8/9, the interval must have length less than 2. In the 2-dimensional case, we will refer to an interval as a subset of Ω of the form $\{x, y : x = x_0, y_0 < y < y_1\}$ for some x_0, y_0, y_1. We note the length term of (4.1.3) can be estimated from below by the integration of the cardinality term from (4.2.5) over all $0 < y < 2L$. Note it is possible for K to contain horizontal boundaries, in which case the cardinality is infinite for a particular $y = y_0$. But the set of "offending" values of y are Lebesque-negligible, so we will only integrate over the non-offending values of y. Also, (i) the log function is concave, which suggests we estimate $|R_i| \log \bar{\sigma}_i^2$ by integration of the $|R_{iy}| \log \bar{\sigma}_{iy}^2$ terms where the subscript iy represents the intersection of i^{th} region and the line $\{(x, y) : x \in \mathbf{R}\}$ and (ii) the term $|R_i|\sigma_i^{*2}/\bar{\sigma}_i^2$ behaves like $|R_i|$, which is actually equal to $\int_j |R_{ij}|$. Unfortunately both (i) and (ii) are only valid when the variances exceed the offset σ_0^2. Thus we also need to decompose the horizontal strips further into those where the variance is below or above σ_0^2. Fortunately, we observe

that if the variance of an interval R_{ij} is below σ_0^2 the length of the interval must be small (less than 2). Consider the horizontal slice containing such an interval. We can substantially decrease the 1-dimensional energy by merging that interval with an adjacent interval. The length term decreases by $L\log(s/2) = 8 \times 10^{12}$ and the integral term changes by no more than $2(2s)^2/\sigma_0^2 = 9s^2 = 9 \times 10^8$. In other words, when the variance of an interval is less than σ_0^2 we incur an error in applying (i) and (ii) but this is more than compensated for by the fact the solution to the 1-dimensional problem is substantially non-optimal for some $y = y_0$.

The above discussion was deliberately vague in order to develop the intuition. We now translate this into a rigorous proof. Denote by R_i the i^{th} region. Set $R_{iy} = R_i \cap \Omega_y = \bigcup_{k\in\mathcal{I}_{iy}} R_{iyk}$ where Ω_y is the horizontal line $\{(x,y) : x \in \mathbf{R}\}$ and $\mathcal{I}_{iy}$ is an indexing set for the intervals comprising R_{iy}. Note that R_{iy} is not necessarily a single interval. Denote by R_{iy}^0 (R_{iy}^1) the union of all intervals R_{iyk} such that the variance is less than (greater than or equal to) 8/9 respectively.

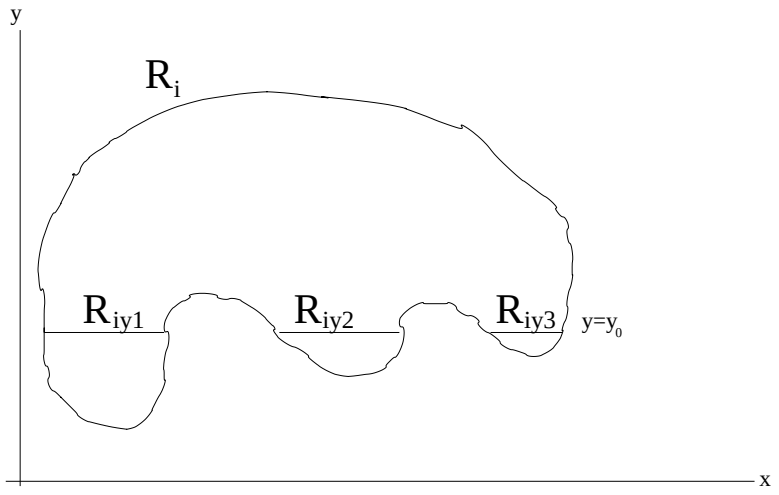


Figure 4.2.2: An example of decomposing a region into horizontal strips.

For example, in Figure 4.2.2 suppose $\sigma_1^{*2} > \sigma_2^{*2} > 8/9 > \sigma_3^{*2}$, where $\sigma_k^{*2} = 1/|R_{iyk}| \int_{R_{iyk}} (g(x) - \mu_i)^2 dx$ and $\mu_i = 1/|R_{iyk}| \int_{R_{iyk}} g(x) dx$, for $k = 1, 2, 3$. If $y = y_0$ then $R_{iy} = R_{iy1} \cup R_{iy2} \cup R_{iy3}$ and $R_{iy}^0 = R_{iy3}$ and $R_{iy}^1 = R_{iy1} \cup R_{iy2}$. It is easily verified that any interval $R_{iyk} \subset R_{iy}^0$ must have length at most 2, otherwise the variance would exceed 8/9. In particular, if Z_y is the number of intervals $R_{iyk} \subset \Omega_y$ with variance less than 8/9, then $|R_{iy}^0| < 2Z_y$. To simplify the notation, set

$$\begin{aligned} R_i^0 &= \bigcup_{0<y<2L} R_{iy}^0, \\ S_i &= \log \bar{\sigma}_i^2 + \frac{\sigma_i^{*2}}{\bar{\sigma}_i^2}, \end{aligned}$$

where σ_i^{*2} and $\bar{\sigma}_i^2$ are given by (4.1.5) and similarly for S_{iy}, S_{iy}^0 and S_{iy}^1. Since g has oscillation $2s$ the quantity

$$\log \bar{\sigma}_R^2 + \frac{\sigma_R^{*2}}{\bar{\sigma}_R^2}$$

is bounded from above by $\log 4s^2 + 1$ for any region R. Let us define the energy contribution E_y from a single horizontal line Ω_y as

$$E_y(K) = \sum_i \frac{|R_{iy}|}{2} S_{iy} + \lambda \#(K_y),$$

where $K_y = K \cap \Omega_y$. The total energy is estimated via

$$
\begin{aligned}
E(K) &= \sum_i \frac{|R_i|}{2} S_i + \lambda \cdot \ell(K) \\
&\geq \sum_i \frac{|R_i^0|}{2} S_i^0 + \sum_i \frac{|R_i^1|}{2} S_i^1 - \sum_i |R_i^0| \frac{9s^2}{2} + \lambda \cdot \ell(K) \\
&\geq \sum_i \int_0^{2L} \frac{|R_{iy}^0|}{2} S_{iy}^0 dy - \sum_i \int_0^{2L} \frac{|R_{iy}^0|}{2} \Big(\log(4s^2) + 1 \Big) dy + \sum_i \int_0^{2L} \frac{|R_{iy}^1|}{2} S_{iy}^1 dy \\
&\quad - \sum_i |R_i^0| \frac{9s^2}{2} + \lambda \cdot \ell(K) \\
&\geq \sum_i \int_0^{2L} \frac{|R_{iy}|}{2} S_{iy} dy - \sum_i \int_0^{2L} \frac{|R_{iy}^0|}{2} \Big(\log(4s^2) + 1 \Big) dy \\
&\quad - 2 \sum_i \int_0^{2L} |R_{iy}^0| \frac{9s^2}{2} dy + \lambda \int_0^{2L} \#(K_y) dy \\
&\geq \sum_i \int_0^{2L} \frac{|R_{iy}|}{2} S_{iy} dy - \int_0^{2L} Z_y (18s^2 + \log 4s^2 + 1) dy + \lambda \int_0^{2L} \#(K_y) dy \\
&\geq \int_0^{2L} \inf E_y + Z_y (L \log \frac{s}{2} - 9s^2 - 18s^2 - \log 4s^2 - 1) dy \\
&\geq \int_0^{2L} \inf E_y dy = 2L \inf E_y,
\end{aligned}
$$

where Z_y is the number of intervals lying in Ω_y where g is constant. The last inequality is valid since $L \log s/2$ dominates all the negative terms. But it is clear that $\inf E(K) \leq 2L \inf_{K'} E'(K')$. By solving the 1-dimensional problem we obtain a set of discontinuities $K' = \{x_0 < \ldots < x_n\}$ and in the 2-dimensional problem we can define K to be the set of all points whose x-coordinate lies in K'. In other words, K is a set of vertical lines. Thus $\inf E(K) = 2L \inf E'(K')$ and moreover in every horizontal strip the corresponding 1-dimensional energy (4.2.5) must realize its minimum value.

This establishes the desired result: for g defined as in Figure 4.2.1 and the parameter values given by (4.2.1-4.2.4) the only optimal segmentation is the set $K = \{(x, y) : x = L\}$.

4.3 Conclusions

The purpose of this chapter was to demonstrate via a mathematical analysis the superiority of the extended Mumford-Shah model over the standard one. The extended model can distinguish two regions with the same mean but different variance. We defined a simple image $g : [0, 2L] \times [0, 2L] \to \mathbf{R}$ consisting of two regions. The left region had zero mean and unit variance, and the right region had zero mean and large variance s^2 where $s^2 \gg 1$. We showed that for the original Mumford-Shah model, it was impossible for the two-region segmentation to be optimal for any choice of parameters s^2, L, λ, but for the extended model the two-region segmentation was optimal given suitable values of these parameters plus the offset σ_0^2.

Chapter 5

A Solution to the Small Sample Problem for Region Merging Algorithms

5.1 Introduction

We recall the work of Crisp and Newsam [26] which is based on a Bayesian approach where the optimal segmentation is defined as the model which is most likely to occur given the image data. In other words, it is a Maximum A-Posteriori estimate of the optimal partition of Ω into regions. The segmentation model $M = (K, \theta)$ consists of a boundary set K and a function θ approximating the image data. From Bayes Law we have

$$\hat{M} = \arg\max_M p(M|g) = \arg\max_M \frac{p(M)p(g|M)}{p(g)},$$

where g is the given data. Since $p(g)$ is assumed constant this is equivalent to

$$\hat{M} = \arg\min_M E(M) = \arg\min_M \Big(-\ln p(M) - \ln p(g|M) \Big). \tag{5.1.1}$$

In (Crisp and Newsam, [26]), the authors work in the discrete domain: the approximating function θ is constant within each pixel $\mathbf{x}$ and defined as a random variable whose pixel intensities follow independent Gaussian distributions. Any two such distributions are identical if they correspond to two pixels in the same region R_i of the segmentation K. Thus

$$\theta(\mathbf{x}) \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x})) = \mathcal{N}(\mu_i, \sigma_i^2) \qquad \text{where } \mathbf{x} \in R_i. \tag{5.1.2}$$

Since θ is completely determined by its means and variances at each pixel, we can encode the same information by using a deterministic two-dimensional vector. Thus we write $\theta(\mathbf{x}) = (\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$ instead of (5.1.2) from now on. Crisp and Newsam [26] show that the problem of minimising (5.1.1) reduces to

$$\begin{aligned} \hat{M} = \arg\min_M \Bigg(& \sum_i \#(R_i) \log\left(\sqrt{2\pi}\sigma_i\right) \\ & + \sum_i \sum_{\mathbf{x}\in R_i} \frac{1}{2\sigma_i^2}(g(\mathbf{x}) - \mu_i)^2 + \lambda \cdot \ell(K)\Bigg), \end{aligned} \tag{5.1.3}$$

where $\#(R_i)$ is the number of pixels in R_i and μ_i, σ_i^2 are the estimated mean and variance parameters corresponding to θ on R_i. It is easily shown that part of the minimization can be done analytically. If K is given then the optimal parameters for θ are the sample mean and variance, i.e.

$$\hat{\theta}(\mathbf{x}) = (\mu^*(\mathbf{x}), \sigma^{*2}(\mathbf{x})) = (\mu_i^*, \sigma_i^{*2}), \quad \mathbf{x} \in R_i, \tag{5.1.4}$$

$$\mu_i^* = \frac{1}{\#(R_i)} \sum_{\mathbf{x}\in R_i} g(\mathbf{x}), \tag{5.1.5}$$

$$\sigma_i^{*2} = \frac{1}{\#(R_i)} \sum_{\mathbf{x}\in R_i} (g(\mathbf{x}) - \mu_i^*)^2, \tag{5.1.6}$$

where R_i is the region containing pixel $\mathbf{x}$. In our notation we assume $\theta = \hat{\theta}$ unless explicitly stated otherwise. By only considering segmentation models where $\theta = \hat{\theta}$ the minimization problem reduces to

$$\hat{K} = \arg\min_K E(K, \hat{\theta}) = \arg\min_K \sum_i \#(R_i) \log(\sigma_i) + \lambda \cdot \ell(K). \tag{5.1.7}$$

We recall the Full Lambda Schedule Algorithm (FLSA) (Crisp and Tao, [27], Redding et al., [67], Robinson et al., [72]) which has already been described in Chapter 2. Redding et al. [67] and Robinson et al. [72] implemented the FLSA for the Mumford-Shah model. Crisp and Newsam [26] extended the FLSA to the above-mentioned Bayesian model with Gaussian distributions for pixels. We denote this extension of the FLSA as FLSA-MAP, since it uses the Maximum A-Posteriori principle to calculate an optimal segmentation given the image data. Unlike the Mumford-Shah model, a subtle difficulty is that for single pixel regions we have $\sigma = 0$, in which case (5.1.3) does not make sense. We view this as a manifestation of the small sample problem. The solution suggested in (Crisp and Newsam, [26]) was to associate with each region two quantities, the true variance σ_i^2 and modified variance $\hat{\sigma}_i^2$, both of which are used to calculate the change in energy when two regions are merged. A simpler more convenient version of this idea was offered by Crisp and Tao in [27]. The idea is to offset the variance estimate so that the sample variance σ^2 is replaced with $\sigma^2 + \sigma_0^2$ for some predetermined value of σ_0^2. In effect, this sacrifices accuracy to avoid the singularity at $\sigma^2 = 0$. Alternatively, we could replace σ^2 with $\max(\sigma^2, \sigma_0^2)$, which is essentially the approach used in Chapter 3. The disadvantage of either approach is that there is no obvious choice of the extra parameter σ_0^2. If σ_0^2 is too small, (5.1.3) is numerically unstable. If σ_0^2 is large, the loss of accuracy can be substantial. We remark that various other solutions to the small sample problem have already been proposed in the literature. For instance, Kanungo et al. [39] proposed that for small regions the sample variance should be weighted against the global variance for the whole image. This is somewhat similar

to (Crisp and Newsam, [26]), except the true and modified variance are in effect being combined into one via a weighted average. In Lee [45], the initial segmentation is determined by a set of *seeds* which consist of one or more pixels. Kanungo's solution has the drawback of specifying an extra parameter to determine the relative weighting between local and global variance. Moreover it does not work in the case of the image being (approximately) globally constant. The solution in (Lee, [45]) requires a non-trivial initialization stage for determining which pixels are seeds. It also requires an assumption on the size of the smallest region. In our view, none of the above solutions to the small sample problem have proved fully convincing and the aim of this paper is to describe a satisfactory solution to the small sample problem. In Section 5.2 we discuss the small sample problem and our proposed solution. Section 5.3 describes the new algorithm which we denote as FLSA-CDF (the reason for the "CDF" will become clear later). In Section 5.4 the utility of the new algorithm is assessed and its performance is compared with the FLSA described in Chapter 2. In Section 5.5 we draw some conclusions.

5.2 Solution to the Small Sample Problem

The reason behind the problems in (5.1.3) when $\sigma^2 = 0$ can be explained by careful consideration of the probability density function (pdf) of a Gaussian random variable. A Gaussian pdf can either be *proper* when $\sigma^2 > 0$ or *degenerate* when $\sigma^2 = 0$. Unless we work with distributions (in the sense of Schwartz, say) we see that degenerate pdfs are not defined everywhere. On the other hand the cumulative distribution function (cdf) is well defined everywhere regardless of the condition $\sigma^2 > 0$ without having to resort to Schwartz distributions. For instance, when $\sigma^2 = 0$, the cdf for $\mathcal{N}(\mu, \sigma^2)$ is $F(z) = H(z - \mu)$ where H is the Heaviside function, i.e. $H(x) = 1$ for non-negative x and 0 otherwise. This suggests it may be advantageous to consider the cdfs instead of pdfs.

In the use of cdfs it is conventional to resolve discontinuities via enforcing right-continuity everywhere. Thus the cdf correpsonding to a random variable Z is defined by

$$F_Z(z) = P(Z \leq z) = \int_{-\infty}^{z} f_Z(y)dy,$$

where f_Z is the pdf of Z (if it exists). By considering the cdf instead of pdf we forfeit the possibility of quantifying the likelihood of some set of observed data, via maximum likelihood estimation or otherwise. Instead we measure a different quantity, namely how much the cdfs corresponding to one segmentation differ from those corresponding to a benchmark segmentation. The obvious choice of benchmark is the trivial segmentation, since that yields the most accurate model of the image, at the expense of greatest complexity. Thus our formulation has similar properties to the Mumford-Shah model, in the sense that fine segmentations correspond to a low data-fidelity penalty and high model-complexity penalty and vice versa for coarse segmentations and we believe this justifies the use of cdfs as a solution to the small sample problem.

