Journal of Mathematical Imaging and Vision 23: 67–86, 2005

Globally Optimal Geodesic Active Contours

BEN APPLETON
Intelligent Real-Time Imaging and Sensing Group, ITEE, The University of Queensland, Brisbane, QLD 4072, Australia
appleton@itee.uq.edu.au

HUGUES TALBOT
CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW 1670, Australia
hugues.talbot@csiro.au

First online version published in July, 2005

Abstract. An approach to optimal object segmentation in the geodesic active contour framework is presented with application to automated image segmentation. The new segmentation scheme seeks the geodesic active contour of globally minimal energy under the sole restriction that it contains a specified internal point p_{int}. This internal point selects the object of interest and may be used as the only input parameter to yield a highly automated segmentation scheme. The image to be segmented is represented as a Riemannian space S with an associated metric induced by the image. The metric is an isotropic and decreasing function of the local image gradient at each point in the image, encoding the local homogeneity of image features. Optimal segmentations are then the closed geodesics which partition the object from the background with minimal similarity across the partitioning. An efficient algorithm is presented for the computation of globally optimal segmentations and applied to cell microscopy, x-ray, magnetic resonance and cDNA microarray images.

Keywords: Geodesic active contour, circular shortest path, optimal segmentation

1. Introduction

Segmentation is a fundamental problem in image analysis. Given a planar image $\mathbb{R}^2 \rightarrow \mathbb{R}^n$ the objective of segmentation is to partition the image into regions that are homogeneous according to some measure, e.g. in object segmentation, the object itself and the background. While many image segmentation techniques have been proposed, a consistent series of approaches have focused on segmentation using edge integration by energy minimisation. In this series a recent notable development is the Geodesic Active Contour (GAC) [10]. In this paper we propose a new algorithm for globally optimal segmentation in planar images based on the GAC energy model.

1.1. Previous Work

Active contours were initially introduced in the form of snakes by Kass et al. [21], and have been widely studied [12, 23]. Kass' classic snake approach simulated the motion of a series of point masses connected by springs and thin plates, with passive forces drawing them toward image features such as edges. This Lagrangian view of curves as splines was replaced by an Eulerian view of curves as level sets in Osher and Sethian's work on front propagation [26], demonstrating topological independence and greater stability. The Osher-Sethian level set method was later refined into the more efficient narrow band scheme presented by Adalsteinsson and Sethian [1]. Level sets have been

applied to active contours for image segmentation by Caselles et al. [9] and Malladi et al. [25]. The Geodesic Active Contour (GAC) was put forward by Caselles et al. [10] as a simplification of the snake energy model with fewer parameters and less sensitivity to the initial contour. They also demonstrated the equivalence between GAC and Riemannian geodesics. The segmentation technique presented in this paper is based on geodesics in a Riemannian space with a metric induced by the image content. Here simple closed curves of minimal energy are used to partition a specific object from the remainder of the image.

Typically in Active Contours a variational framework is used to obtain locally minimal contours by steepest descent of an energy functional. Caselles et al. [10] used a direct first-order time scheme for the level set partial differential equation. The timestep in this scheme was dictated by stability requirements. Goldenberg et al. [20] adapted the Additive Operator Splitting scheme of Weickert et al. [31] to obtain an unconditionally stable update for the narrow band level set method, greatly increasing the speed of the segmentation. However variational descent methods are prone to becoming stuck in local minima caused by noise or irrelevant objects. Methods have also been proposed to reduce the attraction of local minima by modifying the curve flow, such as pressure forces [14], distance forces [15], and Gradient Vector Flow (GVF) snakes [33]. Although these heuristic methods offer improved robustness they remain sensitive to initialisation.

A different approach championed by Cohen and Kimmel [17] is to directly formulate segmentation as a search for minimal geodesics in a Riemannian space. The search for open geodesics may be very efficiently performed using Sethian's Fast Marching Method [27], an optimal-time Eikonal equation solver which grew out of work on narrow band level set methods. Cohen and Kimmel applied the fast marching method to obtain minimal paths for GAC, however their formulation only locates the boundary of an object as the open geodesic connecting two user-specified points. Cohen later extended their work [16] to detecting open and closed contours connecting a number of seed points in an image. In this paper we take a similar approach, using the fast marching method combined with a branch and bound search for connected endpoints to obtain the minimum closed geodesic containing an internal point.

A related area of research formulates object segmentation as the search for a path of minimal energy across a polar transform of the object. Numerous approaches have been researched. Denzler and Niemann [18] presented a gradient descent approach to polar active contours for real-time object tracking. Along with Chen et al. [13] they also give a dynamic programming algorithm to obtain optimal open curves across a radial trellis. Additionally, heuristics have been suggested to obtain 'near-optimal' closed curves. Bamford and Lovell [7] developed an approximate two-pass dynamic programming scheme to ensure continuity of endpoints. The first pass produced a number of candidate endpoints, the best of which was used to compute a closed curve in the second pass. Sun and Pallottino [30] and Appleton and Sun [5] have recently addressed the computation of minimal closed paths in the framework of Circular Shortest Paths. Two main algorithms for CSPs have emerged: the Multiple Backtracking Algorithm/Image Patching Algorithm (MBTA/IPA) hybrid was proposed in [30] for the fast computation of near-optimal closed paths, while the Branch and Bound CSP (BBCSP) algorithm was proposed in [5] for the somewhat slower computation of optimal closed paths. Unfortunately in most cases these polar segmentations are restricted to point-convex objects, limiting their application to only the simplest of object shapes.

1.2. Finding Contours of Globally Minimum Energy

The difficulty in obtaining an optimal segmentation contour with respect to a given metric is threefold. Firstly, the number of contours which must be considered grows exponentially with the image resolution. Secondly, while obtaining an optimal path between two given points is relatively straightforward it is difficult to obtain optimal closed contours. Finally, in the translation to a discrete grid many methods suffer a metrication error resulting in anisotropic segmentations.

In our approach, we represent the image to be segmented as the space $\mathbb{R}^2$ with an associated metric. Typically the metric is an increasing scalar function of the local image gradient at each point in the image. Thus the metric encodes the local homogeneity of the image. The object to be segmented is specified by a single internal point which must be contained inside the segmentation contour. Following Jordan, a simple close curve containing the internal point will then partition the object from the remainder of the image.

Our method produces the simple closed curve of minimal energy or length with respect to the metric, under the restriction that the geodesic must contain a specified internal point. We show that this restriction is a natural requirement to obtain a non-trivial global minimum and is also a natural alternative to various ad-hoc "balloon" forces proposed by other authors, forces that prevent contours from collapsing upon themselves.

Our method is not restricted to convex or point-convex curves. Concave curves are represented in a spiral space identical to the Riemann surface for the logarithm relation. Furthermore our method avoids various metrication errors due to the image grid, and because of its simple initialization (a single point) and lack of parameters (only the metric itself decides the final result) it is highly suitable for automated applications. Thanks to a combination of highly efficient PDE solving and discrete algorithms, our method is itself fast and efficient.

The rest of the paper is structured as follows. In Section 2 we review the basic theory of Geodesic Active Contours, the computation of geodesics by the Fast Marching Method, and Circular Shortest Paths. Section 3 presents the motivation for the proposed method. This motivation inspires the simple approach to segmentation given in this section. In Section 4 a more sophisticated approach to segmentation under the GAC energy model is presented, dealing with the issues of efficiency and optimality. Section 5 discusses the choice of metric and the backtracking scheme to obtain geodesics on a discrete grid. A new metric is proposed and related to existing segmentation schemes utilising polar transforms. Section 7 presents results comparing the new method to the classic Geodesic Active Contour framework. Section 8 demonstrates the properties of the proposed approach. Section 9 concludes.

2. Theoretical Background

In this section we review in more detail geodesic active contours, computing geodesics via the fast marching method, and circular shortest paths, which provide the background to our proposed method.

2.1. Geodesic Active Contours

Geodesic active contours were introduced by Caselles et al. [10, 11] for segmentation in 2D and 3D images. In the planar case they are contours which minimise the integral

$$E[C] = \int_C g(C(s))\,\mathrm{d}s \tag{1}$$

$E[C]$ will be referred to as the energy of the contour $C : [0, L[C]] \to \mathbb{R}^2$, with $L[C]$ the Euclidean length of C. In object segmentation $g : \mathbb{R}^2 \to \mathbb{R}^+$ is a decreasing function of edge strength, typically of the form

$$g = \frac{1}{1 + |\nabla G_\sigma \star I|^p} + \varepsilon \tag{2}$$

where $p = 1$ or 2. $\varepsilon > 0$ is an arc-length penalty term, as if we let $g = \widetilde{g} + \varepsilon$ we obtain

$$E(C) = \int_C \widetilde{g}(C(s))\mathrm{d}s + \varepsilon L(C)$$

Cohen and Kimmel [17] demonstrated the role ε plays in the regularisation of the contour C. The Euler-Lagrange of Eq. (1) is

$$g\kappa = \langle \nabla g, \overrightarrow{N} \rangle$$

hence

$$|\kappa| \leq \frac{|\nabla \widetilde{g}|}{\varepsilon}$$

with $g \geq \varepsilon$ and κ the curvature of C. This places a bound on the maximum curvature of the contour inversely proportional to ε.

2.2. Fast Marching and Geodesics

The length E of a path P in a Riemannian space with heterogeneous metric tensor $\gamma : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ is given by

$$E[P] = \int_P \sqrt{\gamma(P(s))_{ij} dx^i dx^j}$$

where repeated indices imply summation.

For an isotropic metric g this simplifies to

$$E[P] = \int_P g(P(s))\mathrm{d}s \tag{3}$$

So we observe that the geodesic active contour energy (Eq. (1)) is equivalent to the length of a curve in a Riemannian space with an isotropic and heterogeneous metric.

In order to solve the minimisation problem posed by geodesic active contours between two points p_0 and p, we compute the surface of minimal action U_0 from p_0. At each point p of the image plane, $U_0(p)$ corresponds to the minimal energy integrated along a path that starts at p_0 and ends at p.

$$U_0(p) = \inf_{P(L)=p} \{E[P]\}$$

Cohen and Kimmel [17] observed that it may be described by the Eikonal equation

$$|\nabla U_0| = g, \quad U_0(p_0) = 0 \tag{4}$$

They demonstrated that this surface may be efficiently computed using the Fast Marching Method of Sethian [27]. The Fast Marching Method is an optimal time algorithm for solving Eikonal equations on a discrete grid, computing the points of U_0 in increasing order of value. It bears some similarity to Dijkstra's algorithm [19] for obtaining shortest paths in graphs, however it does not display the metrication error inherent to graph algorithms on a discrete grid.

2.3. Circular Shortest Paths

Circular Shortest Paths (CSPs) were introduced by Sun and Pallottino [29, 30] with application to polar object segmentation, crack detection in borehole cores from geophysics and panoramic stereo matching. The discussion here summarises [5].

Consider a discrete image $c : \mathbb{Z}_u \times \mathbb{Z}_v \to \mathbb{R}$, where $c(i, j)$ is the cost of traversing the pixel at position (i, j).

Let G be the directed, vertex-weighted graph with vertex set $V = \mathbb{Z}_u \times \mathbb{Z}_v$. With the notation $a = (a^1, a^2) \in V$ we define the edge set

$$E = \{(a, b) \,|\, a, b \in V,\ a^2 = b^2 - 1,\ |a^1 - b^1| \leq 1\}$$

Such a graph G is known as a *stable acyclic sequential layered graph* [29], or simply as a *trellis* [7].

A shortest path across such a trellis is defined as a sequence of connected vertices

$$P = (a_j \mid a_j \in V,\ j \in \mathbb{Z}^v,\ (a_j, a_{j+1}) \in E)$$

which minimises the path length

$$L(P) = \sum_{j=1}^{v} c(a_j)$$

over all admissible paths P. Such a path may be efficiently computed using dynamic programming.

However circular shortest paths have the added constraint that $|a_1^1 - a_v^1| \leq 1$. This forces the two endpoints a_1, a_v of the path to meet when considering the trellis to be periodically extended. This global type of constraint cannot be embedded in the framework of standard shortest path algorithms whose efficiency hinges upon a local solution for each vertex.

Sun and Pallottino [29, 30] propose three fast approximate solutions and a fourth exact but slow one, which they called the Multiple Search Algorithm (MSA). The MSA solves the CSP problem in $O(u^2 v)$ time by solving u independent shortest path problems, one for each source vertex in the first layer.

Appleton and Sun [5] proposed a more efficient search using a branch and bound formulation of the problem. They use a generalisation of an observation of Sun and Pallottino: for a given set of starting vertices the length of the shortest path across the trellis from those vertices is a lower bound for the length of all circular paths passing through those vertices. This inspires a divide and conquer approach to locating the circular shortest path. A binary search tree is imposed on the set of potential source vertices in the first layer of the trellis. The root of the binary tree is the entire first layer L_1. Each node is then equi-partitioned by its children into two contiguous sets of source vertices. This partitioning is applied recursively down the tree to the leaves, which are all u single vertices of L_1. Each node has value equal to the shortest path across the trellis from the corresponding starting set. Thus each node's value is a lower bound for the corresponding subtree, and hence a lower bound for the length of the circular paths passing through its source set. A best-first branch and bound search is applied to the binary tree, evaluating only those nodes which may contain the CSP. Dynamic programming is used to evaluate each subtree's lower bound, with a priority queue guiding the search. The algorithm halts when a node whose shortest path is circular has the highest priority.

This branch and bound approach was shown to obtain the globally optimal circular path across the trellis and was demonstrated to reduce the order of the search to $O(u^{1.6} v)$ on random cost functions, arguably

a pathological class due to the lack of periodicity and structure common to applications of CSPs.

We restate the branch and bound CSP algorithm here:

Branch and Bound Circular Shortest Path

1. Initialisation
   - Initialise the Root search node as a non-circular path of arbitrary length
   - Enqueue(Root)
2. Priority-First Search: While (*true*)
   - Let $n = Dequeue()$, where n is a node of the binary tree
     If n is circular:
     - Return n
     - Halt

     For each child χ of n:
     - Compute the shortest path across the trellis for the given source vertices specified by χ
     - Store in χ whether the shortest path was circular
     - $Enqueue(\chi)$ with priority equal to the negative shortest path length

3. Motivation

The purpose of object segmentation is to separate an object from the remainder of the image. Approaches to object segmentation typically utilise the dissimilarity of features between the object and the background. Geodesic active contours seek segmentations which minimise the homogeneity of features along the boundary between the object and the background.

Variational methods are often used to evolve contours toward a local energy minimum. Since the first application of level sets to image segmentation by Caselles et al. [9] a great deal of effort has been put into improving the speed of this evolution. Malladi et al. [25] used the narrow band scheme for greater efficiency, and more recently Goldenberg et al. [20] used an additive operator splitting scheme to further increase the speed of segmentation by geodesic active contours. However these approaches require an initial contour and have the potential to become trapped in local minima unrelated to the object of interest.

The segmentation algorithm presented here is motivated by the desire to avoid these spurious local minima. The output of a global minimisation scheme is the best outcome possible within the solution framework, so poor segmentations become apparent as a poor choice of metric or energy functional. The independence to initial conditions improves the robustness to noise and image pathologies as well as automating the segmentation.

The global minimum of Eq. (1) is only well defined for positive metrics $g > 0$. Indeed if a metric is negative in some region then a closed curve of negative energy may be constructed. Repeated loops about such a curve will produce arbitrarily low energy, and hence no global minimum exists. Regions of zero metric permit arbitrarily convoluted contours such as Peano curves at zero cost, hence we require that g be positive everywhere in order that a meaningful global minimum exist. However for $g > 0$ the global minimum of Eq. (1) is an empty curve with zero energy, so approaches to constructing globally minimal closed curves in the existing geodesic active contour framework will fail.

In the following we present a new framework in which we may obtain a non-trivial closed curve of globally minimal energy. We restrict the set of *admissible* contours to those closed curves which contain a specified interior point, which is used to select the object of interest. Without this restriction the curve will be empty so we consider this to be a natural restriction. Observe that the contour of globally minimal energy under this restriction will always be a simple closed curve—as any curve will have positive energy, a contour composed of multiple closed curves would have greater energy than any of its constituent simple closed curves.