Given a segmentation K and a pixel $\mathbf{x}$ we define $F^{K,\mathbf{x}}$ to be a cdf corresponding to the sample parameters $(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$ determined by equations (5.1.5) and (5.1.6). In other words,

$$\begin{aligned} F^{K,\mathbf{x}}(z) &= P(\mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x})) \leq z) \\ &= \int_{-\infty}^{z} \frac{1}{\sigma(\mathbf{x})\sqrt{2\pi}} \exp\Big(- \frac{(y-\mu(\mathbf{x}))^2}{2\sigma^2(\mathbf{x})}\Big) dy. \end{aligned} \tag{5.2.1}$$

For the trivial segmentation T, a pixel with intensity $g(\mathbf{x})$ will have $\mu(\mathbf{x}) = g(\mathbf{x})$ and $\sigma^2(\mathbf{x}) = 0$. Thus for $K = T$ we have

$$F^{K,\mathbf{x}}(z) = H(z - g(\mathbf{x})).$$

We generalize our notation so that F^K denotes the set of cdfs $\{F^{K,\mathbf{x}} : \mathbf{x} \in \Omega\}$, each cdf corresponding to a pixel in Ω. Motivated by the discussion earlier in this section, we would like an energy functional of the form

$$E(K) = D(F^K, F^T) + \lambda \cdot \ell(K)$$

for some metric D. Our problem is to determine a suitable choice of D. The metric D should satisfy the following requirements:

1. Suppose K' is obtained by merging any two adjacent regions of K. Then it should be easy to calculate the quantity $D(F^{K'}, F^T) - D(F^K, F^T)$ so that the quantity $\Delta E = E(K') - E(K)$ is likewise easy to calculate.

2. The distance D should satisfy the usual properties of a metric. In particular, $D(F^K, F^T)$ should be finite for any segmentation K. Note that since the cumulative distribution is defined on $(-\infty, \infty)$ this is not trivial.

3. The metric should be stable in the following sense: Given two segmentations K_1, K_2, if the parameters θ_1, θ_2 corresponding to K_1, K_2 are perturbed by a small amount, say, $\epsilon > 0$, then $D(F^{K_1}, F^{K_2})$ should also be perturbed by a small amount, no greater than $\delta(\epsilon)$ for some function $\delta : \mathbf{R}^+ \to \mathbf{R}^+$. Hence if two regions are merged and their mean and variance are about equal then the difference between $D(F^{K'}, F^T)$ and $D(F^K, F^T)$ should be small. In particular if the two regions have equal mean and variance then the difference should be exactly zero. Conversely, the above difference should be large if either the sample mean or variance is significantly different in the two regions.

The first item suggests if K' is obtained by merging two regions R_1, R_2 of K then $D(F^K, F^{K'})$ only depends on pixels belonging to $R_1 \cup R_2$. Thus we assume that D satisfies

$$D(F^K, F^{K'}) = \sum_{\mathbf{x}\in\Omega} d(F^{K,\mathbf{x}}, F^{K',\mathbf{x}}), \tag{5.2.2}$$

where d is a metric measuring the difference between two cdfs. Note that if $\mathbf{x} \notin R_1 \cup R_2$ then $F^{K,\mathbf{x}} = F^{K',\mathbf{x}}$ and $d(F^{K,\mathbf{x}}, F^{K',\mathbf{x}}) = 0$. Given (5.2.2), the problem of determining D reduces to that of defining d. We have considered various possibilities for the metric $d(\cdot,\cdot)$, including the *Kolmogorov-Smirnov* metric (Utgoff and Clouse, [87]), the *Skorohod* metric (Billingsley, [11]) and the *Kakutani-Hellinger* metric (Anh et al. [5]). However, the first two of these do not satisfy the stability condition and the last is more relevant to comparing pdfs than cdfs. Instead we have opted for one of the simplest metrics for cdfs: the L^1 metric over $(-\infty, \infty)$.

$$d(F, F') = ||F - F'||_{L^1(x)} = \int_{-\infty}^{\infty} |F(t) - F'(t)|dt. \tag{5.2.3}$$

It is easily verified the L^1 metric satisfies the stability condition. Thus we choose to define the metric D by

$$D(F^K, F^{K'}) = \sum_{\mathbf{x}\in\Omega} d(F^{K,\mathbf{x}}, F^{K',\mathbf{x}}) = \sum_{\mathbf{x}\in\Omega} \int_{-\infty}^{\infty} \left| F^{K,\mathbf{x}}(z) - F^{K',\mathbf{x}}(z) \right| dz. \tag{5.2.4}$$

In other words, we have combined the ideas (5.2.2) and (5.2.3). We turn now to the computation of the metric (5.2.4).

Recall that equation (5.2.1) specifies the form of the cdfs involved. It is well known that (5.2.1) can be expressed as

$$F^{K,\mathbf{x}}(z) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\Big(\frac{z-\mu(\mathbf{x})}{\sigma(\mathbf{x})\sqrt{2}}\Big), \tag{5.2.5}$$

where erf is the error function, defined by $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$. If $\sigma(\mathbf{x}) = 0$ then we use the convention that $\mathrm{erf}((z-\mu(\mathbf{x})/\sigma(\mathbf{x})\sqrt{2}) = \mathrm{sign}(z-\mu(\mathbf{x}))$ where $\mathrm{sign}(x) = 1$ for positive x, -1 for negative x and 0 for $x = 0$. Using (5.2.5) it is possible to show that for any segmentations K_1, K_2, pixel $\mathbf{x}$ and μ_i, σ_i^2 corresponding to K_i, we have

$$\begin{aligned} d(F^{K_1,\mathbf{x}}, F^{K_2,\mathbf{x}}) &= |\mu_2 - \mu_1|\mathrm{erf}\Big(\frac{|\mu_2-\mu_1|}{|\sigma_2-\sigma_1|\sqrt{2}}\Big) \\ &\quad + \frac{|\sigma_2-\sigma_1|\sqrt{2}}{\sqrt{\pi}}\exp\Big(-\frac{|\mu_2-\mu_1|^2}{2|\sigma_2-\sigma_1|^2}\Big). \end{aligned} \tag{5.2.6}$$

The details are given in Appendix A. Expression (5.2.6) is clearly finite for all $\Delta\mu, \Delta\sigma$, proving d is always finite. Thus $D(F^{K_1}, F^{K_2})$ is always finite for any two segmentations. Although the choice of L^1-norm is somewhat ad-hoc, we note that a significant advantage of this norm is that $d(F^{K_1,\mathbf{x}}, F^{K_2,\mathbf{x}})$ reduces to a closed-form expression. Returning to (5.2.4), we now have the closed-form expression

$$\begin{aligned} D(F^K, F^T) = \sum_{\mathbf{x}\in\Omega} &\Big[|\mu(\mathbf{x}) - g(\mathbf{x})|\mathrm{erf}\Big(\frac{|\mu(\mathbf{x})-g(\mathbf{x})|}{\sigma(\mathbf{x})\sqrt{2}}\Big) \\ &+ \frac{\sigma(\mathbf{x})\sqrt{2}}{\sqrt{\pi}}\exp\Big(-\frac{(\mu(\mathbf{x})-g(\mathbf{x}))^2}{2\sigma^2(\mathbf{x})}\Big)\Big]. \end{aligned} \tag{5.2.7}$$

However, as explained next, there is still a problem. Although (5.2.7) is an analytic expression, it involves a summation of all the $g(\mathbf{x})$'s which is inconvenient for large regions. In practice, since we are interested in changes in the energy functional rather than absolute values, we will actually calculate quantities of the form $D(F^{K'}, F^T) - D(F^K, F^T)$. In this case the summation simplifies by removing all pixels in Ω that

are not in either of the two regions involved in the merge. However, the computation is still expensive since for N pixels there are $N-1$ region merging operations and the later operations will involve large numbers of pixels leading to a complexity of $\mathcal{O}(N^2)$. In order to make the computation practical, we invoke a simplifying assumption. We assume that

$$D(F^{K'}, F^T) - D(F^K, F^T) = D(F^{K'}, F^K). \tag{5.2.8}$$

Equation (5.2.8) is false because the properties of a metric only imply the triangle inequality, not an equality. This fact is the price we pay for considering cdfs instead of pdfs. In any case our experiments show that for purposes of running the algorithm the errors incurred in (5.2.8) are negligible and do not compound.

With (5.2.8) in place we can now easily calculate the change in energy when two regions are merged. Let the two regions be R_1, R_2 and let their parameters be $\theta_1 = (\mu_1, \sigma_1^2)$ and $\theta_2 = (\mu_2, \sigma_2^2)$, respectively. Denote their common boundary by $\ell(\partial(R_1, R_2))$ and denote the merged region by $R_3 = R_1 \cup R_2$. Then given (5.2.8) we see that

$$\begin{aligned} \Delta E &= E(K') - E(K) \\ &= D(F^{K'}, F^T) - D(F^K, F^T) - \lambda \cdot \ell(\partial(R_1, R_2)) \\ &\approx D(F^{K'}, F^K) - \lambda \cdot \ell(\partial(R_1, R_2)). \end{aligned}$$

Using (5.2.2) this becomes

$$\begin{aligned} \Delta E = \quad & \#(R_1) d(F_{\mathcal{N}(\theta_1)}, F_{\mathcal{N}(\theta_3)}) + \#(R_2) d(F_{\mathcal{N}(\theta_2)}, F_{\mathcal{N}(\theta_3)}) \\ & -\lambda \cdot \ell(\partial(R_1, R_2)), \end{aligned} \tag{5.2.9}$$

where $d(\cdot, \cdot)$ is given by (5.2.6) and θ_3 are the parameters for the new region $R_3 = R_1 \cup R_2$.

It is important to point out a subtle distinction between the optimal and sample parameters of mean/variance for each region. In our earlier work (Crisp and Tao, [27]), we showed that if the segmentation boundary K is fixed then θ minimizes the energy E when

$$\mu_i = \frac{1}{\#(R_i)} \sum_{R_i} g(\mathbf{x}), \tag{5.2.10}$$

$$\sigma_i^2 = \frac{1}{\#(R_i)} \sum_{R_i} (g(\mathbf{x}) - \mu_i)^2. \tag{5.2.11}$$

It turns out in the new formulation, equations (5.2.10), (5.2.11) no longer minimize the energy functional for fixed K and are therefore not optimal. Calculating the optimal values of μ_i, σ_i^2 has proved to be an analytically intractable optimization problem. Hence we take the practical solution of considering only segmentations where θ equals the sample mean/variance for each region determined by K. In other words we solve

$$\hat{M} = \arg\min_{M} E(M) = \arg\min_{K,\theta} E(K, \theta) \tag{5.2.12}$$

subject to (5.2.10) and (5.2.11). Thus E is a function of K alone.

5.3 The Basic Algorithm

As mentioned before, our basic algorithm is a continuation of previous work (Crisp and Newsam, [26], Redding et al., [67], Robinson et al., [72]) motivated by efforts to improve on Koepfler's algorithm [42]. We recall that the main differences are that we avoid the use of the λ-schedule and also our algorithm always locates the globally best merge at each merging operation, at the expense of complex data structures. The basic outline of our algorithm is as follows:

1. Initialize the data with the trivial segmentation, where each pixel is its own region.

2. Calculate the merge cost for each possible region merge. The merge cost is given by setting (5.2.9) equal to 0 and using (5.2.6). Thus for any pair of adjacent regions R_1, R_2 and $R_3 = R_1 \cup R_2$,

$$\begin{aligned}\lambda \;=\; & \frac{1}{\ell(\partial(R_1,R_2))}\Big[\#(R_1)|\mu_3-\mu_1|\mathrm{erf}\Big(\frac{|\mu_3-\mu_1|}{|\sigma_3-\sigma_1|\sqrt{2}}\Big)\\ & +\#(R_1)\frac{|\sigma_3-\sigma_1|\sqrt{2}}{\sqrt{\pi}}\exp\Big(-\frac{|\mu_3-\mu_1|^2}{2|\sigma_3-\sigma_1|^2}\Big)\\ & +\#(R_2)|\mu_3-\mu_2|\mathrm{erf}\Big(\frac{|\mu_3-\mu_2|}{|\sigma_3-\sigma_2|\sqrt{2}}\Big)\\ & +\#(R_2)\frac{|\sigma_3-\sigma_2|\sqrt{2}}{\sqrt{\pi}}\exp\Big(-\frac{|\mu_3-\mu_2|^2}{2|\sigma_3-\sigma_2|^2}\Big)\Big].\end{aligned}$$

3. Find the pair of regions with smallest merge cost and merge them.

4. Repeat until only one region remains.

5. Go over the list of all segmentations and choose the "best one". This corresponds to choosing the best λ.

Despite the surface simplicity of the region merging algorithm, much care is needed in choosing good data structures to avoid inefficient performance in time [27]. For example, given an $N \times N$ image and the trivial segmentation, there are N^2 regions and $\mathcal{O}(2N^2)$ pairs of adjacent regions. Determining the pair of adjacent regions with smallest merge cost would be extremely expensive unless a sorted list is somehow maintained. Similarly each region must keep and update a record of its set of neighbouring regions. A detailed analysis of this algorithm and Koepfler's algorithm is provided in (Redding et al., [67]).

5.4 Test Images and Experiments

Our aim in this section is to compare the performance of the old algorithm described in (Crisp and Newsam, [26], Crisp and Tao, [27]), and the new, the idea being to assess the suitability of our proposed solution to the small sample problem. For convenience we denote the old and new algorithms as FLSA-MAP and FLSA-CDF respectively. In order to assess segmentation performance we would ideally like a set of real images with associated ground truth. However, despite the vast amount of research on the problem of image segmentation there is no generally accepted test suite of images, nor do there exist many standard benchmark tests and algorithms that are easily obtained. We note that progress is being made, see for instance the Berkeley data set (Martin et al., [50]), but that is more suitable for edge detection algorithms rather than segmentation. A good review of the situation with several suggested approaches is given by Zhang in [93]. We have chosen to use synthetic imagery, as is commonly done. Choosing the correct set of synthetic test images for experiments is a difficult task (Zhang, [94]). Many papers display experimental results without justification of why their chosen test images are "correct" for testing purposes. We do not claim that our tests convincingly prove the advantages of our new algorithm; they do give a general indication of under what conditions the new algorithm performs well.

We performed a number of experiments on four synthetic images, shown in figure (5.7.1), and the standard `HOUSE` image, used by (Kanungo et al., [39]). For the synthetic images, the properties of interest are region shape (first image), contrast between adjacent regions of different size (second image), noise level (third image) and size (fourth image). In these experiments we measure both the time taken for the algorithm and its accuracy. The accuracy measure is described in the next subsection. All test images are of size 256×256 with 256 gray levels. Each of the four synthetic images were tested with various parameter settings, and the diagrams in figure (5.7.1) show only images for a particular set of parameter values. In all images except the third, the variance is a parameter. In the third image, the circle radius is a parameter. The first two images have variance 20, and the third and fourth images have radius 28 and variance 60 respectively. The image names from left to right are `SHAPE, CONTRAST, NOISE` and `RADIUS`. For each of the four synthetic images, we tested both FLSA-MAP and FLSA-CDF. Recall that the FLSA-MAP requires a variance offset σ_0^2 to be specified. Since there is no rigorous way of choosing σ_0^2 we specify a range of values and consider the results from them all. The following representative values were used: 10^{-i} where i is an integer between 0 and 9. We use an offset of zero to denote the FLSA-CDF algorithm since a zero offset has no meaning for FLSA-MAP. This is merely for convenience of displaying the results (Tables 5.7.1 - 5.7.9). Note that the FLSA-MAP does not "converge" to the FLSA-CDF as the offset approaches zero.

5.5 The Accuracy Measure

The accuracy metric is defined as follows: a pixel is called an "edge pixel" if any of its four (less than four if it lies in $\partial\Omega$) neighbours is in a different region. For each edge pixel in the ground truth, measure the Euclidean distance to the closest edge pixel in the actual segmentation and vice versa (reverse the roles of ground truth and actual segmentation). Then sum up all the distances to yield the total

"score". A score of zero indicates a perfect or near perfect segmentation (note that zero indicates that each pixel is correctly classified as edge-pixel or non-edge-pixel, although the segmentations may still be slightly different, but we believe the error is negligible, since the horizontal-vertical discretization of Ω implies a measurement error anyway). This is the "Boundary Distance Measure" used by Kanungo et al. [39]. A high score implies a serious mismatch in the segmentation, for instance one region is completely missing or the boundary of a region "oscillates" much more than it should. Since there is no ground truth for the `HOUSE` image, we decided to compare it subjectively with the results in (Kanungo et al., [39]).

5.6 Experimental results

5.6.1 The `SHAPE` Image

The first image, called `SHAPE` (Figure 7), consists of ellipses with various eccentricities with mean 100 against a background of mean 200. We tested the image with various amounts of noise, more specifically, variance = 0,20,40,60. The algorithm was instructed to stop at 6 regions. The time and accuracy results are in Tables 5.7.1 and 5.7.2 respectively, where σ_0^2 represents the variance offset and 'var' represents the variance. Timewise, FLSA-CDF was better for non-zero variance and worse with zero variance. For variance of 0 or 20, no circles were lost. Obviously we had a perfect segmentation with variance 0. With variance of 40, FLSA-MAP lost one or all ellipses with offset $\sigma_0^2 = 10^{-3}$ or less. In fact FLSA-CDF outscored FLSA-MAP for all values of offset. With variance of 60, FLSA-CDF also outscored FLSA-MAP for all values of offset since it lost only most of the bottom ellipse. FLSA-MAP could not retain any ellipse, with any choice of variance offset except one ellipse when $\sigma_0^2 = 10^{-1}$.

5.6.2 The CONTRAST Image

The second image, called CONTRAST, is a 4x4 chessboard with small circles. The black and white squares have mean 100, 200 respectively and the circles have mean 150. Again the image was tested with variance levels 0,20,40,60. The algorithm was to stop at 20 regions. The time and accuracy results are in Tables 5.7.3 and 5.7.4 respectively, where σ_0^2 represents the variance offset and 'var' represents the variance. Again FLSA-CDF did better than FLSA-MAP timewise with non-zero variance, but worse with zero variance. For variance of 0, both FLSA-CDF and FLSA-MAP (any offset) recovered the exact segmentation. For variance of 20, FLSA-MAP lost three circles out of four when offset was 10^{-9} or 10^{-8}. With any other offset FLSA-MAP retained all circles, as did FLSA-CDF. With variance 40, FLSA-CDF outscored FLSA-MAP even with the best offset for the latter. FLSA-CDF lost none of the circles. With offset of 10^{-9} FLSA-MAP lost all circles and squares. With offset 10^{-8} or 10^{-7} FLSA-MAP only kept two and five large squares respectively. The symbol '-X' in Table 5.7.4 indicates at least one square was lost, clearly much worse than losing any number of circles. With an offset 10^{-6} or greater, FLSA-MAP lost 3 or 4 circles, but kept the squares. In some of these cases the shape-errors in the squares outweighed the number of circles lost. With variance of 60, FLSA-CDF's superiority was not as pronounced. Although FLSA-CDF kept the squares, it lost all circles and was outscored by FLSA-MAP when $\sigma_0^2 > 10^{-2}$.

5.6.3 The NOISE Image

The third image, called NOISE, consists of 9 circles of equal radii on a background, all with mean 128. The background has variance 5 and each circle has a different level of variance ranging from 20-100. The algorithm was to stop at 10 regions. We tested four different versions of the image where the radius of the circles changed: the radius was 4,12,20,28. The results are in Tables 5.7.5 and 5.7.6 respectively, where σ_0^2 represents the variance offset and 'rad' represents the radius of the circles. We counted a circle as being "retained" if at least one pixel is classified as being different from the background. This is because the difference between a single pixel and no pixel results in an order-of-magnitude difference in the accuracy metric defined above. Note that for images NOISE and RADIUS (subsection 5.6.4) the columns are in order of *decreasing* radius/variance respectively so that the difficulty of segmenting the image increases from left to right, which is consistent with the tables for the first two images.

FLSA-CDF's time-superiority is evident from Table 5.7.5. FLSA-MAP takes significantly less time when the offset exceeds the critical value of about 10^{-7} to 10^{-6}, but is still nowhere near as good as FLSA-CDF. Also note that the ground truth is different for different values of radius (unlike variance). On one hand, we would expect that for small radius the scores to be lower since there are less pixels in the ground truth and the accuracy measure sums up pixel distances. On the other hand one could argue that for small radius, the circles are harder to distinguish (especially ones with small variance) so the scores should be higher. Thus comparing the accuracy for different values of the radius is not really meaningful. However, it still makes sense to compare FLSA-MAP and FLSA-CDF for the same value of radius. FLSA-CDF is comparable with FLSA-MAP with offset of $\sigma_0^2 = 10^{-9}$, but worse when $10^{-8} \leq \sigma_0^2 \leq 10^{-3}$, and better when $\sigma_0^2 \geq 10^{-2}$. Thus FLSA-CDF's "superiority" (if any) is not as pronounced as for the first two images.

Example outputs for Image NOISE with radius 12 are given in Figure 5.7.2. The 2nd-4th diagrams are segmentation masks corresponding to a variance offset of $10^{-9}, 10^{-5}, 10^{-2}$ respectively. The final diagram is the segmentation mask corresponding to FLSA-CDF.

Since FLSA-CDF performed worse than the other images we decided to investigate the effect of removing the upper left circle, resulting in Image NOISE2 (not shown). This time the algorithm was instructed to retain only nine regions. We did not display the time taken since it is not significantly different from before. Here we counted the upper left circle as missing even though the ground truth did not have it. Thus any algorithm must "lose" at least one circle. This is to keep the data for Image NOISE2 consistent with the data for Image NOISE. If FLSA-CDF were allowed to keep ten regions instead of nine it would lose only one circle scoring 957, a 10-fold improvement. Again FLSA-CDF did poorly compared to the first two images. Note that in both Images NOISE and NOISE2, FLSA-MAP did better than FLSA-CDF (assuming correct offset), but not by an order of magnitude (except the case of radius 12). This was because FLSA-CDF's fractal-like boundaries were worse than FLSA-MAP's even though they recovered the same number of "circle-like objects". Of course, the scores in Image NOISE2 were lower than in Image NOISE since the removal of one circle would leave less error.

5.6.4 The RADIUS Image

The last image, called RADIUS, consists of 9 circles of equal variance but different radii ranging from 4 to 36, on a background of variance 5. The circles and background all had the same mean. We tested four different versions of the image where the variance of the circles changed: the variance was 20,40,60,80. Again the algorithm was to stop at 10 regions. With variance of 20 FLSA-CDF lost 1 circle but FLSA-MAP also lost at least one circle for any offset. Thus FLSA-CDF was clearly better. With variance of 40 FLSA-MAP was generally better than FLSA-CDF unless the

offset exceeded 10^{-3}. FLSA-CDF was only comparable with FLSA-MAP with an offset of 10^{-9}. With variance 60 or 80 FLSA-CDF outscored FLSA-MAP only when the offset exceeded 10^{-3}. With lower values of offset FLSA-MAP did significantly better than FLSA-CDF, even when the offset was 10^{-9}. Example outputs for Image `RADIUS` are given in Figure 5.7.3. The 2nd-4th diagrams are segmentation masks corresponding to variance offset equal to $10^{-8}, 10^{-5}, 10^{-3}$ respectively. The final diagram is the segmentation mask corresponding to FLSA-CDF.

5.6.5 The `HOUSE` Image

The `HOUSE` image is a standard image, shown in Figure 5.7.4. Note that, unlike the synthetic images, the correct number of regions is not obvious. Thus, we have instead specified a value of λ, namely $\lambda = 1.585 \approx \log(3)/\log(2)$. Kanungo et al. [39] showed that the Minimum Description Length functional yields an approximately linear weighting between model complexity and accuracy (ignoring a few other terms such as the code length for encoding coefficients of polynomials and so on) and that $\lambda \approx \log(3)/\log(2)$ is a reasonable approximation to the "correct" value of scale parameter. We noted that the algorithm always ran in under 5 seconds on a Sun Microsystems machine regardless of whether variance offset was used or not. With offset of 10^{-3} the image was heavily oversegmented (768 regions), with an offset of 10^{-2} the image was "about right" (61 regions), and with an offset of 10^{-1} a number of regions were over-merged (10 regions), which indicated a very small margin of error for guessing σ_0. The algorithm without variance offset resulted in a segmentation close to the result for offset $= 10^{-2}$. In Figure 5.7.4 the 2nd-4th diagrams correspond to variance offset equal to $10^{-3}, 10^{-2}, 10^{-1}$ respectively. The final diagram is the segmentation mask corresponding to FLSA-CDF. We believe our segmentation results are comparable with Kanungo's. In particular, the final diagram of Figure 5.7.4 seems to indicate our algorithm is slightly better than Kanungo's at eliminating small regions, although some of our regions have been overmerged.

5.7 Conclusions

We have described a new segmentation algorithm called the FLSA-CDF algorithm. It is based on the algorithm described in (Crisp and Newsam, [26], Crisp and Tao, [27]), which we denote by FLSA-MAP. The FLSA-CDF algorithm proposes a new solution of the small sample problem. It avoids the problem of guessing a suitable value of variance offset parameter, and runs significantly faster than the FLSA-MAP, except in the trivial case of an image with zero noise. The offset parameter is eliminated by considering the cumulative distribution function instead of the probability density to calculate merge costs. Removing this parameter is a significant advantage since the tests show that the best choice of offset parameter varies from image to image and it is difficult to justify one value over another. The new algorithm has some disadvantages: it resorts to a number of approximations such as equations (5.2.8), (5.2.10), and (5.2.11).

From the experimental results, we conclude the new algorithm is more robust to difficult images with poor SNR/contrast, but given an "easy image" there exist values of offset parameter such that FLSA-MAP outscores FLSA-CDF, assuming that there exists some solution to the choice of scale parameter λ (for instance in the ground truth, we know the number of regions and can set λ accordingly). We note that generally in the latter case, FLSA-MAP only "just" outperforms FLSA-CDF (i.e. not by an order of magnitude) but the superiority of FLSA-CDF over FLSA-MAP is very significant in the former. For example in Image `RADIUS` FLSA-CDF really outshines FLSA-MAP for variance of 20 but FLSA-CDF does poorly against FLSA-MAP with variance of 80 or 60. In Image `SHAPE`, FLSA-CDF dominates completely for variance of 60 but for variance of 20 it is just outscored for FLSA-MAP with offset 10^{-6} or higher (FLSA-MAP has fewer "small shape errors" than FLSA-CDF). In Image `NOISE` FLSA-CDF does not do well at all and we believe this is because changing the radius of circles does not really correspond to good or poor SNR/contrast. We believe that in the latter case the reason for slightly worse

performance without variance offset in "easy images" is due to the approximation in (5.2.8) and the fact that sample mean/variance no longer optimizes the new energy functional with respect to θ. We noted that the old algorithm ran much slower when segmenting a noisy image with few regions, as compared to an image with a high amount of structure, such as the `HOUSE` image. A significant weakness of the new algorithm is the inability to handle multi-band images. The reason for this is our inability to generate a closed form expression for the difference of two cdfs according to some integral expression such as (5.2.3) since otherwise the algorithm would take too much time updating merge costs at each merging iteration. The multivariate normal distribution is given by

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}[x-\mu]'\Sigma^{-1}[x-\mu]\Big) \tag{5.7.1}$$

for mean vector μ and covariance matric Σ. Its cdf is generally not an analytic expression (Tong, [84]). In other words, it cannot be expressed in terms of simple functions, including error functions as for the 1-dimensional case. However, it is possible to investigate the effect of approximating (5.7.1) by a simpler, more tractable expression. We think this is a promising avenue of future research.

As a final remark, we note that the idea of considering the cdf of a probability density can apply equally well to many distributions, not just the Guassian. For example, if the order parameter of a Gamma distribution tends to infinity, the resulting density becomes a Dirac delta, the same result as a Gaussian with zero variance.

$\sigma_0^2 \backslash$var	0	20	40	60
0	18.0	5.1	9.0	5.5
10^{-9}	8.2	68.9	29.8	38.5
10^{-8}	20.8	22.5	23.0	41.1
10^{-7}	8.2	30.6	34.3	36.0
10^{-6}	8.2	33.3	36.2	37.9
10^{-5}	8.2	33.0	66.9	23.2
10^{-4}	7.9	18.8	100.2	18.0
10^{-3}	7.9	10.9	65.3	11.2
10^{-2}	7.9	17.8	38.3	17.3
10^{-1}	7.8	24.0	49.7	54.0
10^{-0}	7.5	20.3	44.9	57.5

Table 5.7.1: Time taken for Image `SHAPE`.

$\sigma_0^2 \backslash$var	0	20	40	60	0	20	40	60
0	0	12	553	2.1×10^4	0	0	0	0
10^{-9}	0	102	3.4×10^5	3.5×10^5	0	0	-5	-5
10^{-8}	0	22	3.6×10^5	3.5×10^5	0	0	-5	-5
10^{-7}	0	16	3.4×10^5	3.5×10^5	0	0	-5	-5
10^{-6}	0	6	3.4×10^5	3.5×10^5	0	0	-5	-5
10^{-5}	0	6	3.6×10^4	3.4×10^5	0	0	-1	-5
10^{-4}	0	6	3.6×10^4	2.6×10^5	0	0	-1	-5
10^{-3}	0	8	3.5×10^4	2.5×10^5	0	0	-1	-5
10^{-2}	0	8	603	2.5×10^5	0	0	0	-5
10^{-1}	0	10	712	1.4×10^5	0	0	0	-4
10^{-0}	0	10	741	2.5×10^5	0	0	0	-5

Table 5.7.2: Accuracy results/No. of ellipses lost for Image `SHAPE`.

$\sigma_0^2 \backslash$var	0	20	40	60
0	7.0	5.3	6.2	5.3
10^{-9}	4.2	10.8	22.3	32.2
10^{-8}	7.2	10.5	19.1	36.4
10^{-7}	4.3	10.6	13.7	33.2
10^{-6}	4.3	9.8	12.0	25.4
10^{-5}	4.2	9.0	11.1	27.5
10^{-4}	4.1	7.0	10.8	13.6
10^{-3}	4.2	5.7	7.7	7.1
10^{-2}	4.2	5.6	7.1	6.1
10^{-1}	4.2	5.8	7.1	7.6
10^{-0}	3.9	5.9	6.8	8.0

Table 5.7.3: Time taken for Image `CONTRAST`.

$\sigma_0^2 \backslash$var	0	20	40	60	0	20	40	60
0	0	126	1.1×10^3	6.9×10^3	0	0	0	-4
10^{-9}	0	1.2×10^3	5.8×10^5	5.8×10^5	0	-3	-X	-X
10^{-8}	0	1.1×10^3	1.8×10^5	5.8×10^5	0	-3	-X	-X
10^{-7}	0	319	1.3×10^5	5.8×10^5	0	0	-X	-X
10^{-6}	0	179	3.3×10^3	5.7×10^5	0	0	-4	-X
10^{-5}	0	79	1.6×10^3	5.1×10^5	0	0	-4	-X
10^{-4}	0	88	1.5×10^3	5.0×10^5	0	0	-4	-X
10^{-3}	0	95	2.0×10^3	1.5×10^4	0	0	-4	-4
10^{-2}	0	119	1.4×10^3	6.0×10^3	0	0	-3	-4
10^{-1}	0	121	3.2×10^3	5.6×10^3	0	0	-3	-4
10^{-0}	0	197	4.1×10^3	5.8×10^3	0	0	-4	-4

Table 5.7.4: Accuracy results/No. of circles lost for Image `CONTRAST`.

off \ rad	28	20	12	4
0	6.8	6.2	5.8	5.5
10^{-9}	65.5	68.6	51.2	39.9
10^{-8}	65.2	69.7	55.8	32.3
10^{-7}	55.4	47.0	50.5	52.7
10^{-6}	23.6	28.9	32.5	31.4
10^{-5}	14.2	20.0	24.2	24.4
10^{-4}	14.7	17.3	20.9	16.8
10^{-3}	18.5	21.6	24.4	27.2
10^{-2}	18.4	22.6	23.9	25.4
10^{-1}	21.4	22.2	24.1	26.2
10^{-0}	21.8	22.8	24.0	26.4

Table 5.7.5: Time taken for Image `NOISE`.

σ_0^2\ rad	28	20	12	4	28	20	12	4
0	1.5×10^4	1.3×10^4	9.0×10^3	1.0×10^4	-1	-1	-1	-3
10^{-9}	1.6×10^4	1.3×10^4	8.9×10^3	6.8×10^3	-1	-1	-1	-2
10^{-8}	2.9×10^3	1.9×10^3	958	3.4×10^3	0	0	0	-1
10^{-7}	2.6×10^3	1.3×10^3	808	3.3×10^3	0	0	0	-1
10^{-6}	1.8×10^3	1.0×10^3	765	3.3×10^3	0	0	0	-1
10^{-5}	1.5×10^3	961	743	128	0	0	0	0
10^{-4}	1.6×10^3	1.0×10^3	564	98	0	0	0	0
10^{-3}	1.4×10^4	1.2×10^4	8.8×10^3	3.4×10^3	-1	-1	-1	-1
10^{-2}	4.3×10^4	4.3×10^4	2.8×10^4	1.0×10^4	-3	-3	-3	-3
10^{-1}	2.1×10^5	2.1×10^5	8.0×10^4	5.0×10^5	-7	-7	-6	-8
10^{-0}	2.8×10^5	1.9×10^5	8.0×10^4	5.0×10^5	-7	-7	-6	-8

Table 5.7.6: Accuracy results /No. of circles lost for Image `NOISE`.

$\sigma_0^2 \backslash$ rad	28	20	12	4	28	20	12	4
0	1.9×10^3	3.0×10^3	9.1×10^3	6.8×10^3	-1	-1	-2	-3
10^{-9}	1.4×10^3	835	614	3.5×10^3	-1	-1	-1	-2
10^{-8}	1.1×10^3	872	402	115	-1	-1	-1	-1
10^{-7}	1.0×10^3	677	334	94	-1	-1	-1	-1
10^{-6}	932	610	321	84	-1	-1	-1	-1
10^{-5}	863	573	310	71	-1	-1	-1	-1
10^{-4}	855	559	312	60	-1	-1	-1	-1
10^{-3}	1.3×10^3	964	545	3.1×10^3	-1	-1	-1	-2
10^{-2}	1.7×10^4	2.9×10^4	1.9×10^4	7.1×10^3	-3	-3	-3	-3
10^{-1}	1.6×10^5	1.7×10^5	6.4×10^4	4.3×10^4	-7	-7	-6	-8
10^{-0}	2.3×10^5	1.5×10^5	6.4×10^4	4.3×10^4	-7	-7	-6	-8

Table 5.7.7: Accuracy results/No. of circles lost for Image `NOISE2`.