3.1. Geodesic Active Contours About a Point

Let Φ be the set of all simple closed curves containing p_{int}. For simplicity in the following we place the origin of the image axes at $p_{\text{int}} = (0, 0)$. We wish to find the curve of minimal energy in Φ, the set of closed curves containing the origin. Existing methods for computing (globally) minimal geodesics typically rely on the computation of a minimal open curve between two endpoints. The constructions presented in this section convert the problem of computing closed geodesics to that of computing open geodesics. We first present a simplified version of our technique, demonstrating the broad approach without introducing excessive complexity.

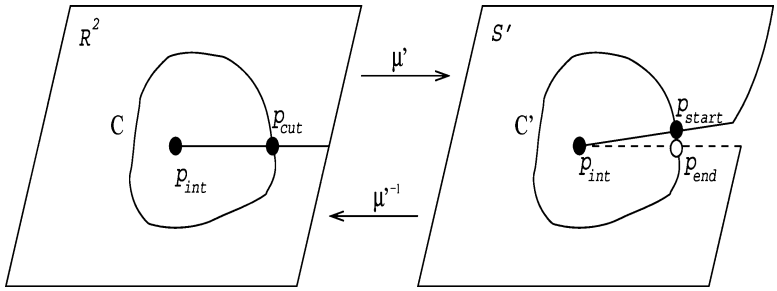


Figure 1. A minimal closed geodesic in the image plane $\mathbb{R}^2$ passing through p_{cut} and the corresponding open geodesic in S' between p_{start} and p_{end}.

We identify the two shortcomings of this simple approach before presenting a more sophisticated construction which overcomes both.

Consider a Riemannian space S', formed from the image plane $\mathbb{R}^2$ by cutting infinitesimally beneath the non-negative x-axis:

$$\forall x \geq 0, \quad \lim_{\epsilon \to 0^-} (x, \epsilon) \neq (x, 0)$$

i.e. the two sides of the cut are disconnected. Figure 1 depicts this space. The mappings $\mu' : \mathbb{R}^2 \to S'$ and $\mu'^{-1} : S' \to \mathbb{R}^2$ may now be introduced:

$$\mu'(x, y) = (x, y) \in S', \quad \mu'^{-1}(x, y) = (x, y) \in \mathbb{R}^2$$

As $\mathbb{R}^2$ and S' have differing topologies this mapping is discontinuous along the cut. We identify the metric g in the two spaces, so that the metric at any point in S' is the same as for the image plane $\mathbb{R}^2$:

$$g(\mu'(p)) = g(p)$$

A closed curve $C : [0, L] \to \mathbb{R}^2$ containing the origin must pass through the positive x-axis at least once. In this simple approach we consider the restricted set of curves $\Phi' \subset \Phi$ which pass through the cut *exactly* once. Let p_{cut} be the unique point of intersection between C and the positive x-axis. Then $C' = \mu'(C)$ is the equivalent open curve in S' with endpoints p_{start} and p_{end} on either side of the cut. For completeness we define the virtual endpoint

$$p_{\text{end}} = \lim_{\epsilon \to 0} \mu'(C(L - \epsilon))$$

The endpoints p_{start} and p_{end} are identified as p_{cut} under the mapping μ' back to $\mathbb{R}^2$. In this simple construction we obtain the restricted minimal curve $C_{\min} = \arg\inf\{E[C] | \, C \in \Phi'\}$.

For a single p_{cut} we define the surface of minimal action $U : S' \to \mathbb{R}^+$:

$$U(p) = \inf\{E[C'] | C' \subset S', \; C'(0) = p_{\text{start}}, \; C'(L) = p\}$$

Following Cohen and Kimmel [17] we observe that U may be described as the viscosity solution to the following Eikonal equation:

$$|\nabla U| = g, \quad U(p_{\text{start}}) = 0$$

Given U the minimal geodesic C' connecting p_{start} and p_{end} may be described as a gradient descent path:

$$\frac{\partial C'}{\partial s} = -\frac{\nabla U}{|\nabla U|}, \quad C'(L) = p_{\text{end}}$$

The corresponding geodesic $C|_{p_{\text{cut}}} = \mu'^{-1}(C')$ is then the minimal closed geodesic in Φ' passing through p_{cut}. The minimal closed geodesic in Φ' is then the least of these restricted geodesics.

This simple construction yields the minimal curve from the set Φ' of closed curves which contain the point p_{int} and pass through the x-axis only once. The restriction that the curve pass through the cut exactly once is described here as 'cut-convexity', an unfortunate artifact of the simple framework presented in this section. The exhaustive search through all cut points for the minimal closed path is potentially quite inefficient, in parallel with the Multiple Search Algorithm [30].

4. General Algorithm

4.1. Representing Cut-Concave Contours

Previously we cut the image plane $\mathbb{R}^2$ along the positive x-axis to obtain the space S', such that paths from one side of the cut to the other were forced to travel around the internal point p_{int}. In this space we were able to represent any simple closed curve containing p_{int} which passed through the cut exactly once. Now we wish to extend our class of admissible curves to include cut-concave curves, those which pass through the cut multiple times. To compute such curves we must allow the curve to pass through the positive x-axis, while still being able to differentiate between a path that has passed around p_{int} and one which has not.

Here we construct a space S in which we may represent any simple closed curve containing p_{int} as a path. To do so we augment the space S' such that the information about whether a simple closed curve contains p_{int} is embedded into the space itself. Consider the space $S = S' \times \mathbb{Z}$, where S' is the image plane $\mathbb{R}^2$ cut along the x-axis as before and $\mathbb{Z}$ corresponds to the signed number of times that the path has passed anti-clockwise around the origin. We connect each side of the cut in each plane S' to the upper and lower planes to form a helical surface, creating a space identical to the Riemann surface of the natural logarithm relation [24]. Thus any admissible closed curve in $\mathbb{R}^2$ corresponds to a path on the helical surface S with its endpoints exactly one layer apart.

The space S and the corresponding representation of a closed curve as a path are depicted in Fig. 2. Note that the perceived slope of S in this figure is solely for viewing and does not affect the metric g.

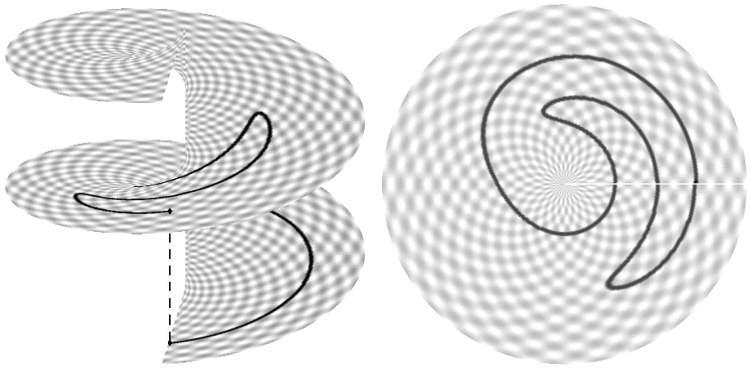
There are many equivalent ways to represent an admissible closed curve C in the image plane $\mathbb{R}$ as an open path in S. Without loss of generality we require that the path begin along the cut at level zero and end at the corresponding point one layer above (Fig. 2). In the case of a curve that passes through the cut multiple times, we select the unique starting point such that the curve lies in the non-negative half-space $S' \times \mathbb{Z}^+$. Open paths in S satisfying these criteria are described as *admissible* in the following. A proof of the existence and uniqueness of this choice of starting point is given in the Appendix. It follows that this representation induces a one-to-one correspondence between the set of admissible closed curves in $\mathbb{R}^2$ and the set of admissible open paths in S.

From an admissible open path in S we may construct the corresponding closed curve in $\mathbb{R}^2$ by the projection $\nu : S \to \mathbb{R}^2$:

$$\nu(x, y, z) = (x, y) \tag{5}$$

4.2. Obtaining Cut-Concave Contours

By searching for the minimum admissible open path in S we may obtain the corresponding minimum admissible closed curve in $\mathbb{R}^2$. For a given p_{cut} the minimum admissible open path in S may be found on a discrete grid using the fast marching method. Let $p_{\text{start}} = (p^1_{\text{cut}}, 0, 0)$ and $p_{\text{end}} = (p^1_{\text{cut}}, 0, 1)$ be the coordinates of the two discrete endpoints of the geodesic. Set $U(p_{\text{start}}) = 0$ and



(a) (b)

Figure 2. The helical representation of a concave curve. (a) Two layers of the helical surface S. A path equivalent to a cut-concave closed curve has been overlayed. (b) The projection of S to the image plane $\mathbb{R}^2$.

compute the surface of minimal action on S, halting when it reaches p_{end}. $U(p_{\text{end}})$ is then the energy of the corresponding geodesic active contour in $\mathbb{R}^2$, with the geodesic extracted by gradient descent of U from p_{end} before being projected back onto $\mathbb{R}^2$ by the mapping ν. Only the points in the non-negative half-space $S' \times \mathbb{Z}^+$ need to be computed due to the earlier choice of p_{start}. For boundary conditions, set

$$U(x \geq 0, 0, 0) = \begin{cases} 0 & x = p^1_{\text{start}} \\ \infty & \text{otherwise} \end{cases}$$

Observe that the choice of space S ensures that curves may pass through the cut any number of times before reaching p_{end}.

4.3. Efficient Search for the Cut Point

The exhaustive search over all possible cut points p_{cut} proposed earlier is inefficient. Here we adapt the Branch and Bound Circular Shortest Path (BBCSP) algorithm [5] described earlier to efficiently search the set of cut points, obtaining the simple closed curve of globally minimal integral. If efficiency is preferred over exactness, the MBTA/IPA algorithm [30] can be used instead to obtain near-optimal geodesic active contours in log-linear time.

We impose a binary search tree on the discrete set of cut points lying in the image, denoted $P^{\text{root}}_{\text{cut}}$. Each node in the binary tree corresponds to a subset of the cut points in the image. Nodes are equi-partitioned by their children into two contiguous halves, down to the leaves of the tree which are all single-pixel nodes. For each node of the binary tree, P_{cut} is a set of potential cut points p_{cut}, with P_{start} and P_{end} the corresponding set of discrete endpoints in S. The minimal geodesic is simply the shortest path between these two sets of endpoints P_{start} and P_{end}. This tree allows us to consider a large number of possible cut points simultaneously, efficiently determining the cut point which gives the minimal closed geodesic.

For any node of the binary tree, the minimal (open) geodesic connecting P_{start} and P_{end} is a lower bound on all closed geodesics passing through that node's cut set P_{cut}. This geodesic and its length may be computed using the fast marching method. Active nodes are stored in a priority queue for efficient access to the minimum. The branch and bound search proceeds by progressively splitting the minimal node to search its children. The search halts when the minimal geodesic for the node with lowest bound is closed.

4.4. General Algorithm

Globally Optimal Geodesic Active Contours

1. Initialisation
   - Initialise the Root search node as an open geodesic of arbitrary length
   - Enqueue(Root)
2. Priority-First Search: (Infinite loop)
   - Let $n =$ Dequeue(), where n is a node of the binary tree
     If n has been flagged as *closed*:
     - Return the minimal closed geodesic corresponding to n
     - Halt
   - Compute the surface of minimal action U in the space S for the cut-set of node n
     - Halt the computation early when at least one element of each child's P'_{end} has been evaluated
   - Locate the end of the geodesic: $p_{\text{end}} = \arg\inf\{U(p_{\text{end}}) \mid p_{\text{end}} \in P_{\text{end}}\}$
   - Obtain the minimal geodesic C_{min} for node n by gradient descent of U from p_{end} to p_{start}

   For each child χ of n:
   - Partition P_{cut} of n to give the subset P'_{cut} of χ
   - Assign to χ a lower bound $\inf\{U(p_{\text{end}}) \mid p_{\text{end}} \in P'_{\text{end}}\}$
   - Flag χ's minimal geodesic as closed if p_{start} and p_{end} are connected in the discrete grid
   - $Enqueue(\chi)$ with priority equal to its negative lower bound

4.5. Optimality

Here we prove the optimality of the general algorithm given in this paper. Lemma 1 establishes the correctness of using the energy of the minimal geodesic between two sets of endpoints as a lower bound for the set of closed geodesics passing through those endpoints. Lemma 2 proves the optimality of the closed geodesics constructed above.

Define the set of admissible curves with respect to a cut set P_{cut} as the set of all simple curves containing p_{int} with both endpoints in the cut set P_{cut}.

Lemma 1. *Let $P'_{\text{cut}} \subseteq P_{\text{cut}}$ be two cut sets and let C' and C denote their minimal admissible curves respectively. Then $E[C] \leq E[C']$.*

Proof: Let $\Psi(P_{\text{cut}})$ be the set of admissible curves with respect to P_{cut} and define $\Psi(P'_{\text{cut}})$ likewise. Then as $P'_{\text{cut}} \subseteq P_{\text{cut}}$ we have $\Psi(P'_{\text{cut}}) \subseteq \Psi(P_{\text{cut}})$, and so $\inf E(\Psi(P_{\text{cut}})) \leq \inf E(\Psi(P'_{\text{cut}}))$. Hence the energy of the minimal admissible curve passing through P_{cut} is a lower bound for the energy of all admissible curves through P'_{cut}. □

Corollary 1. *Observe that if P_{cut} is a single point then $\Psi(P_{\text{cut}})$ is the set of all simple closed curves passing through P_{cut}. Hence by Lemma 1 the energy of the minimal admissible curve through P_{cut} is a lower bound on all closed curves through P_{cut}.*

Lemma 2. *The general algorithm presented above obtains the globally minimal closed curve containing p_{int}.*

Proof: Observe that the cut sets of the search nodes in the queue partition the cut P_{cut}. By the corollary of Lemma 1 the minimum lower bound of the nodes in the priority queue is a lower bound on the energy of the globally minimal closed curve containing p_{int}. So if the minimal curve of the node with lowest bound is closed then this closed curve is the globally minimal closed geodesic. As this is the halting criteria our algorithm obtains the simple closed curve of minimal energy containing p_{int}. □

5. Image Segmentation Based on Globally Optimal Geodesic Active Contours

In this section we discuss the choice of metric for image segmentation and the backtracking scheme to obtain geodesics. A new metric is introduced and related to existing research on polar object segmentation. The numerical problems of computing geodesics on a discrete grid are also investigated.

5.1. Metrics

By introducing an optimal segmentation scheme we have removed the dependence upon initial conditions of geodesic active contours. The segmentation is therefore determined solely by the placement of the internal point p_{int} and the choice of metric. Here we introduce an alteration to the metric commonly used in GAC Eq. (2), the inclusion of a weighting term inversely proportional to the radius r from p_{int}

$$g' = \frac{1}{r}\left(\frac{1}{1 + |\nabla G_\sigma \star I|^p} + \epsilon\right) \tag{6}$$

i.e. $g' = \frac{g}{r}$. The reasons for the inclusion of this term are quite natural. Firstly, without this term a trivial segmentation may be found in the continuous limit. For a smooth and bounded metric g a sufficiently small circle C containing p_{int} will have energy

$$\begin{aligned} E[C] &= \lim_{r\to 0} \oint_C g(C(s))\,\mathrm{d}s \\ &= \lim_{r\to 0} g(p_{\text{int}})L(C) = \lim_{r\to 0} g(p_{\text{int}})2\pi r = 0 \end{aligned}$$

Thus the energy of this contour approaches zero as its size approaches zero. This situation is analogous to the problem of the global minimum of Eq. (1) being an empty curve. However, with the introduction of the $\frac{1}{r}$ term such a small circle C will instead have non-zero energy:

$$\begin{aligned} E[C] &= \lim_{r\to 0} \oint_C g'(C(s))\,\mathrm{d}s \\ &= \lim_{r\to 0} \frac{g(p_{\text{int}})}{r} 2\pi r = 2\pi g(p_{\text{int}}) \end{aligned}$$

The introduction of the $\frac{1}{r}$ term also introduces scale invariance to the energy functional. A problem in early work on active contours was the attraction of curves toward the global minimum of zero energy, causing snakes to shrink when placed in a homogeneous region. While many heuristic methods have been proposed to avoid this and other undesirable minima [14, 15, 33], the introduction of the $\frac{1}{r}$ term removes this problem by ensuring that small contours are not preferred *per se* over large contours.