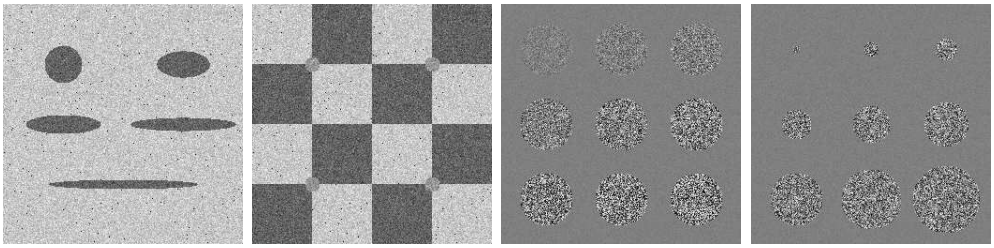


Figure 5.7.1: Four different synthetic images to be segmented.

$\sigma_0^2 \backslash$var	80	60	40	20
0	5.8	6.0	6.4	6.3
10^{-9}	42.4	43.6	44.0	65.1
10^{-8}	38.0	39.1	41.2	46.6
10^{-7}	44.9	44.0	45.4	49.4
10^{-6}	22.4	22.1	23.0	24.9
10^{-5}	16.2	15.4	16.3	18.3
10^{-4}	16.0	16.8	16.2	16.8
10^{-3}	18.3	19.2	19.4	20.4
10^{-2}	19.3	20.4	21.7	23.2
10^{-1}	19.5	21.5	21.2	22.5
10^{-0}	19.3	22.0	21.7	22.4

Table 5.7.8: Time taken for Image `RADIUS`. σ_0^2 = variance offset, var = variance of added noise.

$\sigma_0^2 \backslash$var	80	60	40	20	80	60	40	20
0	9.7×10^2	1.2×10^3	4.8×10^3	7.9×10^3	0	0	-1	-1
10^{-9}	5.6×10^2	8.1×10^2	4.5×10^3	1.0×10^5	0	0	-1	-6
10^{-8}	4.7×10^2	7.1×10^2	4.2×10^3	2.6×10^4	0	0	-1	-3
10^{-7}	4.3×10^2	6.5×10^2	1.3×10^3	1.7×10^4	0	0	0	-2
10^{-6}	3.8×10^2	5.2×10^2	1.0×10^3	7.4×10^3	0	0	0	-1
10^{-5}	3.9×10^2	5.3×10^2	9.6×10^2	5.8×10^3	0	0	0	-1
10^{-4}	4.1×10^2	5.5×10^2	1.0×10^3	6.4×10^3	0	0	0	-1
10^{-3}	6.4×10^2	8.8×10^2	4.5×10^3	1.8×10^4	0	0	-1	-2
10^{-2}	2.3×10^3	6.4×10^3	5.2×10^4	1.0×10^5	0	-1	-5	-6
10^{-1}	1.0×10^5	1.2×10^5	6.9×10^4	1.0×10^5	-7	-6	-5	-6
10^{-0}	1.6×10^5	1.2×10^5	6.9×10^4	1.0×10^5	-7	-6	-5	-6

Table 5.7.9: Accuracy results /No. of circles lost for Image `RADIUS`.

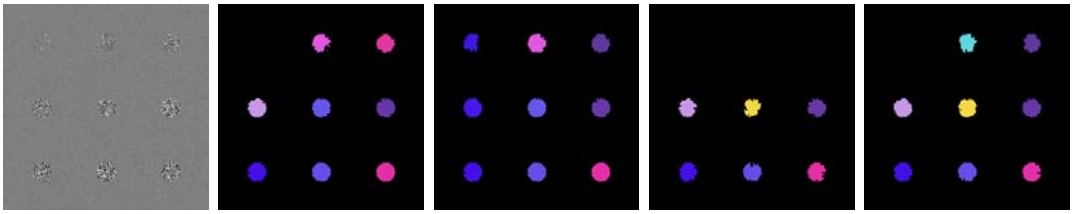


Figure 5.7.2: Comparison of algorithms FLSA-MAP and FLSA-CDF for Image `NOISE`.

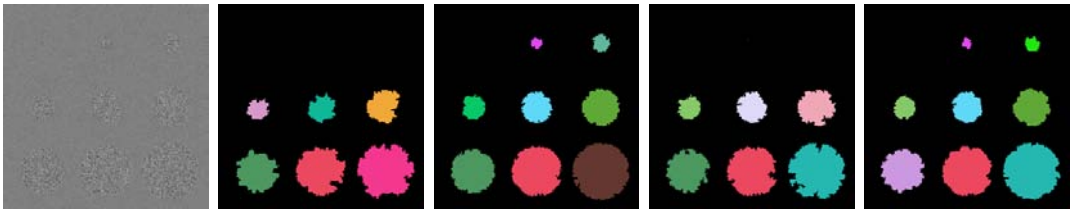


Figure 5.7.3: Comparison of algorithms FLSA-MAP and FLSA-CDF for Image `RADIUS`.

Figure 5.7.4: Comparison of algorithms FLSA-MAP and FLSA-CDF for Image `HOUSE`.

Chapter 6

The Modelling of Images with Texture

6.1 Introduction

Texture is an inherent property of most real images and hence texture segmentation of images is an important problem. A fundamental difference between textured and smooth images is that texture can only be defined across a neighbourhood of pixels, instead of a single pixel. Indeed this demonstrates one of the difficulties of texture segmentation. Typically the image data are presented in the form of a matrix, with each element recording a gray value intensity. Hence there is no texture information in a single element. Another difficulty of texture segmentation stems from the inability to easily define texture. Roughly speaking, texture can be characterized by repeating patterns of intensity differences. For instance an object can be described as striped, furry, uniform or checkerboard and so on, but it is difficult to represent such notions using mathematical formulae. This has been attempted: for instance, the use of co-occurrence matrices[1] was introduced by Haralick et al. [38] to model

[1]The original authors used the term "gray-tone spatial dependency matrices", but this is now obsolete.

spatial properties of an image. However, it requires extensive computation, and moreover, many "texture features" derived from this matrix do not correspond to visual perception of the human eye. At the opposite extreme, Tamura [81] defined six texture features derived directly from visual perception, but the features were measured "psychometrically" (Coombs et al., [24]) rather than mathematically.

However current research has shown much promise and many methodologies to texture segmentation have been proposed, with some success. Of course it will be impossible to cover the vast literature in detail and we only concentrate on the important aspects.

Simon Barker [8] classified segmentation methodologies into two main categories. In model-based segmentation the optimal segmentation is defined via a Bayesian sense, usually via the Maximum A-Posteriori (MAP) principle. To model spatial dependencies between adjacent pixels, the Markov Random Field (MRF) is used to specify the probabilities of all possible configurations on a lattice. The inter-spatial dependence between pixels is captured by the Markov property via specification of neighbourhoods. The neighbourhood of a pixel $\mathbf{x}$ is usually defined as the set of all pixels within a ball of radius $r > 0$ centred at $\mathbf{x}$, but excluding $\mathbf{x}$ itself. We will not discuss MRF theory in detail and instead refer the reader to [9]. In feature-based segmentation, the original image is transformed to another domain so the data reflect texture characteristics instead of individual pixel values. The transform data are further processed, for instance, via K-means clustering.

Model-based segmentation has proved attractive in theory, since there are few problems with setting arbitrary thresholds and it is easy to specify a Markov Random Field via an "energy potential" due to the well-known Hammersley and Clifford Theorem (Besag, [9]). However, the search for efficient algorithms for computing the minimizer has proved elusive, except for trivial models. Hence it is necessary to settle for a local minimum. One of the most common techniques is to update pixel intensities one at a time using Iterated Conditional Modes (ICM) (Besag, [10]), or

simulated annealing (Geman and Geman, [33]) thanks to its ease of implementation. Simulated annealing avoids the problem of local minima by allowing a temporary increase of energy. In fact, simulated annealing is a special case of ICM with temperature fixed at zero (Barker, [8]). Geman and Geman [33] proved simulated annealing always results in convergence to the global minima (under certain conditions). Unfortunately the result is only of theoretical interest since it requires the temperature be reduced extremely slowly. One of the simplest and well-known models is the Ising model (Chandler, [22]), later developed into the Potts model. In both models the random process is a lattice and energy is calculated by summing contributions from individual pixels or adjacent pairs of pixels. The model introduced by Geman and Geman [33] adds a line process and is a precedent of the celebrated Mumford-Shah model [61].

In feature-based segmentation a vector is typically associated with each pixel, each element of a vector representing some component of texture. A simple example is that used by Laws [43]. The original image is convolved with a small mask, from which statistics such as variance can be computed for each pixel. An even simpler method by Unser [85] proposes the use of Hadamard 2x2 masks to estimate local derivatives in horizontal, vertical and diagonal directions. In the use of co-occurrence matrices (Haralick et al., [38]) each entry of the matrix represents the number of times two gray levels appear in two pixels separated by a fixed distance and orientation. Thus the matrix does not represent texture directly, but it can be used to extract features such as energy, entropy and contrast (Soares et al., [78]). It has been used extensively (Davis et al., [28, 29], Haralick et al., [38], Soares et al., [78], Unser [86], Welch, [91]). Its main drawback is the computational cost, which increases quadratically with the number of gray levels of an image. There are many other ways to derive texture features and we refer to (Reed and Du Buf, [68]) for a more complete survey. Feature-based methods have proved useful in practice, since good segmentations can be readily computed in reasonable time (Barker, [8]).

Our main motivation is to extend the region merging algorithms of Koepfler (Koepfler et al., [42]), originally intended for Mumford-Shah segmentation. We propose the use of feature-based methods, simply because this was also proposed by Koepfler (Koepfler et al., [42]). According to the Mumford-Shah model [61] (we work in the discrete domain) the correct segmentation is one that minimizes

$$E(u, K) = \sum_{\mathbf{x} \in \Omega} (u(\mathbf{x}) - g(\mathbf{x}))^2 + \lambda \cdot \ell(K), \tag{6.1.1}$$

where $u(x)$ is constant on each region of $\Omega \setminus K$. Implicit in (Koepfler et al., [42]) is the assumption that the image is represented as a scalar or vectorial function with each channel representing some meaningful quantity such as colour, gray-level or wavelet transform coefficients. Furthermore, the channel data must be smooth (without texture or noise). However, the authors point out that in the vectorial case texture discrimination can be achieved by assigning suitable texture features to a number of "channels". They cite a number of papers (Malik and Perona, [47], Marr [48], Voorhees and Poggio, [88]) where texture features are derived from local operators such as convolutions or derivatives, similar to the ideas of Unser [85]. Koepfler et al. themselves propose the use of the Haar Wavelet transform, useful for multiscale analysis.

Recall that Crisp and Newsam [26] developed an "Extended Mumford-Shah" (EMS) model to account for images contaminated by white noise. This was discussed in Chapter 3. In that chapter we derived the following energy functional, where the variance was always assumed to be bounded from below by some predetermined constant σ_0^2:

$$E(K) = \sum_{i\in\mathcal{I}} \#(R_i)\frac{\log\bar{\sigma}_i^2}{2} + \sum_{i\in\mathcal{I}}\sum_{\mathbf{x}\in R_i} \frac{(g(\mathbf{x})-\bar{\mu}_i)^2}{2\bar{\sigma}_i^2} + \lambda\cdot\ell(K), \qquad (6.1.2)$$

where $\mathcal{I}$ is an indexing set for all regions and

$$\begin{aligned}\bar{\mu}_i &= \frac{1}{\#(R_i)}\sum_{\mathbf{x}\in R_i} g(\mathbf{x}),\\ \bar{\sigma}_i^2 &= \max(\sigma_0^2, \sigma_i^{*2}) = \max\left(\sigma_0^2, \frac{1}{\#(R_i)}\sum_{\mathbf{x}\in R_i}(g(\mathbf{x})-\bar{\mu}_i)^2\right).\end{aligned}$$

Essentially, equation (6.1.2) avoids excessive penalization of regions with large oscillation, provided they fit a normal distribution. The original Mumford-Shah functional enforces all regions to be smooth, thus it cannot distinguish two regions with same mean but different variance. This is demonstrated in (Crisp and Newsam, [26]).

It is worth demonstrating the difference between calculating mean/variance directly and defining them in transform domain. In a simple synthetic image consisting of two regions (Figure 6.1.1, first diagram) applying our algorithm using the EMS model yields roughly the correct segmentation (second diagram). If we instead compute the mean and variance for each pixel in a 3x3 window, say, then the resulting transform images are still noisy (third and fourth diagrams), and hence unsuitable for Koepfler's segmentation algorithm.

However, the EMS model in (Crisp and Newsam, [26]) could not discriminate between different textures since pixel intensities were assumed independent. This latter assumption was necessary to facilitate the mathematical analysis of the model and algorithm implementation. For instance, it allowed us to derive analytic formulae for optimal mean and variance parameters. In the next section we propose how to modify the EMS model to account for texture segmentation. Section 6.3 describes the experimental results and Section 6.4 is a short summary.

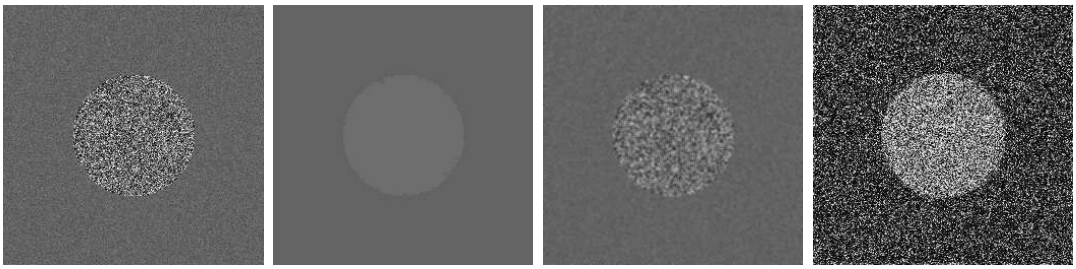


Figure 6.1.1: The difference between the EMS model and the use of transform domain for mean and variance.

6.2 Modifying the EMS Model to Account for Textures

We propose to combine the EMS model with the use of the Haar wavelet-transform to derive texture features. More specifically we convolve the original image with the masks shown in Figure 6.2.1:

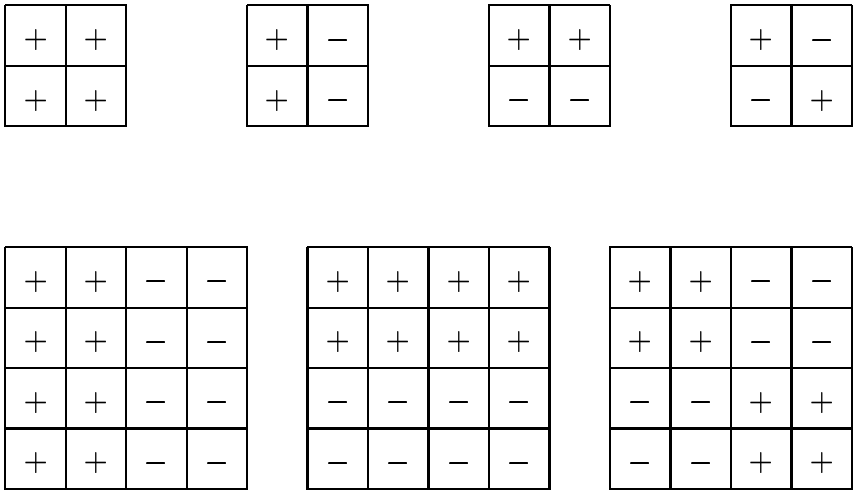


Figure 6.2.1: Seven masks used for texture segmentation.

For each mask, the plus/minus signs represent the number $\pm 1/N$ where N is the number of plus signs within that mask. Thus the first mask returns the average gray value of four pixels. The other 2x2 masks measure horizontal, vertical and diagonal derivatives. In fact these are the same masks used by Unser [85]. The three 4x4 masks are dilated versions of the latter three 2x2 masks. After applying the masks, the channels are convolved with a 3x3 Gaussian filter. We then take the absolute values of the channel outputs. We prefer this to Koepfler's half-wave rectification [42] to reduce the number of channels. Thus there are seven masks in total. These channels are then treated as a vector random variable where correlation can exist between different bands of the same pixel, but not between different pixels. Strictly speaking, correlation does exist between pixels since pixel neighbourhoods are required to calculate the texture features. In practice we ignore this to simplify the implementation of our region merging algorithm.

6.3 Experimental Results

We tested a number of synthetic and real images, described below. We use the region merging algorithm described in Chapters 2 and 3 (Crisp and Newsam, [26] and references therein). Recall that our algorithm computes a unique segmentation given any value of scale parameter λ. Equivalently we can compute a unique segmentation given a specific number of regions, which is done in the experiments below. As noted in the previous chapter, one difficulty with this algorithm is that modelling textures implies the use of multiple data channels, and it is therefore impossible to use the FLSA-CDF to solve the problem of determining a reasonable value of variance offset. In the experiments below, the variance offset is equal to 1.0 unless stated otherwise.