The $\frac{1}{r}$ term may also be related to segmentation of the polar transform of an object, as previously investigated by Denzler and Niemann [18]. One choice of energy functional for polar segmentation expressed as a continuous integral is of the form [7]:

$$E[C] = \int_C g(C(r,\theta))\,\mathrm{d}\theta \tag{7}$$

As $\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\mathrm{d}\theta^2$ and $\mathrm{d}r$ is omitted in Eq. (7), it may equivalently be written by a projection of $\mathrm{d}s$

onto $\overrightarrow{\theta}$:

$$E[C] = \int_C \frac{g(C(s))}{r} \langle \mathrm{d}s, \overrightarrow{\theta} \rangle$$

Note the close similarity to Eq. (6).

5.2. Backtracking Methods

Once the surface of minimal action U has been computed the minimal geodesics may be obtained as the gradient descent paths of U. In the continuous theory the only local minimum of U is the starting set, so gradient descent paths are guaranteed to converge to this set. Unfortunately the implementation on a discrete grid is problematic. Many existing methods suffer from inaccuracy due to grid bias or stability problems, due to the introduction of false minima in the interpolation of U or the creation of vortices in the interpolation of ∇U. Here we consider previous schemes and present a simple approach which is both stable and isotropic.

Cohen and Kimmel [17] present two approaches. The first of these is to view the minimal geodesics between two points p_0 and p_1 as the set of points p_g that satisfy

$$U_0(p_g) + U_1(p_g) = \inf_{p \in \mathbb{R}^2} \{U_0(p) + U_1(p)\}$$

where U_0 and U_1 are the minimal action surfaces to p_0 and p_1 respectively. As this approach is applied to discrete data the function $U_0 + U_1$ is instead thresholded with a value larger than its infimum in order that the resulting path be connected. This produces a fat path which must then be either thinned or refined by curve evolution.

Their second, simpler approach is a back propagation procedure implemented by network descent of the surface of minimal action. The tracking point is only allowed to move in a 4-connected or 8-connected grid, jumping to the neighbour of minimal action at each step. The resulting gradient descent scheme is unconditionally stable as the discrete surface of minimal action computed using a first-order Fast Marching Method has no local minima except for the starting set. Unfortunately it is anisotropic due to the metrication error of the gradient used in the descent, producing incorrect contours even in the continuous limit.

Sethian [28] proposed a second order improved Euler's method to integrate the geodesic for global first order accuracy. However to interpolate gradients at continuous points from a discrete grid he suggests using a bilinear interpolation of ∇U. This interpolation scheme is inconsistent with the minimisation framework as it causes the descent at a point to depend upon the gradient at higher values of U. In practice the authors have found this to also produce vortices in the interpolated gradient field, occasionally trapping the gradient descent in endless loops.

Kimmel [22] recommends computing the gradient descent path on a cubic surface interpolated from the values of U on discrete mesh points. However this scheme can create false local minima in the interpolated minimal action surface, again trapping the gradient descent process.

Here we propose a scheme for rectilinear grids which overcomes these stability and metrication problems. We bilinearly interpolate the surface U and analytically determine the gradients at continuous points. Without loss of generality let the grid step $h = 1$. Let $[U_{ij}]_{22}$ be the values of the computed surface U on the corners of the grid square in which the continuous point lies:

$$[U_{ij}]_{22} = \begin{bmatrix} U(\lfloor x \rfloor, \lfloor y \rfloor) & U(\lfloor x \rfloor, \lceil y \rceil) \\ U(\lceil x \rceil, \lfloor y \rfloor) & U(\lceil x \rceil, \lceil y \rceil) \end{bmatrix} \tag{8}$$

and thus define the bilinear interpolation scheme

$$\begin{aligned} U(x, y) &= U_{ij} P^i(x - \lfloor x \rfloor) P^j(y - \lfloor y \rfloor), \\ P(s) &= [1 - s, s] \end{aligned} \tag{9}$$

As earlier, repeated indices imply summation.

The gradient ∇U in the interior of grid squares is simply

$$\nabla U(x, y) = \left[U_{ij} \frac{\partial P^i}{\partial x}(x - \lfloor x \rfloor) P^j(y - \lfloor y \rfloor), \; U_{ij} P^i(x - \lfloor x \rfloor) \frac{\partial P^j}{\partial y}(y - \lfloor y \rfloor) \right] \tag{10}$$

The gradient on grid lines and grid points may be defined using one-sided directional derivatives. The one-sided directional derivative of $U(\vec{x})$ in the direction of the non-zero vector $\vec{\theta}$ is defined by

$$\nabla_{\vec{\theta}} U(\vec{x}) \equiv \lim_{\varepsilon \to 0^+} \frac{U(\vec{x} + \varepsilon \vec{\theta}) - U(\vec{x})}{\varepsilon |\vec{\theta}|}$$

For simplicity in the following we define $\nabla_{\vec{0}} U = 0$. We then wish to define $\nabla^- U$ at any point as

$$\nabla^- U \equiv \vec{\theta} \nabla_{\vec{\theta}} U, \quad \vec{\theta} = \arg\min_{|\vec{\theta}|=1} \{\nabla_{\vec{\theta}} U\}$$

For the interpolation scheme of Eq. (9) this may be derived separably in each dimension using a simple switching scheme. Define

$$\nabla^- U = [U_x^-, U_y^-]$$

with

$$U_{\hat{x}}^- = \begin{cases} \lambda \nabla_{\vec{\theta}} U, \lambda = \arg\min_{\lambda \in \mathbb{R}} \{\nabla_{\lambda \hat{x}} U\} & x \in \mathbb{Z} \\ \nabla U \cdot \hat{x} & \text{otherwise} \end{cases}$$

where $\hat{x} = [1, 0]^T$. U_y^- is defined analogously. Observe that where $U(x, y)$ is smooth $\nabla^- U = \nabla U$.

Bilinear interpolation of the surface from the values at discrete grid points introduces no new local extrema. Begin by considering the function $U(x, y)$ in the interior of a grid square. $U(x, y)$ is a second order polynomial in this region so we may analyse its Hessian $\nabla^2 U$:

$$\det \begin{pmatrix} \frac{\partial^2 U}{\partial x^2} & \frac{\partial^2 U}{\partial x \partial y} \\ \frac{\partial^2 U}{\partial y \partial x} & \frac{\partial^2 U}{\partial y^2} \end{pmatrix} = -\left(\frac{\partial^2 U}{\partial x \partial y} \right)^2 \leq 0$$

where $\frac{\partial^2 U}{\partial x^2} = \frac{\partial^2 U}{\partial y^2} = 0$ due to the bilinearity of the interpolation scheme. As all higher derivatives are zero this contravenes the requirement for the existence of an extremum.

Now consider the values of U along a line connecting two neighbouring grid points. Without loss of generality consider the case of two grid points (x', y) and $(x' + 1, y)$. Along this line U is the linear function $U(x' + \lambda, y) = [U(x', y) \quad U(x' + 1, y)][1 - \lambda \quad \lambda]^T$ for $\lambda \in [0, 1]$. Then $\frac{\partial^2 U}{\partial x^2}$ and all higher derivatives are zero, contradicting the requirement for the existence of a local extremum.

Finally consider a grid point $\vec{x} \in \mathbb{Z}^2$ which is not a discrete extremum, in the sense that it has two 4-connected neighbours $\vec{x_1}, \vec{x_2}$ with $U(\vec{x_1}) < U(\vec{x}) < U(\vec{x_2})$. Then we may consider the one-sided directional derivatives in the directions $\vec{\theta_1} = \vec{x_1} - \vec{x}$ and $\vec{\theta_2} = \vec{x_2} - \vec{x}$. We observe from Eq. (9) that $\nabla_{\theta_1} U = U(\vec{x_1}) - U(\vec{x}) < 0$ and $\nabla_{\theta_2} U = U(\vec{x_2}) - U(\vec{x}) > 0$, and hence that $U(\vec{x})$ is not a local extremum.

At saddle points in the interior of grid squares, $\nabla^- U = \vec{0}$ while $\det(\nabla^2 U) < 0$. The descent at these points follows the eigenvector of the Hessian $\nabla^2 U$ with negative eigenvalue.

This gradient interpolation scheme is simple to implement. Contour tracking is performed using a second order improved Euler's method for global first order accuracy. The resulting contours demonstrate that this formulation is isotropic.

6. Complexity Analysis

Here we analyse the complexity of the algorithm presented in Section 4 following the argument from [5]. In general Branch and Bound techniques have poor worst-case behaviour on pathological data [34], so here we analyse the average running time of our algorithm. For simplicity we consider the case of a square image with sidelength M.

The halting condition for our algorithm is that the minimal closed geodesic is shorter than or equal to the minimal geodesics for all other cut sets. Hence our algorithm must split each cut set which admits a geodesic shorter than the minimal closed geodesic. We call such geodesics *short geodesics*.

This binary tree of cut sets must be evaluated along all branches from the root (the full cut set) to those nodes which split the short geodesics. The priority-first search ensures that only these nodes are evaluated, minimising the number of geodesic computations required. Short geodesics whose endpoints are far from meeting will be split near the root of the search tree, and will tend to be split incidentally along the way to other geodesics. On the other hand, short geodesics whose endpoints are nearly connected quite possibly share a large portion of their length with the minimal closed geodesic, and so are worth investigating.

Let $N(n)$ be the number of search nodes to be evaluated in the subtree rooted at node n. Let $R \equiv P_{\text{cut}}^{\text{root}}$ denote the root of the search tree, with χ_1 and χ_2 its children. Due to the data dependant complexity of Branch and Bound we use the simple model of the search dynamics proposed by Zhang and Korf [34]. This model assumes that each node has a constant probability β of being split given that its parent has been split. Then the expected number of geodesic computations to perform is

$$\mathbb{E}(N(R)) = 2 + \beta \mathbb{E}(N(\chi_1)) + \beta E(N(\chi_2))$$

Given that children symmetrically partition the cut sets of their parents we have $\mathbb{E}(N(\chi_1)) = \mathbb{E}(N(\chi_2))$ and so

$$\mathbb{E}(N(R)) = 2 + 2\beta\mathbb{E}(N(\chi_1))$$

The cut in a discrete $M \times M$ image has length at most M. Then the binary search tree has at most $\lceil \lg M \rceil + 1$ levels with the root always split. Recursing we obtain

$$\begin{aligned}\mathbb{E}(N((R)) &= 2 + 2\beta(2 + 2\beta(\cdots)) \\ &= 2\left(1 + 2\beta + (2\beta)^2 + \cdots + (2\beta)^{\lceil \lg M \rceil}\right)\end{aligned}$$

So

$$\mathbb{E}(N(R)) = \begin{cases} 2 \cdot \dfrac{(2\beta)^{\lceil \lg M \rceil + 1} - 1}{2\beta - 1} & \beta \neq \frac{1}{2} \\ 2(1 + \lceil \lg M \rceil) & \beta = \frac{1}{2} \end{cases}$$

The cost of splitting a search node is the cost of running the Fast Marching Method over an $M \times M$ image, $O(M^2 \lg M)$ [27].

For $\beta > \frac{1}{2}$, $\mathbb{E}(N(R)) \approx 2 \cdot \frac{(2\beta)^{\lceil \lg M \rceil + 1}}{2\beta - 1} \leq \frac{8\beta^2}{2\beta - 1} \cdot M^{\lg(2\beta)}$ for large M, giving an overall time complexity of $O(M^{2+\lg(2\beta)} \lg M)$.

For $\beta = \frac{1}{2}$, we obtain an overall time complexity of $O(M^2 \lg^2 M)$ directly.

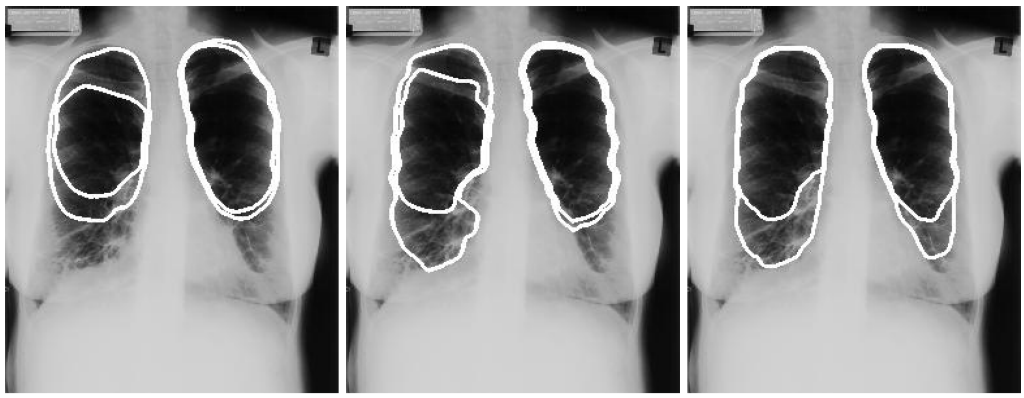
For $\beta < \frac{1}{2}$, $\mathbb{E}(N(R)) \approx \frac{2}{1-2\beta}$ for large M, giving an overall time complexity identical to the Fast Marching Method at $O(M^2 \lg M)$.

7. Results

In this section we present results for the segmentation of image data in the form of microscope, x-ray, MRI and cDNA images. All tests were performed on a 700 MHz Toshiba P-III laptop with 192MB of RAM under the Linux operating system. The algorithm has been implemented in the C language.

7.1. Comparison to Classic Geodesic Active Contours

Here we compare the segmentations produced by the globally optimal GAC scheme presented in this paper to two iterative approaches. The first of these is the classic GAC scheme proposed by Caselles, Kimmel and Sapiro [10], a variational approached based on curve evolution via a level set embedding. The second of these is the Gradient Vector Flow (GVF) scheme of Xu and Prince [32]. The test image is a chest x-ray where the application is to find the boundaries of both smooth and textured portions of lungs. Textured portions can sometimes indicate inflamation or other diseases. The metric uses a simple morphological texture model. Figure 3(a) is a segmentation by the classic GAC scheme. Observe that the contours become stuck in local minima such as that produced by the collar bone projected onto the right lung. Figure 3(b) is a segmentation using the GVF scheme. Although it presents an improvement over the classic GAC scheme it has failed to segment large portions of both lungs. Figure 3(c) is a segmentation



(a) (b) (c)

Figure 3. Comparison of segmentation by iterative methods to globally optimal GAC. The interior contour contains the smooth portions of the lung, while the exterior contour encompasses both smooth and textured portions. The textured portions of lungs often indicate inflamation. (a) Segmentation by gradient descent GAC. (b) Segmentation by Gradient Vector Flows. (c) Segmentation by globally optimal GAC.

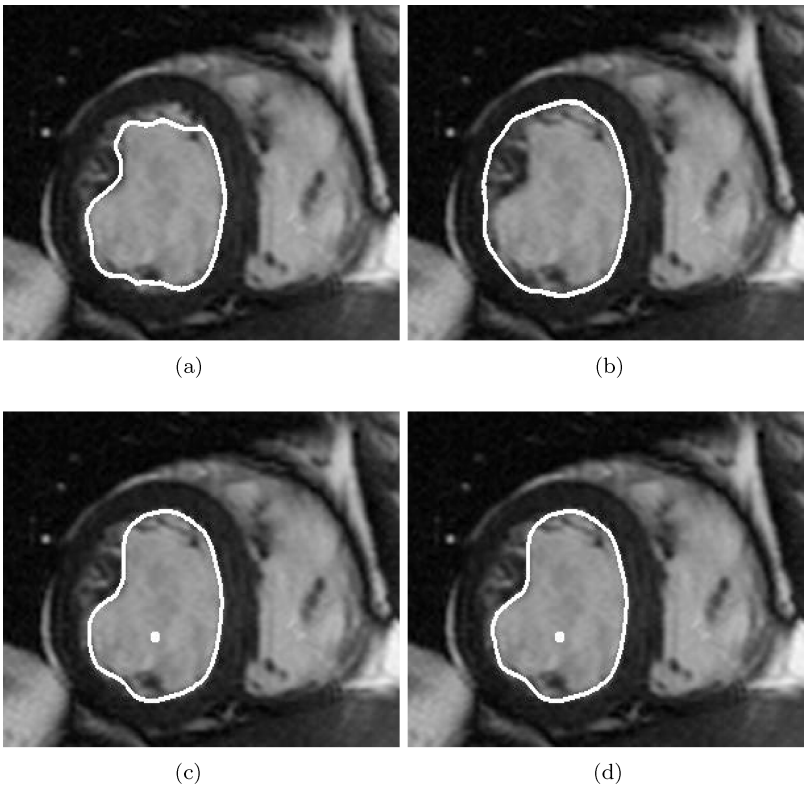


Figure 4. (a) Classic GAC segmentation (narrow-band, explicit time step): 21.5 seconds. (b) Multiscale GAC segmentation (narrow-band, implicit time step): 3.1 seconds. (c) Globally optimal GAC segmentation: 2.5 seconds. (d) 'Near-optimal' GAC segmentation via MBTA/IPA-like approach: 0.35 seconds.

using the new optimal scheme. The metric is the same as that used in Fig. 3(a) before radial weighting. This segmentation offers a significant improvement, having avoided the local solutions which trapped the contours in the two iterative approaches.