6.3.1 Synthetic Image

For our first experiment, we tested a synthetic image consisting of two different regions, shown in Figure 6.3.1 (left). Both regions have the same striped texture but the bottom region is contaminated by white noise. This synthetic image is similar to one used by Koepfler [42] where two similar regions differ only by a weak "secondary channel". Using the seven channels defined above we easily recover two regions as shown in Figure 6.3.1 (right). An interesting aspect of Koepfler's experiment on texture segmentation is that he deliberately suppresses the gray value channel information. For purposes of classifying different textures, mean gray value is irrelevant. For example if a region has a particular texture then it will still have the same texture if a constant is added to all pixel intensities. Indeed, using the gray value channel information can sometimes result in an incorrect segmentation. To demonstrate this, we tested the same synthetic image but with an increased signal-to-noise ratio, as shown in Figure 6.3.2 (left). Using all channels except the gray value, we recover the correct segmentation (Figure 6.3.2, middle). With the gray value channel added, the result is nonsense (Figure 6.3.2, right).

We also tested a two-dimensional synthetic image where the image used in the previous experiment is combined with its "transpose" (Figure 6.3.3, left and middle diagrams). For both dimensions, the seven masks in Section 6.2 are computed, resulting in fourteen channels in total. We discard the gray-value channels, leaving only twelve channels. Using these twelve channels results in the desired four-region segmentation, (Figure 6.3.3, right) albeit with jagged boundaries. We note a similar phenomenon has been observed with Koepfler's original algorithm (Morel and Solimini, [58], Chapter 5) although no explanation was given. We conjecture that this may have something to do with the fact that more than one channel is contributing important information (which was not the case with the previous experiment). Despite the result of the above experiments in Figures 6.3.2 and 6.3.3, we believe suppressing the gray value channel is somewhat artificial. Clearly, suppressing the

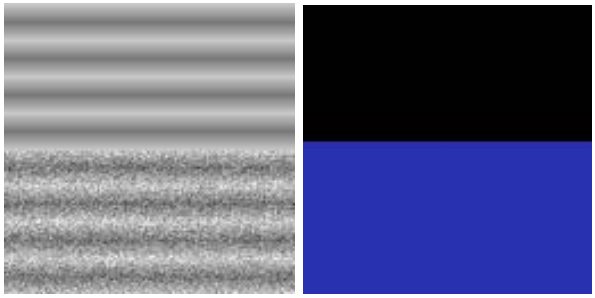


Figure 6.3.1: An image with striped texture with strong noise.

Figure 6.3.2: An image with striped texture with weak noise.

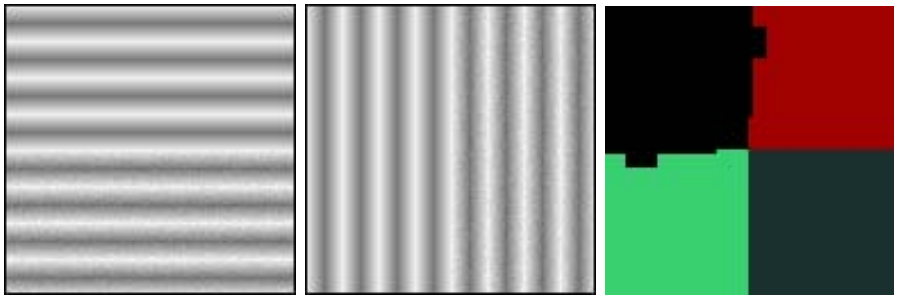


Figure 6.3.3: A two-dimensional image with striped texture with weak noise.

Figure 6.3.4: The Cameraman image.

gray value channel prevents us from segmenting piecewise constant functions correctly. This is inconvenient from a theoretical perspective, since the main motivation for this work is to generalize the Mumford-Shah model. Indeed, Simon Barker [8] correctly pointed out that the choice of which features or channels to measure is ad-hoc and difficult to justify theoretically. Nevertheless, suppressing irrelevant channels is quite useful in practice, since one can obtain better segmentations with less computational burden. For instance if we had *a priori* information that regions differed in terms of textures rather than average gray value, it would not be surprising if suppressing the gray value channel leads to superior results for a particular test suite.

6.3.2 The Cameraman Image

The next experiment was performed on the standard image Cameraman (Figure 6.3.4, left). This is very popular for image compression (Wakin et al., [89], Mertins, [55]) but it is also used for segmentation and denoising algorithms (Chan et al., [20]). We ran the algorithm to obtain a 2-region segmentation using the gray value channel (Figure 6.3.4, middle) only and all seven channels (Figure 6.3.4, right). Using only the gray value information, the man is merged with the background to the right. Using all seven channels, we obtain a reasonable segmentation, comparable with the

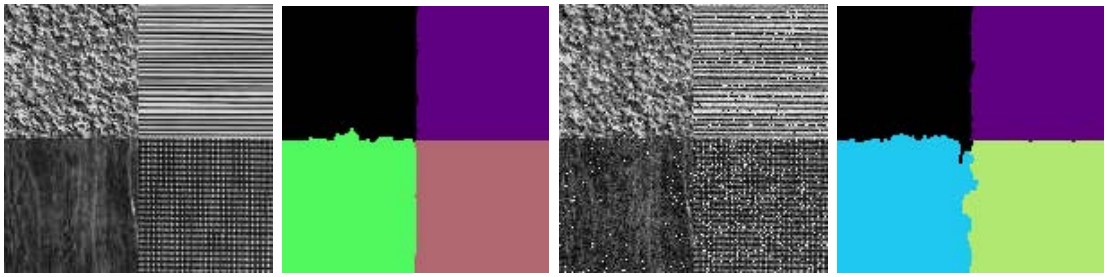


Figure 6.3.5: A Brodatz mosaic.

result in (Chan et al., [20]).

6.3.3 A Brodatz Mosaic

The next experiment was performed on a Brodatz mosaic (Figure 6.3.5). A popular experiment for texture segmentation is to divide a square into a mosaic of subsquares consisting of different Brodatz textures. Indeed a number of such experiments are demonstrated in (Koepfler et al., [42]). We tested a synthetic image consisting of four different Brodatz textures with and without white noise. The textures can be described as noisy (top left), strong horizontal (top right), weak vertical (bottom left) and grid (bottom right). In the case without noise, we recover near perfect boundaries except for the interface between the top left and bottom left regions (Figure 6.3.5, 1st and 2nd diagrams). This can be explained by the rough interface between the two textures. Note that the top right and bottom right textures are much more regular. When noise was added, we were still able to recover the correct regions, albeit with jagged boundaries (Figure 6.3.5, 3rd and 4th diagrams).

6.4 Conclusions

We briefly surveyed algorithms for texture segmentation. Algorithms can be classified as model-based or feature-based. Our new proposed algorithm is feature-based and combines the ideas of modelling white noise contamination and defining simple texture features based on the Haar wavelet transform. Much of these ideas are based on the original work of Koepfler [42]. We pointed out that Koepfler's texture features do not allow one to obtain the correct segmentations for images contaminated by white noise. We found that suppressing the gray-value output sometimes improves the segmentation results obtained in practice, as was the case in Koepfler's original experiment, but we pointed out that this is theoretically doubtful since the ability to segment cartoon images correctly is lost. We propose as future work the possibility of considering other, more complex, texture features (such as those derived from co-occurrence matrices) to enable us to obtain reasonable segmentations of complex images.

Chapter 7

Selection of an Optimal Value of the Scale Parameter for the Extended Piecewise Constant Mumford-Shah Functional

7.1 Introduction

In previous chapters we had considered the following problem: given an image and the value of a scale parameter, find a reasonable segmentation, one that is locally optimal in some sense. We now discuss the important issue of finding the "optimal" value of the scale parameter itself.

The problem of choosing scale values of parameters is pervasive in virtually all energy functionals used in image segmentation. In Chapter 1 we discussed the variational formulation of the image segmentation problem and discussed numerous methodologies (Snakes, Active Contours, Level Sets) proposed in the literature. All of these are based on selection of a model M that minimizes an energy functional of the form

$$E(M) = \sum_{i=1}^{N} \lambda_i T_i \tag{7.1.1}$$

consisting of two or more terms where each λ_i is a scale parameter and each T_i is a term designed to enforce some property of a desired segmentation such as "data fidelity" or "smoothness of boundaries" and so on. Assuming the energy is scaled with $\lambda_1 = 1$, this leaves one or more scale parameters which control the trade-off between the various properties of a segmentation. We note that some authors have studied functionals without scale parameters. For example Kanungo et al. [39] derive an energy functional from the principle of Minimum Description Length to derive the terms T_i. The main disadvantage of this is that the functional becomes very complex and consists of many terms. In the literature, energy functionals are almost always based on equation (7.1.1). We will only consider the piecewise constant Mumford-Shah model which only contains one scale parameter.

The simplest way to choose the scale parameter is to determine the proper values experimentally, running the algorithm a number of times before the user is satisfied with the end result. Alternatively a multiscale hierarchy of segmentations can be computed first. This allows the user to view all segmentations "off-line" after running the algorithm once only. One of the earliest efforts in context of the Mumford-Shah functional is, of course, Koepfler's algorithm [42], which has already been discussed in considerable detail in this thesis. We recall the basic properties of Koepfler's algorithm: starting with a fine segmentation, regions are successively merged until it is no longer possible to decrease an energy functional. The energy functional depends on a scale parameter λ. Given a schedule Λ, that is, a finite list of values for the scale parameter, the algorithm can compute a *hierarchy* $\mathcal{K}$ of segmentations, that is, a finite list of segmentations with each one corresponding to a specific value $\lambda \in \Lambda$. Furthermore $\mathcal{K}$ and Λ satisfy the property of causality. That is, if segmentations K_1, K_2 are computed for λ_1, λ_2, with $\lambda_1 < \lambda_2$ then $K_2 \subset K_1$. In the Full Lambda Schedule Algorithm (FLSA), described in Chapter 2, we show

that it is possible to obtain a full schedule, where $\Lambda = [0, \infty]$, the set of all positive reals, eliminating the need to select $\mathcal{K}$ a-priori. In this case each segmentation in the hierarchy corresponds to an interval $\lambda \in (\lambda_1, \lambda_2]$ and the causality property implies that for $\lambda_1 < \lambda_2$, we have $K_2 \subset K_1$ or $K_2 = K_1$.

In some applications it is desirable to eliminate the need for human interaction altogether. In other words, an algorithm is required to not only construct the multiscale heirarchy of segmentations but also automate the selection of the parameter value. Recently there has been an increasing interest in multiscale approaches to Image Processing problems such as segmentation and restoration. However many algorithms are multiscale in nature, not with respect to the scale parameters λ_i but with respect to something else! For example the algorithm proposed by Chan and Esedoglu [18] is a multiscale generalization of the algorithm by Song and Chan [79]. Both papers seek a local minimizer of the Mumford-Shah functional via the Level Set method. The latter considers the effect of altering the state of a single pixel, but the former considers instead neighbourhoods of pixels of various sizes. Chan and Esedoglu argue that the multiscale nature of their algorithm allows them to remove noise at different scales of the boundary interface. Thus, given a value of scale parameter, they are more likely to produce a reasonable segmentation. But neither paper discusses the selection of the scale parameter itself. To the best of our knowledge, we are unaware of any significant efforts to address the issue of automatic scale parameter selection for the Mumford-Shah model (or any of its variants). Indeed, the selection of scale parameter is a somewhat controversial issue. Some authors (Marr, [48], Petrovic and Vendergheynst, [66]) insist that image segmentation is inherently a multiscale problem and there is no "unique" correct scale on which to analyze an image. We do not entirely agree with this. Although all realistic images have features at different scales, we believe some scales are more important than others. For instance, if one looks at an ebony chessboard, the division of the board into 64 squares is more prominant than the ebony-texture within each square.

Thus the former large-scale segmentation should be ranked higher than the latter small-scale segmentation. But if we were only shown a single square then the texture within that square would stand out more clearly. This brings us to the fundamental idea of this chapter: every segmentation can be given a score which reflects how important it is relative to every other segmentation. The segmentation with the best score out of the hierarchy is chosen as the optimal. Note that the scoring function also allows us to rank, say, the best N segmentations out of a hierarchy, where N is an integer.

Recall that in the standard and extended Mumford-Shah models, the scale parameter λ plays the role of controlling the trade-off between two quantities and there are many well-known techniques for defining an optimal trade-off between two quantities. For instance, in (Tao and Crisp, [82]) we mentioned the possibility of using L-curves. The basic principle is that for any value of scale parameter λ, the solution "size" (complexity) and "error" (accuracy) can be computed, hence a graph of accuracy versus complexity can be drawn, with each point on the graph corresponding to a particular value of λ. Hansen (Hansen, [36]) defines the optimal value of λ as that corresponding to the "corner" of the graph. Ideally, the graph should contain only one obvious corner, resembling the corner of the letter 'L', hence the name "L-curve". A common definition of best corner is one with greatest curvature (Hansen and O'Leary, [37]). In the case of image segmentation, we can associate with each merge a "merge number" (the order in which they occur) and a "merge cost". When merging two regions, the merge cost is the critical value of scale parameter λ that results in no change of the value of the energy functional. The merge cost is given by (2.4.3), from Chapter 2. However, we reported in (Tao and Crisp, [82]) that the experimental results obtained were not satisfactory. Moreover, the use of L-curves is difficult to justify theoretically since it was originally intended for linear problems of the form $Ax = B$ where A is ill-conditioned. This is not really related to the Mumford-Shah functional.

We therefore need a different solution to the problem of ranking each merge. We propose two possibilites in the next subsection: measuring the significance of each merge and modelling the merge cost. We will show that the former, while simpler, is not really satisfactory but the latter yields better results.

7.2 Significance of Merges

Recall that Koepfler's algorithm can be extended by using the FLSA (Chapter 2) or its variants, FLSA-MAP (Chapter 3) and FLSA-CDF (Chapter 5). In this way we can determine a unique segmentation K given an image g and value of scale parameter $\lambda \in \mathbf{R}$. We thus have the hierarchy $=\{K_0, K_1, \ldots K_N\}$ where N is the number of pixels and each K_i corresponds to some interval $(\lambda_i, \lambda_{i+1}]$. From each K_i, the segmentation K_{i+1} can be obtained by a single region merge and the value λ_{i+1} is called the merge cost. Furthermore, the values of λ_i and the full hierarchy $\mathcal{K}$ of segmentations can be efficiently computed. Since the image has N pixels, we have $N-1$ merging operations, or simply "merges", indexed with a merge number i where $i \in \{1, 2, \ldots, N-1\}$. The i-th merge corresponds to changing the $(N-i+1)$-region segmentation into the $(N-i)$-region segmentation. Our problem is to determine the correct value of λ or equivalently, the correct interval $(\lambda_i, \lambda_{i+1}]$ for some $0 \leq i \leq n-1$.

The most obvious idea would be to define the correct segmentation as that corresponding to the largest interval, i.e. where $\lambda_{i+1} - \lambda_i$ is maximum. One difficulty is that, by definition the right segmentation would always be the "blank" one-region segmentation $K = \phi$ since the corresponding interval is (λ_0, ∞) for some $\lambda_0 \in \mathbf{R}$! A possible solution is to define an artificial threshold $T > 0$ and consider only scale parameter values in $(0, T)$ instead of $(0, \infty)$. But this is clearly undesirable, since T behaves as another scale parameter to be determined. Instead, we will simply discount the possibility of obtaining the one-region segmentation. This is no great loss since the 1-region segmentation carries no useful information in any event. Thus

the algorithm is expected to return a segmentation with two or more regions, and it is up to the user to decide if the segmentation carries any useful information. The theory behind this is that if the 1-region segmentation is correct, then there is no real image structure and we would expect all region merges to be equal in significance, to within the usual statistical "chance variations". In other words we would expect a random number of regions to be reported as optimal. Thus a large number of regions imply no significant image structure but a small number of regions probably indicates some important information in the image. Admittedly this is somewhat artificial but we believe the advantages of the automatic selection of scale parameter outweighs the disadvantages.

In our experiments we found that the merge cost increases in the long run with respect to the merge number. However the merge cost does not increase monotonically as demonstrated by the following simple example: suppose one is given two pairs of adjacent regions AB and CD. If merging AB is cheaper than merging CD then the former must be merged before the latter. If however we only have three regions A, B, D all adjacent to each other, one can merge AB to obtain a larger region C and then merge C with D. It is then possible for the merge CD to be cheaper than that for AB. We define the significance of a region merge M as

$$\max(0, \lambda(M) - \sup_{M'} \lambda(M')),$$

where $\lambda(M)$ is the merge cost for the merge M and the supremum is taken over all previous merges M'. This is equivalent to taking the "envelope" of the merge cost graph (the infemum of all monotonically increasing functions not lower than the original graph) and defining the significance as the difference between adjacent elements. If merge M_i is the most significant merge, then the $(N - i + 1)$-region segmentation is correct. By definition, this implies the correct segmentation must have between 2 and N regions.

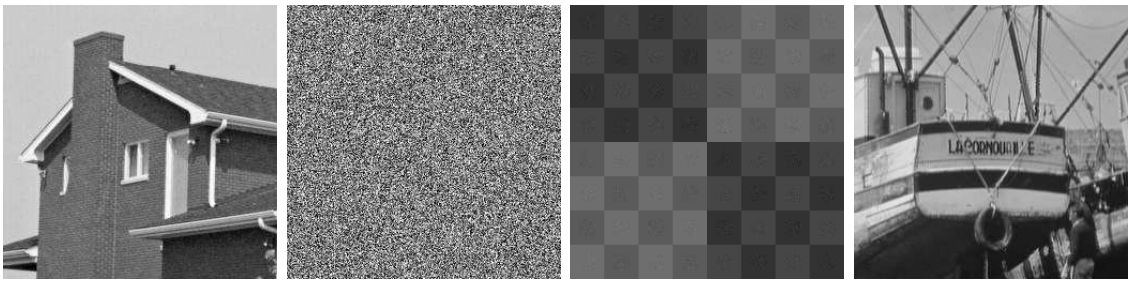


Figure 7.3.1: Four images to be segmented.

7.3 Experimental Results for the Significance of Merges

We ran the FLSA-MAP and FLSA-CDF algorithms and tested four images, called "House", "Gaussian Noise", "Multiscale" and "Boat" (Figure 7.3.1). The House and Boat images are standard in the literature. The Gaussian Noise image is obtained by using pixel values from i.i.d. $\mathcal{N}(128, 64)$ distributions. The Multiscale image is constructed as follows: a 256×256 image is divided into a 2×2 large chessboard whose mean "black-square" and "white-square" gray values differ by 40. Each of these squares is divided into a 4*4 sub-chessboard whose mean black- and white-square gray values differ by 20. Each of the 64 squares is of constant gray value except that a white noise is added to each pixel within a small circle of radius 8, centred at the centre of the square. Thus there are three levels of detail in an image: the division into large squares, division into small squares, and adding white noise to a small circle within each small square.

We tested the FLSA-MAP with offset of $10^{-6}, 10^{-3}, 1$ and FLSA-CDF. We will not be concerned with the proper selection of offset for the FLSA since we argued that the use of FLSA-CDF provides a reasonable solution. For notational convenience we use an offset of zero to denote the use of the FLSA-CDF, as was done in Chapter 5. The results are displayed in Figures 7.3.2-7.3.5, with the graphs for merge cost on the left and the significance of merges on the right (middle and right for Figure 7.3.3).

For convenience of displaying thes results we showed the base-ten logarithm of the merge costs since the smallest and largest non-zero values differ by several orders of magnitude. Note that the curves do not "start" at merge number zero since the earliest merges correspond to a merge cost of $\lambda = 0$ or $\log(\lambda) = -\infty$. In the merge significance graphs we displayed the number of regions on the x-axis, which is equivalent to reversing the merge number on the x-axis. This is because the most significant merges tend to occur with fewer regions. Optimal segmentations obtained using the FLSA-CDF using the significance of merges for the House (2 regions), Gaussian Noise (5 regions), Multiscale (4 regions) and Boat (2 regions) images are shown in Figure 7.3.6.

We considered the obtained results in (Figures 7.3.2-7.3.5) to be unsatisfactory. Firstly, for the Gaussian Noise image, the "optimal" number of regions is less than 30 with almost all pixels belonging to a single region. In other words, both FLSA-MAP and FLSA-CDF incorrectly assume that there is significant structure in the image. Moreover, the number of regions obtained for the Boat image is two. The resulting segmentation is nonsense. A more subtle problem can be discerned through careful consideration of the Multiscale image. It is possible to divide only some of the four quadrants of the image into 16 smaller squares. In other words, the scale of the segmentation differs at various parts of the image. This results in a 19-, 34- or 49- region segmentation, shown in Figure 7.3.7.

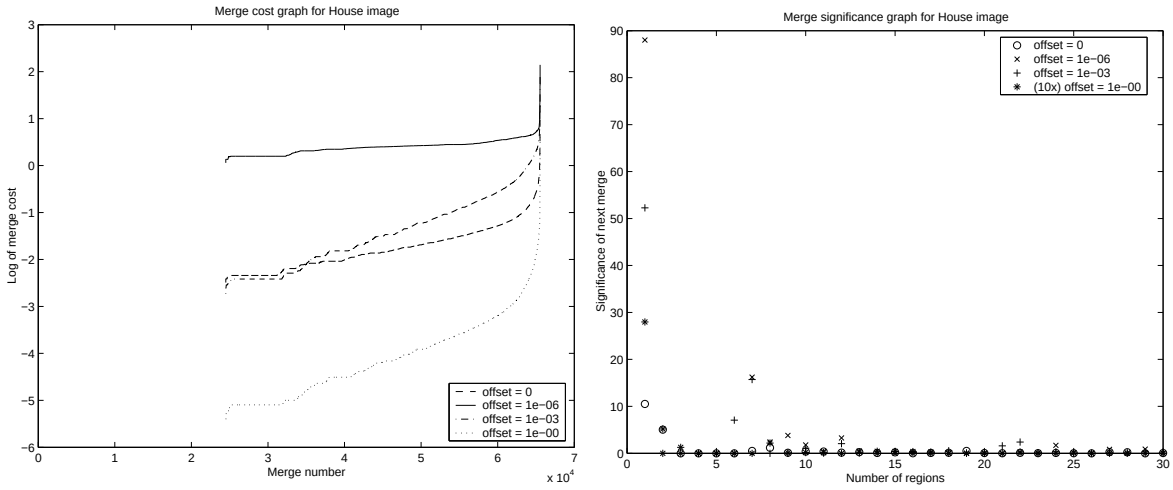


Figure 7.3.2: Graphs for merge cost and significance of merges for the House image.

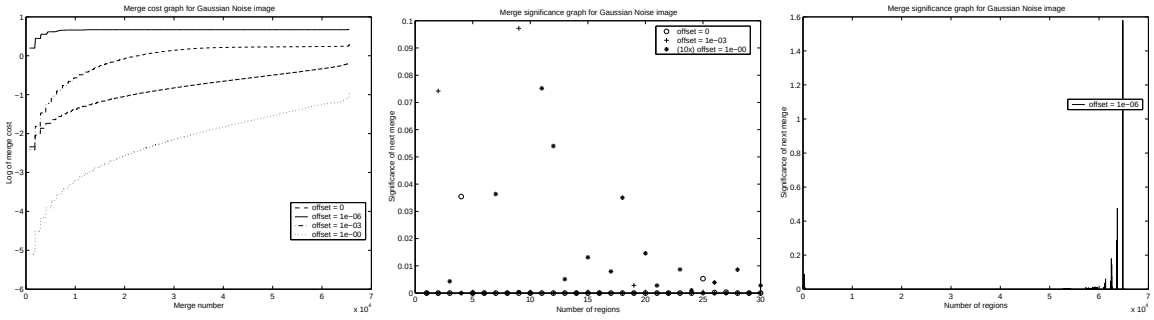


Figure 7.3.3: Graphs for merge cost and significance of merges for Gaussian Noise image.

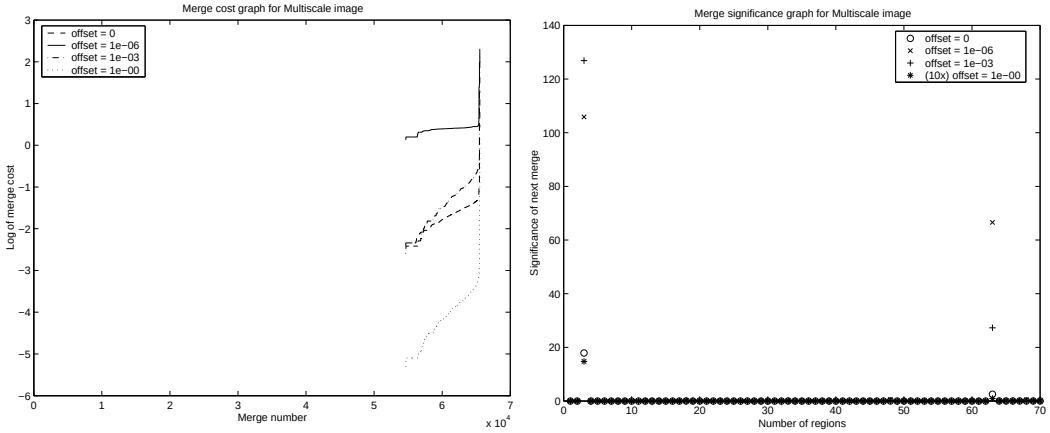


Figure 7.3.4: Graphs for merge cost and significance of merges for the Multiscale image.

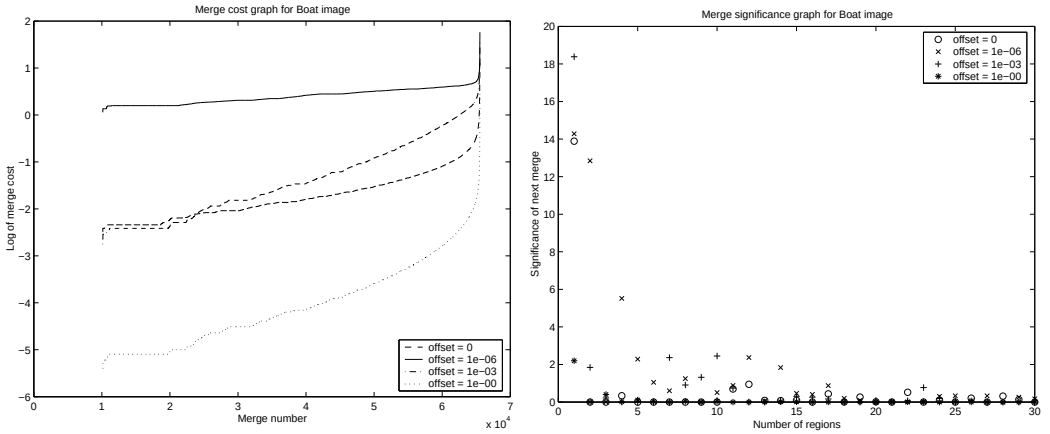


Figure 7.3.5: Graphs for merge cost and significance of merges for the Boat image.

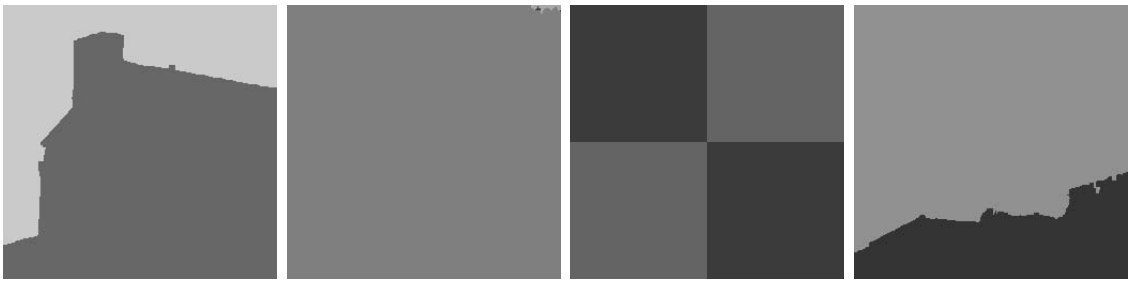


Figure 7.3.6: Optimal segmentations obtained using Significance of merges for the images in Figure 7.3.1.

Figure 7.3.7: Segmentations at different scales for the Multiscale image.

But according to the results these segmentations are not considered significant at all. The results suggest that the largest intervals idea is biased towards segmentations with fewer regions. We believe this is confirmed by the the following observation from our experimental results: the graph of merge cost over merge number grows more than linearly, which accounts for the above bias.

7.4 Modelling the Merge Cost

We need a better heuristic for selecting a correct value of scale parameter. Essentially, we need a more appropriate model of how the merge costs λ_i grow over each iteration. Morel and Solimini [58] state that the number of regions should be proportional to $1/\lambda^2$. This is corroborated by Lemma 5.5 of their proof of the Mumford-Shah conjecture for the Mumford-Shah model which gives an explicit bound for the number of regions α for a 2-normal segmentation:

$$\alpha \leq \frac{288|\Omega|\mathrm{osc}^4(g)}{C_{iso}^2\lambda^2} = \frac{C}{\lambda^2} \tag{7.4.1}$$

for some constant C, where C_{iso} and $\mathrm{osc}(g)$ have the same meaning as in Chapter 3. In the adaptation of this proof for our extended model in Chapter 3 we have a similar inverse square law. Note that equation (7.4.1) alone does not prove that the number of regions behaves like $1/\lambda^2$, since ideally, one would need a bound from both directions. That is,

$$\frac{C_{lower}}{\lambda^2} \leq |R| \leq \frac{C_{upper}}{\lambda^2}.$$

However, Morel and Solimini maintain that their numerical simulations justify the assumption that $|R|$ and λ can be related by an inverse square law. We observe that the inverse square law can be rewritten as

$$\lambda \propto |R|^{-1/2}.$$

This can be interpreted as saying $\lambda/|R|^{-1/2}$ should ideally be constant over time. Obviously we cannot expect a graph of $|R|$ against $\lambda/|R|^{-1/2}$ to be approximately constant for any image. We will not attempt to justify the inverse square law, but simply use it as a heuristic, based on the work of Morel and Solimini.

We therefore propose the following: instead of plotting the values of λ against merge number we plot the ratio $\lambda/|R|^{-1/2}$ against merge number. We denote by ϕ the function that returns the ratio $\lambda/|R|^{-1/2}$ given the merge number. The maximum value of $\lambda/|R|^{-1/2}$ is the most significant "indication of structure" (in some sense) and is therefore the one corresponding to the "optimal" number of regions.

7.5 Experimental Results for Modelling the Merge Cost

We tested the same four images as those used in Section 7.3. For all images, we found that either (i) the optimal number of regions is 64 or less, or (ii) the optimal number of regions is large and the curve is approximately "bell-shaped". The latter case is observed when either segmenting the Gaussian Noise image with any offset or segmenting either the House or Boat image with an offset of 10^{-6}. Note that for the bell-shaped curves, we have suppressed the effect of many small-term fluctuations for convenience of displaying the results. The fluctuations are characterized by the fact that there were many values of R where the ratio $\lambda/R^{-1/2}$ is significantly lower

than the neighbouring values. We plotted only the envelopes of the curves defined by

$$ENV_\phi(x) = \min(\sup_{y \le x} \phi(y), \sup_{y \ge x} \phi(y)). \tag{7.5.1}$$

Roughly speaking, this corresponds to the infimum of all "Λ-shaped" functions not lower than the original graph. Note that for bell-shaped curves, taking the envelope in equation (7.5.1) and multiplying by a constant does not affect where the maximum occurs. The results are shown in Figures 7.5.1-7.5.4 and are much better than those shown for the significance of merges. Note that in some cases, the ratio $\lambda : R^{-1/2}$ differed by an order of magnitude for various offsets so we have multiplied this ratio by 10 for ease of comparison.

Figure 7.5.1 shows the graph of $\lambda/R^{-1/2}$ using 30 regions or less (offset $\neq 10^{-6}$) or the envelope (offset $= 10^{-6}$) for the House image. Figure 7.5.2 shows the graph of the envelope of $\lambda/R^{-1/2}$ for the Gaussian Noise image. Figure 7.5.3 shows the Graph of $\lambda/R^{-1/2}$ using 70 regions or less for the Multiscale image. Figure 7.5.4 shows the Graph of $\lambda/R^{-1/2}$ using 40 regions or less (offset $\neq 10^{-6}$) or envelope (offset $= 10^{-6}$) for the Boat image. Note that in some cases the reason for showing only the graphs for up to a specified number of regions is because the envelope of the graphs is flat for a larger number of regions. For the House image, the FLSA-CDF reported two regions, as did FLSA-MAP with an offset of 10^{-3} or 10^{0}. With an offset of 10^{-6}, FLSA-MAP incorrectly guessed a large number of regions. For the Gaussian Noise image, both FLSA-MAP and FLSA-CDF correctly reported a large number of regions as optimal, indicating there is no structure in the image. In Figure 7.5.5 the obtained segmentation looks almost the same as the original, and one must look closely to determine that small regions of constant gray value indeed exist. For the Multiscale image, FLSA-MAP and FLSA-CDF recognise the 19-, 34- and 49- region segmentations as being significant. FLSA-MAP reported the four-region segmentation as the optimal segmentation except for an offset of 10^{-6}.