7.2. Computational Speed

Here we compare the computational speed of Caselles' classic GAC scheme [10], the proposed globally optimal GAC scheme, and their variants. The test example is a 301×221 magnetic resonance image of a human heart, the goal being to segment the endomyocardium. Figure 4(a) depicts the result of the classic GAC segmentation implemented by level set evolution. It converges in 21.5 seconds. This implementation uses a narrow-band approach with explicit time step 90% of the maximum given by the Courant-Freidrichs-Levy stability condition. Figure 4(b) depicts the result of a multiscale GAC segmentation, converging in 3.1 seconds. This implementation uses a narrow-band approach with an implicit time step $\tau = 20$ as described by Goldenberg et al. [20], embedded in a multiscale framework for efficiency and improved convergence. Figure 4(c) shows the result of the globally optimal GAC segmentation presented here. The algorithm takes 2.5 seconds. Figure 4(d) depicts the result of a 'near-optimal' GAC segmentation via an MBTA/IPA-like approach in the same framework as the simple algorithm proposed in Section 3. The computation time for this approach is 0.35 seconds.

Both iterative segmentations (a–b) were initialised with a circle inside the endomyocardium centred on the internal point depicted in segmentations (c–d).

7.3. Computational Complexity

Here we present experimental results which verify the simple complexity analysis of Section 6. In

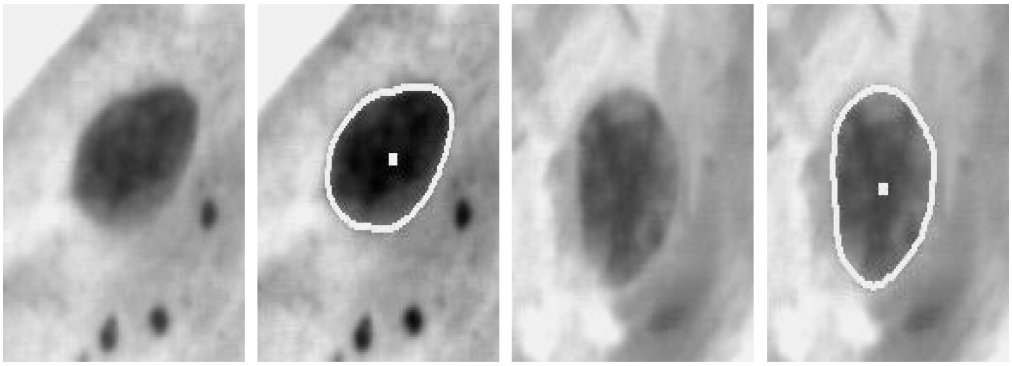


Figure 5. Two examples from the CSSIP Papanicolou stained cell image database [6]. Figures (a) and (c) depict the original cell images while figures (b) and (d) show the corresponding segmentations.

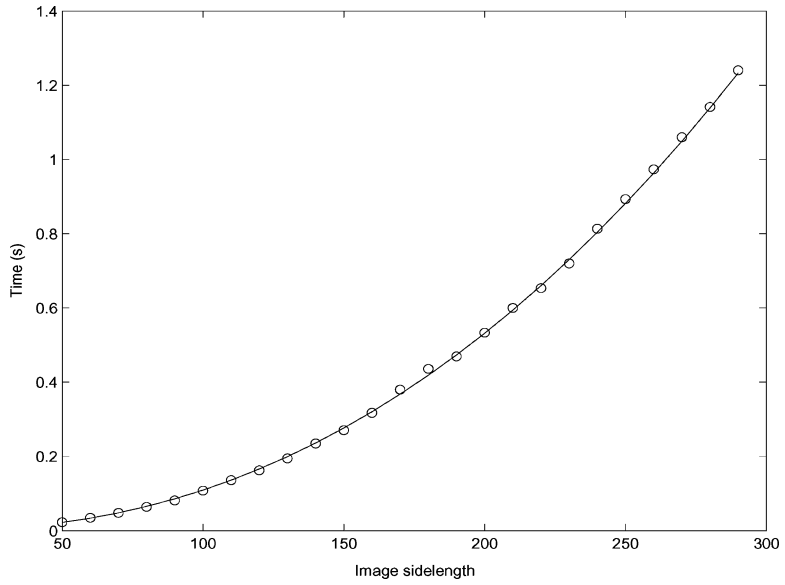


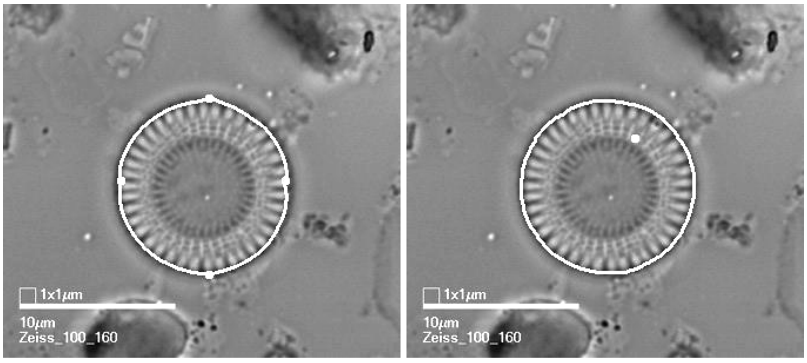
Figure 6. Average segmentation time for $M \times M$ images over a range of sizes. The fitted curve is proportional to $M^P \lg M$ with $P = 2.08$.

order to estimate the average running time of this method it has been used to segment 500 microscope images from a database of Papanicolou stained cells [6]. Figure 5 shows two images from this dataset.

The results depicted in Fig. 6 show the average running time for square images ranging from size 50×50 to 300×300. Each image was resized to a square of the desired sidelength M before segmentation, with the computation times averaged to produce these timings. Also shown is a fitted curve of the form $KM^P \lg M$ with K a constant and power $P = 2.08$ corresponding to a splitting probability $\beta = 0.53$. These results clearly agree with the simple complexity analysis of Section 6, and demonstrate that this method requires an amount of computation that is near linear in the number of pixels.

8. Discussion

The segmentation framework put forward here compares favourably with standard iterative approaches to active contours (Fig. 3). For efficiency the fast marching method implemented here is the first order approximation to the Eikonal equation (Eq. (4)) proposed



(a) (b)

Figure 7. Segmentation of a microscope image of *Cyclostephanos Dubius*, a diatom. (a) Piecewise-geodesic segmentation based on the work of Cohen and Kimmel. Four points were required to segment this object. (b) Segmentation by a minimal closed geodesic under the same metric. This method is robust to poor placement of the internal point.

in [27]; higher order schemes may be used to improve the precision of the segmentation. Conversely for situations where greater speed is required an approximate method based on MBTA/IPA described in [30] can be used.

These computational results demonstrate that the framework presented here is competitive with variational GAC segmentation schemes while it is arguably easier to implement than the AOS numerical scheme for GAC [20].

8.1. Comparison to Cohen's Related Approach

This approach is related to the work of Cohen and Kimmel [17] and Cohen [16] on segmentation using geodesics between seed points. However these methods are not generally suitable for object segmentation, as they require a number of points on the boundary of the object of interest. Figure 7 depicts a microscope image of *Cyclostephanos Dubius*, a diatom. Figure 7(a) demonstrates the piecewise-geodesic segmentation of Cohen and Kimmel while Fig. 7(b) demonstrates the segmentation method proposed in this paper. Dots denote boundary points or the internal point p_{int} respectively. Poor placement of the geodesic endpoints in the piecewise-geodesic schemes will result in poor segmentations as the contours are forced to pass through them. By comparison, the method proposed in this paper requires only the internal point p_{int} in order to segment an object. The contour is not required to pass through this point and the resulting segmentation is robust to poor localisation.