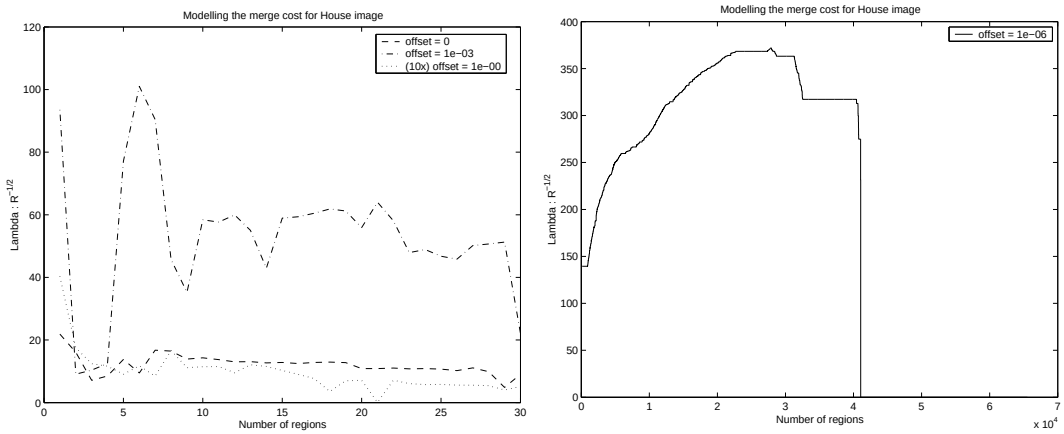


Figure 7.5.1: Graph of $\lambda/R^{-1/2}$ for the House image.

With such an offset, the four-region segmentation is ranked fifth, behind the 19- 34- 49- and 64- region segmentations. For the Boat image, FLSA-CDF reports 12 regions, better than the 2-region segmentation, albeit not perfect. It recognizes a number of "blobby regions" on the lower part of the boat but misses the long lines at the top, which is not surprising since in our proof of the main theorem we showed that long skinny regions do not occur. It is futile to discuss the correct number of regions without a proper ground truth image. However we should mention that Koepfler et al. [42] chose to display the result for 50 and 200 regions. Results for the CDF algorithm using Modelling the merge cost are shown in Figure 7.5.5 for the House(4 regions), Gaussian Noise (14297 regions), Multiscale (4 regions) and Boat (12 regions).

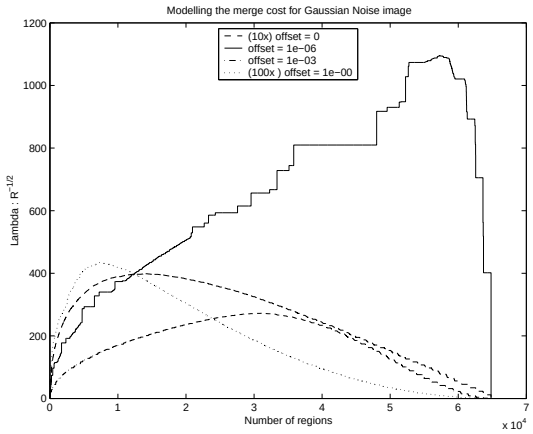


Figure 7.5.2: Graph of $\lambda/R^{-1/2}$ for the Gaussian Noise image.

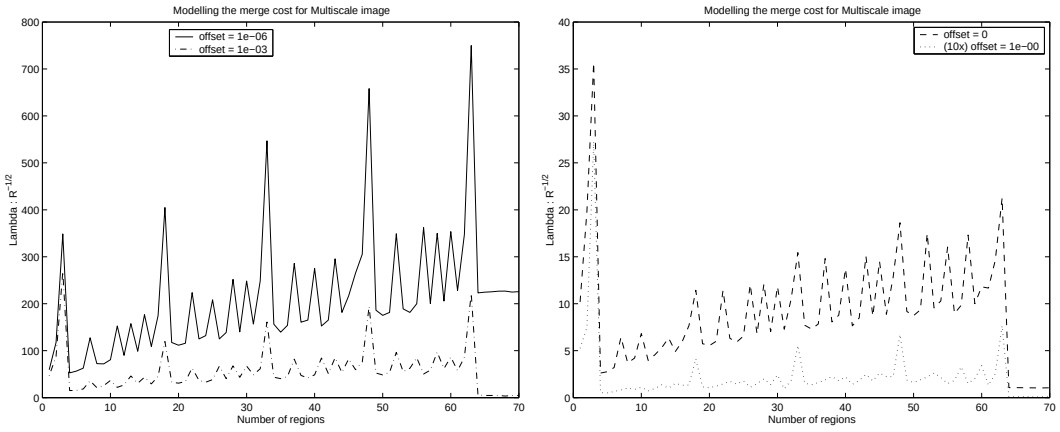


Figure 7.5.3: Graph of $\lambda/R^{-1/2}$ for the Multiscale image.

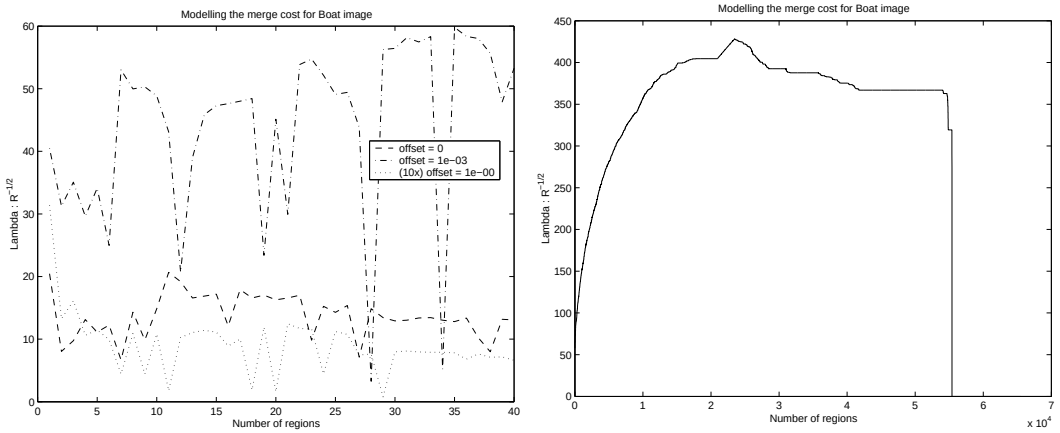


Figure 7.5.4: Graph of $\lambda/R^{-1/2}$ for the Boat image.

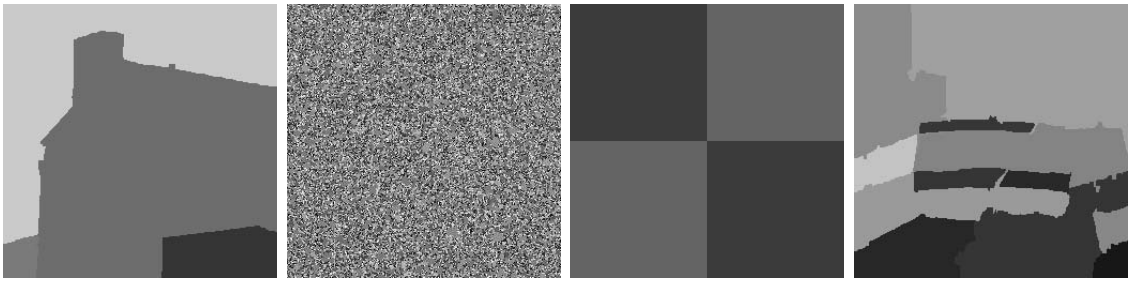


Figure 7.5.5: Optimal segmentations obtained using Modelling the merge cost for the images in Figure 7.3.1.

7.6 Conclusions

To determine a segmentation corresponding to the optimal value of scale parameter λ is an important but difficult problem. Koepfler et al. showed that a relatively simple algorithm can be used to obtain a hierarchy of segmentations satisfying causality, i.e. segmentations corresponding to greater values of λ are a subset of those corresponding to lesser values. However, Koepfler et al. did not state any reasonable solution for the automatic selection of λ. Nor are we aware of any significant efforts in the literature to address this. In Chapter 5 we discussed the idea of using the cumulative distribution function and the corresponding FLSA-CDF. This avoids the use of the regularization parameter σ_0^2 in the FLSA. Our experimental results suggest both FLSA-CDF and FLSA-MAP report the correct number of regions, provided the variance offset is correctly guessed for the latter. We considered two different criteria for determining the most significant merge, namely: (i) largest difference between λ corresponding to a merge and any previous merge and (ii) the maximum value of the ratio $\lambda/|R|^{-1/2}$, justified by Morel and Solimini's consideration that the number of regions should behave as λ^{-2}. The first criterion is simpler, but we considered it unsatisfactory. The second yields better results.

Chapter 8

Conclusions

The central themes of this thesis are an analysis of an extended Mumford-Shah model and the development of a new region-merging algorithm for this model. In Chapter 1 we gave an overview of various algorithms for finding a segmentation of a given image. In Chapter 2 we described the piecewise constant Mumford-Shah model and Koepfler's region merging algorithm in detail. This model allows an efficient representation of simple images whose regions are approximately constant. We also proposed some improvements to Koepfler's original region merging algorithm. We eliminated the requirement of selecting a "Lambda-schedule" prior to region merging, and we also demonstrated that it was possible to find the globally best merge at each stage of the region merging algorithm, instead of just approximating the best merge by considering one region only. This considerably facilitated the theoretical analysis of the algorithm.

However, the piecewise constant Mumford-Shah model is unsuitable for images corrupted by noise or texture. We proposed a new extended model to account for images corrupted by white noise. This was examined in detail in Chapter 3, the central chapter of this thesis. The basic idea was the following: each pixel value in the same region was represented as independent and identically distributed normal random variables. An image model consisted of the partition of the image domain

into regions and the specification of the model parameters within each region. The probability of the data given the model was obtained by multiplying the individual probabilities for each pixel, and the prior probability of the model itself was defined as being proportional to $\exp(-\lambda \cdot \ell(K))$ where $\ell(K)$ is the length of K and λ is the scale parameter. By defining the energy to be the negative logarithm of the probability of the model given the data, and applying Bayes Law, we showed it was possible to convert the problem of maximizing a probability function into that of minimizing an energy functional. A number of important properties of the original Mumford-Shah model were shown to be true also for the extended model. In Chapter 4, we gave an example image g where a unique minimizer could be explicitly computed.

One difficulty we noted with the extended model was that an extra parameter was required to regularize the problem. The difficulty stemmed from the following consideration: a normal random variable with zero variance was an improper distribution, namely a Dirac distribution centred on the mean of the random variable. Thus when taking the negative logarithm of the probability in order to obtain an energy functional, we found the latter was unbounded from below. We therefore added an extra parameter, called the variance offset, to fix this problem. We showed that a theorem by Morel and Solimini, developed for the piecewise constant Mumford-Shah model, was also applicable to the extended model, with minor changes. In Chapter 5, we proposed an alternative solution which did not require an extra parameter: by considering the cumulative distribution function instead of the probability density, we avoided having to work with the Dirac distribution altogether since a variance of zero merely implies a cumulative distribution function which equates to that of a Heaviside step function. We showed that better segmentations were obtained in less time.

The disadvantages of this approach are: (i) a number of approximations are required to quickly calculate the change of energy when merging two regions, (ii) the method is not directly applicable to multiband images, since the calculation of the cumulative distribution function is intractable. However, we proposed for future research the possibility of approximating the cumulative distribution function with a simpler expression.

Following the original paper of Koepfler et al. [42] we argued that with suitable definition of data channels, we could also achieve texture segmentation. We showed that we could segment images corrupted with both texture and noise. This was discussed in Chapter 6. However, the methods employed were rather simple and we hope that investigation of more sophisticated methods could provide a promising avenue of future research. Finally we considered the problem of automatically selecting a stopping value of scale parameter in Chapter 7, an issue somewhat neglected in the literature.

Appendix A

The Distance Between Two Error Functions for the Metric d Defined in Chapter 5

We use the standard result

$$\int_a^b \mathrm{erf}(z)dz \;=\; z\mathrm{erf}z + \frac{e^{-z^2}}{\sqrt{\pi}}\Big|_a^b$$

to verify that if $d(\cdot,\cdot)$ is defined by equation (5.2.3) and if $F^{K_1,\mathbf{x}}, F^{K_2,\mathbf{x}}$ correspond to the cumulative distribution functions (cdfs) of two normal distributions $\mathcal{N}(\theta_1(\mathbf{x})), \mathcal{N}(\theta_2(\mathbf{x}))$ then

$$\begin{aligned} d(F^{K_1,\mathbf{x}}, F^{K_2,\mathbf{x}}) \;=\; & |\mu_2-\mu_1|\mathrm{erf}\Big(\frac{|\mu_2-\mu_1|}{|\sigma_2-\sigma_1|\sqrt{2}}\Big) \\ & + \frac{|\sigma_2-\sigma_1|\sqrt{2}}{\sqrt{\pi}}\exp\Big(-\frac{|\mu_2-\mu_1|^2}{2|\sigma_2-\sigma_1|^2}\Big). \end{aligned} \tag{A.0.1}$$

The proof follows. In it we use the convention that $\mathrm{erf}((z)/0) = \mathrm{sign}(z) = 1$ for positive z, -1 for negative z and 0 for $z = 0$.

The cdf of a normal distribution $\mathcal{N}(\theta) = \mathcal{N}(\mu, \sigma^2)$ is given by

$$\begin{aligned} F_{\mathcal{N}(\theta)}(z) &= \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(- \frac{(y-\mu)^2}{2\sigma^2}\Big) dy \\ &= \int_{-\infty}^{(z-\mu)/\sigma\sqrt{2}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\ &= \frac{1}{2}\int_{-\infty}^{0} \frac{2}{\sqrt{\pi}} e^{-t^2} dt + \frac{1}{2}\int_{0}^{(z-\mu)/\sigma\sqrt{2}} \frac{2}{\sqrt{\pi}} e^{-t^2} dt \\ &= \frac{1}{2} + \frac{1}{2}\mathrm{erf}\Big(\frac{z-\mu}{\sigma\sqrt{2}}\Big). \end{aligned}$$

Since $d(F^{K_1,\mathbf{x}}, F^{K_2,\mathbf{x}}) = d(F^{K_2,\mathbf{x}}, F^{K_1,\mathbf{x}})$ we can assume without loss of generality that $\sigma_1 \leq \sigma_2$. Then

$$\begin{aligned} d(F^{K_1,\mathbf{x}}, F^{K_2,\mathbf{x}}) &= \int_{-\infty}^{\infty} \Big| \frac{1}{2} + \frac{1}{2}\mathrm{erf}\Big(\frac{x-\mu_1}{\sigma_1\sqrt{2}}\Big) - \frac{1}{2} - \frac{1}{2}\mathrm{erf}\Big(\frac{x-\mu_2}{\sigma_2\sqrt{2}}\Big)\Big| dx \\ &= \frac{1}{2}\int_{-\infty}^{\infty} \Big| \mathrm{erf}\Big(\frac{x-\mu_1}{\sigma_1\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{x-\mu_2}{\sigma_2\sqrt{2}}\Big)\Big| dx. \end{aligned}$$

The curves $\mathrm{erf}(x-\mu_1/\sigma_1\sqrt{2}), \mathrm{erf}(x-\mu_2/\sigma_2\sqrt{2})$ must either (i) coincide ($\mu_1 = \mu_2, \sigma_1 = \sigma_2$) or (ii) intersect an infinite number of times ($\mu_1 \neq \mu_2, \sigma_1 = \sigma_2 = 0$) or (iii) are parallel ($\mu_1 \neq \mu_2, \sigma_1 = \sigma_2 \neq 0$) or (iv) intersect exactly once ($\sigma_1 \neq \sigma_2$). In the first two cases equation (A.0.1) is trivial so we assume either (iii) or (iv). Let K be the point of intersection of the two curves, where $K = \infty$ in case (iii) and $K < \infty$ in case (iv). K will be the value where

$$\frac{1}{2} + \frac{1}{2}\mathrm{erf}\Big(\frac{K-\mu_1}{\sigma_1\sqrt{2}}\Big) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\Big(\frac{K-\mu_2}{\sigma_2\sqrt{2}}\Big)$$

which reduces to

$$K = \frac{\mu_1\sigma_2 - \mu_2\sigma_1}{\sigma_2 - \sigma_1}.$$

Therefore

$$\begin{aligned} d(F^{K_1,\mathbf{x}}, F^{K_2,\mathbf{x}}) &= \pm\frac{1}{2}\int_{-\infty}^{K} \operatorname{erf}\Big(\frac{x-\mu_2}{\sigma_2\sqrt{2}}\Big) - \operatorname{erf}\Big(\frac{x-\mu_1}{\sigma_1\sqrt{2}}\Big)dx \\ &\quad \pm\frac{1}{2}\int_{K}^{\infty} \operatorname{erf}\Big(\frac{x-\mu_2}{\sigma_2\sqrt{2}}\Big) - \operatorname{erf}\Big(\frac{x-\mu_1}{\sigma_1\sqrt{2}}\Big)dx \qquad \text{(A.0.2)} \end{aligned}$$

with the above-mentioned value for K. We must determine the plus/minus signs via comparison of $\operatorname{erf}((x-\mu_1)/\sigma_1\sqrt{2})$ and $\operatorname{erf}((x-\mu_2)/\sigma_2\sqrt{2})$. Since $\sigma_1 \leq \sigma_2$ we have that $x < K$ implies the graph of the first cdf must be below that of the second, and vice versa for $x > K$ which implies that (A.0.2) reduces to

$$\begin{aligned} d(F^{K_1,\mathbf{x}}, F^{K_2,\mathbf{x}}) &= \frac{1}{2}\int_{-\infty}^{K} \operatorname{erf}\Big(\frac{y-\mu_2}{\sigma_2\sqrt{2}}\Big) - \operatorname{erf}\Big(\frac{y-\mu_1}{\sigma_1\sqrt{2}}\Big)dy \\ &\quad -\frac{1}{2}\int_{K}^{\infty} \operatorname{erf}\Big(\frac{y-\mu_2}{\sigma_2\sqrt{2}}\Big) - \operatorname{erf}\Big(\frac{y-\mu_1}{\sigma_1\sqrt{2}}\Big)dy. \qquad \text{(A.0.3)} \end{aligned}$$

For notational convenience we also introduce

$$L = \frac{K-\mu_1}{\sigma_1\sqrt{2}} = \frac{K-\mu_2}{\sigma_2\sqrt{2}} = \frac{\mu_1-\mu_2}{\sqrt{2}(\sigma_2-\sigma_1)}.$$

Now we calculate

$$
\begin{aligned}
&\int_K^\infty \mathrm{erf}\Big(\frac{y-\mu_1}{\sigma_1\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{y-\mu_2}{\sigma_2\sqrt{2}}\Big)dy \\
=\ & \lim_{T\to\infty}\int_K^T \mathrm{erf}\Big(\frac{y-\mu_1}{\sigma_1\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{y-\mu_2}{\sigma_2\sqrt{2}}\Big)dy \\
=\ & \lim_{T\to\infty}\int_L^{(T-\mu_1)/\sigma_1\sqrt{2}} (\sigma_1\sqrt{2})\mathrm{erf}(y)dy \\
& -\int_L^{(T-\mu_2)/\sigma_2\sqrt{2}} (\sigma_2\sqrt{2})\mathrm{erf}(y)dy \\
=\ & \lim_{T\to\infty}\sigma_1\sqrt{2}\Big[y\mathrm{erf}y + \frac{e^{-y^2}}{\sqrt{\pi}}\Big]_L^{(T-\mu_1)/\sigma_1\sqrt{2}} \\
& -\sigma_2\sqrt{2}\Big[y\mathrm{erf}y + \frac{e^{-y^2}}{\sqrt{\pi}}\Big]_L^{(T-\mu_2)/\sigma_2\sqrt{2}} \\
=\ & \lim_{T\to\infty}(T-\mu_1)\mathrm{erf}\Big(\frac{T-\mu_1}{\sigma_1\sqrt{2}}\Big) - (T-\mu_2)\mathrm{erf}\Big(\frac{T-\mu_2}{\sigma_2\sqrt{2}}\Big) \\
& +(\sigma_2-\sigma_1)\sqrt{2}\Big(L\mathrm{erf}L + \frac{e^{-L^2}}{\sqrt{\pi}}\Big) \\
=\ & (\mu_2-\mu_1)\Big(1+\mathrm{erf}\Big(\frac{\mu_2-\mu_1}{(\sigma_2-\sigma_1)\sqrt{2}}\Big)\Big) \\
& +\frac{(\sigma_2-\sigma_1)\sqrt{2}}{\sqrt{\pi}}\exp\Big(-\frac{(\mu_2-\mu_1)^2}{2(\sigma_2-\sigma_1)^2}\Big). \qquad \text{(A.0.4)}
\end{aligned}
$$

A similar calculation gives

$$
\begin{aligned}
&\int_{-\infty}^K \mathrm{erf}\Big(\frac{y-\mu_2}{\sigma_2\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{y-\mu_1}{\sigma_1\sqrt{2}}\Big)dy \\
=\ & -\int_{-K}^\infty \mathrm{erf}\Big(\frac{-y-\mu_2}{\sigma_2\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{-y-\mu_1}{\sigma_1\sqrt{2}}\Big)dy \\
=\ & -\int_{-K}^\infty \mathrm{erf}\Big(\frac{y-(-\mu_1)}{\sigma_1\sqrt{2}}\Big) - \mathrm{erf}\Big(\frac{y-(-\mu_2)}{\sigma_2\sqrt{2}}\Big)dy \\
=\ & (\mu_1-\mu_2)\Big(1+\mathrm{erf}\Big(\frac{\mu_1-\mu_2}{(\sigma_2-\sigma_1)\sqrt{2}}\Big)\Big) \\
& +\frac{(\sigma_2-\sigma_1)\sqrt{2}}{\sqrt{\pi}}\exp\Big(-\frac{(\mu_1-\mu_2)^2}{2(\sigma_2-\sigma_1)^2}\Big). \qquad \text{(A.0.5)}
\end{aligned}
$$

Therefore (A.0.1) follows from (A.0.3), (A.0.4) and (A.0.5).

Bibliography

[1] D. Adalsteinsson and J. Sethian. A fast level set method for propagating interfaces. *J. Comp. Phys.*, 118:269–277, 1995.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems.* Oxford University Press, 2000.

[3] L. Ambrosio and V. Tortorelli. Approximation of functionals depending on jumps by elliptic functionals via gamma convergence. *Commun. Pure Appl. Math.*, 43:999–1036, 1990.

[4] L. Ambrosio and V. Tortorelli. On the approximation of free discontinuity problems. *Boll. Un. Mat. Ital*, 6-B:105–123, 1992.

[5] V. V. Anh, Q. Tieng, D. Bui, and G. Chen. The Hellinger-Kakutani metric for pattern recognition. In *Proceedings of IEEE ICIP*, 1997.

[6] G. Aubert and P. Kornprobst. *Mathematical problems in image processing: partial differential equations and the calculus of variations.* Springer-Verlag, 2002.

[7] R. Bajcsy and M. Tavakoli. Computer recognition of roads from satellite pictures. *IEEE Trans. Syst., Man, Cybern*, SMC-6(9):612–637, 1976.

[8] Simon Barker. *Image segmentation using Markov random field models.* PhD thesis, University of Cambridge, July 1988.

[9] J. Besag. Spatial interaction and the statistical analysis of lattices. *J. Royal Statist. Soc.*, B36:192–236, 1974.

[10] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statist. Soc*, B55:25–37, 1986.

[11] P. Billingsley. *Convergence of probability measures.* John Wiley and Sons, New York, 1968.

[12] A. Blake and A. Zisserman. *Visual reconstruction.* MIT Press, Cambridge M.A., 1987.

[13] G. Box and G. Tiao. *Bayesian inference in statistical analysis.* Addison-Wesley Pub. Co., Reading, Mass., 1973.

[14] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. In *IEEE Int. Conf. on computer vision*, Boston, USA, 1995.

[15] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22:61–79, 1997.

[16] A. Chambolle. Image segmentation by variational methods: Mumford-Shah functional and the discrete approximations. *SIAM Appl. Math*, 55(3):827–863, 1995.

[17] A. Chambolle and G. Dal Maso. Discrete approximation of the Mumford-Shah functional in dimension two. *M2AN*, 33(4):651–672, 1999.

[18] T. Chan and S. Esedoglu. A multiscale algorithm for Mumford-Shah image segmentation. Technical report, UCLA, December 2003.

[19] T. Chan and S. Esedoglu. Aspects of total variation regularized l^1 function approximation. Technical report, UCLA Mathematics Department, 2004.

[20] T. Chan, S. Esedoglu, and M. Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. Technical report, UCLA, 2004.

[21] T. Chan and L. Vese. An active contour model without edges. In *Int. Conf. on scale space theories in computer vision*, pages 141–151, 1999.

[22] D. Chandler. *Introduction to modern statistical mechanics.* Oxford University Press, 1987.

[23] D. Chopp. Computing minimal surfaces via level set curvature flow. *J. of Comp. Phys.*, 106:77–91, 1993.

[24] C. Coombs, R. Dawes, and A. Tversky. *Mathematical psychology - an elementary introduction.* Prentice Hall, Englewood Cliffs, New Jersey, 1970.

[25] D. Cremers, C. Schnorr, and J. Weickert. Diffusion-snakes combining statistical shape knowledge and image information in a variational framework. In *IEEE Workshop on Variational and Level set methods*, Vanconver, July 2001.

[26] D. Crisp and G. Newsam. A fast efficient segmentation algorithm based on region merging. In *Proceedings of the IVCNZ*, pages 180–185, November 2000.

[27] D. Crisp and T. Tao. Fast region merging algorithms for image segmentation. In *Proceedings of the Fifth Asian Conference on Computer Vision*, pages 412–417, Melbourne Australia, 2002.

[28] L. Davis, M. Clearman, and J. Aggarwal. An empirical evaluation of generalized co-occurrence matrices. *IEEE Trans. PAMI*, 3(2):214–221, 1981.

[29] L. Davis, S. Johns, and J. Aggarwal. Texture analysis using generalized co-occurence matrices. *IEEE Trans. PAMI*, 1(3):251–259, 1979.

[30] H. Derin and H. Elliott. Modelling and segmentation of noisy and textured images. *IEEE Trans PAMI*, 1987.

[31] R. Dubes and A. Jain. Random field models in image analysis. *Journal of Appl. Statistics*, 16(2):131–164, 1989.

[32] L. Evans and R. Gariepy. *Measure theory and fine properties of functions.* CRC Press, 1992.

[33] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. PAMI*, 1984.

[34] E. De Georgi. Gamma-convergenza e g-convergenza. *Boll. Un. Mat. Ital*, 5(14-A):213–220, 1977.

[35] G. Giraldi, L. Goncalves, and A. Oliveira. Dual topologically adaptable snakes. In *Fifth Joint Conference on Information Sciences*, volume 2, pages 103–106, Feb 2002.

[36] P. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM review*, 34:561–580, 1993.

[37] P. Hansen and D. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14:1487–1503, 1993.

[38] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. Systems Man. Cybernet.*, 3(1):610–621, 1973.

[39] T. Kanungo, B. Dom, W. Niblack, D.Steele, and J.Cheinvald. MDL based multi-band image segmentation using a fast region merging scheme. Technical report, IBM Research Report RJ-9960 (87919), 1995.

[40] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. J. of Comp. Vision*, 1988.

[41] S. Kichenassamy, A. Kumar, P. Olver, A. Tannenbaum, and A. Yezzi. Gradient flows and geometric active contour models. In *IEEE Int. Conference on Computer Vision*, pages 810–815, Boston, USA, 1995.

[42] G. Koepfler, C. Lopez, and J. Morel. A multiscale algorithm for image segmentation by variational method. *SIAM. J. Numer. Anal*, 1994.

[43] K. Laws. Textured image segmentation. Technical Report USCIPI 940, Dept. of Elec. Eng. Image Processing Institute, Univ. of Southern California, Los Angeles, 1980.

[44] Y. Leclerc. Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, 3:73–102, 1989.

[45] T. Lee. A minimum description length-based image segmentation procedure and its comparison with a cross-validation-based segmentation procedure. *Jnl. American Statistical Association*, 2000.

[46] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation as the search or the best description of images in terms of primitives. Technical Report MS-CIS-90, University of Pennsylvania, 1990.

[47] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *J. of the Opt. Soc. of America*, 7(5):923–932, 1991.

[48] D. Marr. *Vision.* Freeman and Co., 1982.

[49] A. Martelli. Edge detection using heuristic search methods. *CGIP*, 1:169–182, 1972.

[50] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[51] G. Dal Maso. *An introduction to gamma convergence.* Birkhäuser, Boston, 1993.

[52] T. McInerney. *Topologically adaptable deformable models for medical image analysis*. PhD thesis, University of Toronto, 1997.

[53] T. McInerney and D.Terzopoulos. Topologically adaptable snakes. In *Fifth Int. Conf. on Computer Vision*, pages 840–845, Cambridge, MA, USA, June 1995.

[54] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis: a survey. *Medical Image Analysis*, 1998.

[55] A. Mertins. Image compression via edge-based wavelet transform. *Optical Engineering*, 38:991–1000, 1999.

[56] J. Morel and S. Solimini. Segmentation of images by variational methods: a constructive approach. *Revista Matematica de la Universidad Complutense de Madrid*, 1:169–182, 1988.

[57] J. Morel and S. Solimini. Segmentation d' images par méthods variationnelle: une preuve constructive d'existence. *C.R. Academie de Science de Paris Serie I*, 308:465–470, 1989.

[58] J. Morel and S. Solimini. *Variational methods in image segmentation with seven image processing experiments*. Birkhäuser, 1994.

[59] J. Morel and S. Solimini. *Variational methods in image segmentation*. Birkhäuser, 1995.

[60] J. Muerle and D. Allen. Experimental evaluation of techniques for automatic segmentation of objects in a complex scene. In G. Cheng et al., editor, *Pictorial Pattern Recognotion*, pages 3–13, Thompson, Washington, 1968.

[61] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 1989.

[62] N. Nordstrom. *Variational edge detection*. PhD thesis, Univ. of California, 1990.

[63] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Jnl. of Computational Physics*, 79:12–49, 1988.

[64] T. Pavlidis and Y. Liow. Segmentation of pictures and maps through functional approximation. *Comp.Gr and Im.Proc*, 1:360–372, 1972.

[65] W. Perkins. Area segmentation of images using edges. *IEEE PAMI*, 2(1), Jan 1980.

[66] A. Petrovic and P. Vandergheynst. Multiscale variational approach to simultaneous image regularization and segmentation. In *Proceedings of ISPA*, Rome, Italy, September 2003.

[67] N. Redding, D. Crisp, D. Tang, and G. Newsam. An efficient algorithm for Mumford-Shah segmentation and its application to SAR imagery. In *Digital Image Computing: Techniques and Applications*, pages 35–41, 1999.

[68] T. Reed and J. Hans Du Buf. A review of recent texture segmentation and feature extraction techniques. *CVGIP: Image understanding*, 57(3):359–372, May 1993.

[69] T. Richardson. *Scale independent piecewise smooth segmentation of images via variational methods.* PhD thesis, MIT, Cambridge MA, Feb 1990.

[70] T Richardson and S Mitter. Approximation, computation and distortion in the variational formulation. *Geometry-driven diffusion in computer vision*, 1994.

[71] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[72] D. Robinson, N. Redding, and D. Crisp. Implementation of a fast algorithm for segmenting SAR imagery. Technical report, DSTO, 2002.

[73] A. Rosenfeld and C. Kak. *Digital picture processing*, volume 1. Academic Press, New York, 1982.

[74] A. Rosenfeld and C. Kak. *Digital picture processing*, volume 2. Academic Press, New York, 1982.

[75] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[76] C. Samson, L. Blanc-Feraud, G. Aubert, and J. Zerubia. A level set model for image classification. In *Int. Conf. on scale space theories in computer vision*, pages 306–317, 1999.

[77] J. Sethian. *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision and materials science*. Cambridge University Press, 1999.

[78] J. Soares, C. Renno, A. Formaggio, C. Yanasse, and A. Frery. An investigation of the selection of texture features for crop discrimination using SAR imagery. *Remote Sensing of Environment*, 59(2):234–247, 1997.

[79] B. Song and T. Chan. A fast algorithm for level set based optimization. Technical report, UCLA, 2002.

[80] D. Strong and T. Chan. Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Problems*, 19:S165–S187, 2003.

[81] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Trans. Systems Man. Cybernet.*, 8(6):460–473, 1978.

[82] T. Tao and D. Crisp. A useful bound for region merging algorithms in a Bayesian model. In *ACCV*, pages 95–100, Adelaide, Australia, 2003.

[83] T. Tao, D. Crisp, and J. van der Hoek. Mathematical analysis of an extended Mumford-Shah model for image segmentation. To appear in the Journal of Mathematical Imaging and Vision.

[84] Y. Tong. *The multivariate normal distribution.* Springer-Verlag, New York, 1990.

[85] M. Unser. Local linear transforms for texture measurements. *Signal. Process.*, 11:61–79, 1986.

[86] M. Unser. Sum and difference histograms for texture classification. *IEEE Trans. PAMI*, 8(1):118–125, 1986.

[87] P. E. Utgoff and J. A. Clouse. Decision tree induction based on efficient tree restructuring. Technical Report 95-18, Univ. of Mass., Amherst, 1995.

[88] H. Voorhees and T. Poggio. Computing texture boundaries from images. *Nature*, 333, 1988.

[89] M. Wakin, J. Romberg, H. Choi, and R. Baraniuk. Rate-distortion optimized image compression using wedgelets. In *IEEE International Conference on Image Processing*, September 2002.

[90] R. Weiss and M. Boldt. Geometric grouping applied to straight lines. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1986.

[91] R. Welch, K. Kuo, and S. Sengupta. Cloud and surface textural features in polar regions. *IEEE Trans. Geosc. Remote Sens.*, 28(4):520–528, 1990.

[92] A. Yezzi, A. Tsai, and A. Willsky. A statistical approach to snakes for bimodal and trimodal imagery. In *IEEE Int. Conf. on computer vision*, pages 898–903, Corfu, Greece, 1999.

[93] Y. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–46, 1996.

[94] Y. Zhang. A review of recent evaluation methods for image segmentation. In *Int. Symposium on signal processing and its applications*, pages 148–151, Malaysia, 13–16 Aug 2001.

[95] S. Zhu and A. Yuille. A unified theory for image segmentation: region competition and its analysis. Technical Report 95-7, Harvard Robotics Laboratory, 1995.

[96] S. Zhu and A. Yuille. Region competition: unifying snakes, region growing and Bayes/MDL for multiband image segmentation. *IEEE Trans PAMI*, 1996.

[97] S. Zucker. Region growing: childhood and adolescence. *Comp. Graphics and Image Proc.*, 5:382–399, 1976.