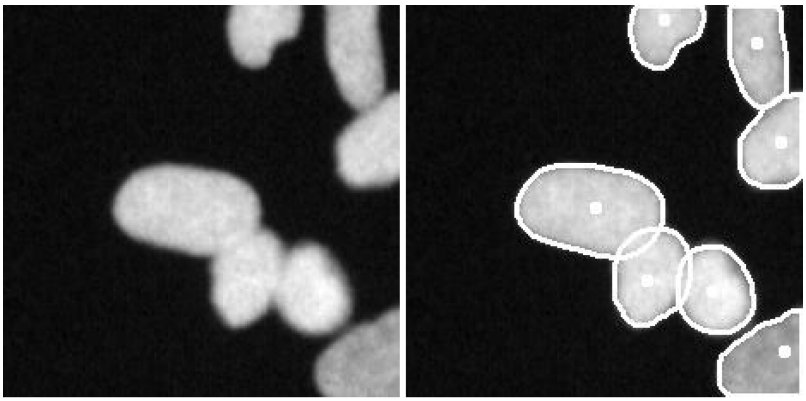
8.2. Object Separation

Object boundaries are often difficult to detect along transitions to adjacent objects with similar features. Geodesic active contours use the assumed regularity of the segmentation contour to avoid leaking through such gaps. Figure 8(a) depicts some cell clusters in a microscope image. They have been separated using the approach presented in this paper (Fig. 8(b)). Observe that on the indistinct border between cells the segmentation contours approximately interpolate the cell contours, overlapping suitably where appropriate.

The segmentation of these cells demonstrates a fundamental limitation or feature of the framework presented here. Contours composed of multiple closed curves will have greater energy than any of their constituent simple closed curves. While a variational approach to GAC allows the formation of multiple closed curves, the optimisation approach is implicitly constrained to find a simple closed curve. However as demonstrated in Fig. 8 it may still be applied to produce multiple closed curves from a set of seed points.

8.3. Relation to Polar Trellis Segmentation

In Section 5.1 we considered a relation between this method and object segmentation on a polar trellis. A major limitation of the polar transform approach to segmentation is that it may only be applied to point-convex objects. The method proposed here overcomes this limitation by considering all simple closed curves containing the specified internal point. Here we demonstrate



(a) (b)

Figure 8. Globally optimal GAC applied to object separation. The cells (a) are separated despite the weak intensity gradient between them (b).

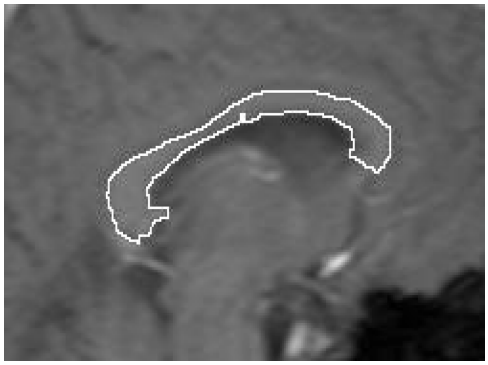


Figure 9. Segmentation of a magnetic resonance image of the corpus callosum in a human brain.

the ability of our scheme to segment objects with convoluted or concave boundaries. Figure 9 depicts a magnetic resonance image of the corpus callosum in a human brain. Here the segmentation contour is point-concave and passes through the x-axis multiple times.

8.4. Constraining the Segmentation

In some applications it is desirable to bias or constrain the segmentation to take advantage of prior information. In situations where an approximate contour is known this information may be used to spatially weight the metric. Reducing the metric in the proximity of this contour will bias the solution to make use of the corresponding prior information. A more direct approach is to block the propagation of the surface of minimal action from undesired areas. This can be used to force the contour to contain an arbitrary connected shape or to restrict its progress to a bounding box or region. In addition to constraining the optimisation this can also significantly reduce the computational load.

8.5. Segmentation Quality Measure

The energy of the segmentation contour may be used as a measure of segmentation quality. A high energy contour contains a large proportion of weak or non-existent edges and is indicative of a poor segmentation. This measure may be used in a post-processing step to grade segmentation contours and further investigate poor segmentations.

Figure 10 demonstrates an application of both spatial constraints and the use of segmentation quality measure.

Figure 10(a) is a fluorescence microarray image for cDNA functional analysis. This type of microarray is used to analyse the expression levels of large collections of genes. From an image analysis point of view, the objective is to segment the visible dots and measure their brightness. The microarray dots form a grid whose parameters are known approximately, however due to the printing process there is some jitter in the placement of the spots. While the majority of spots are circular it is desirable to obtain the shape of all spots rather than impose the assumption of circularity. Many spots are not visible as they produced little or no fluorescence. Using the algorithm presented in this paper each spot is individually segmented with the contour constrained to include a small range of jitter about its expected bounding box. Additionally the energy of the optimal contour in each grid box gives the similarity of the spot to the background, so a simple threshold is used to

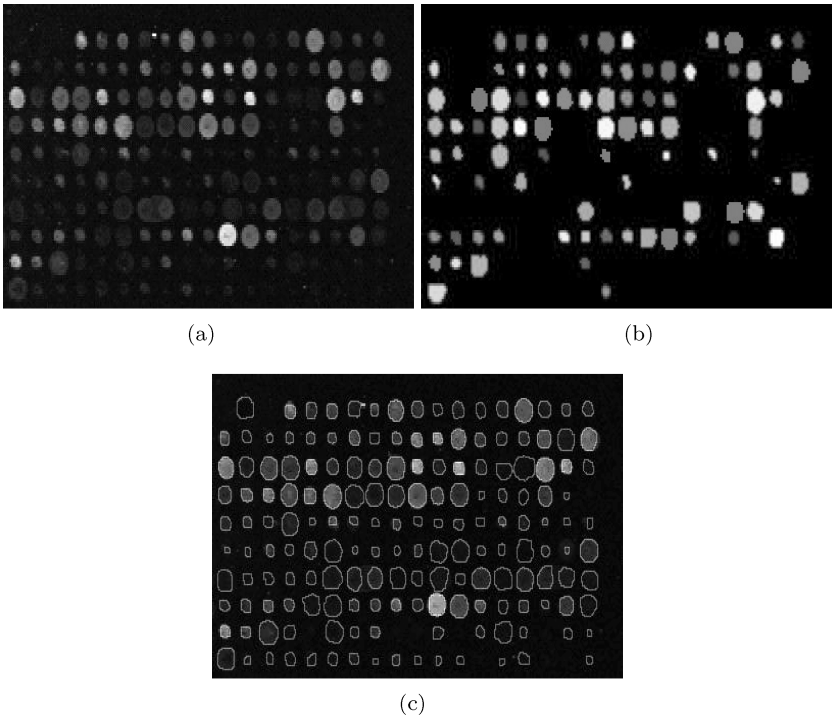


Figure 10. Segmentation of a cDNA microarray image. This image was acquired with an Axon GenePix microarray scanner, and may be obtained from [8]. (a) The original cDNA microarray image. (b) Segmentation via Seeded Region Growing. (c) Segmentation via globally optimal GAC.

discard spots with little or no response. The results are compared to a segmentation via Seeded Region Growing [3], in which each spot is segmented within its grid square. The seeds in this method are the centre point of the grid square and the four corners. The segmentation via globally optimal GAC misses fewer spots and gives generally better contours than the seeded region growing method.

8.6. Limitation to 2D

The proposed method uses shortest path methods extensively, which prevents its natural extension to higher-dimension images. However the authors are currently working on a different approach altogether which yields the same results but is not limited to 2D [4].

9. Conclusion

We have presented a novel framework for object segmentation under the geodesic active contour energy model and demonstrated its usefulness with various applications. We have also developed an efficient algorithm for obtaining optimal segmentations under the new framework.

The new segmentation method seeks the geodesic active contour of minimal energy under the sole restriction that it contains a specified internal point p_{int}. This internal point selects the object of interest and may be used as the only input parameter to yield a highly automated, optimal object segmentation scheme. The image to be segmented is represented as a Riemannian space S with a metric g induced by the image content. The metric is an isotropic and decreasing function of the image gradients, thus encoding the local homogeneity of the image. Using this space segmentation is achieved by locating the minimal closed geodesic containing p_{int}. The image plane $\mathbb{R}^2$ is augmented to form a spiral space S identical to the Riemannian surface for the natural logarithm relation. This space naturally embeds the information of whether a closed contour contains p_{int} without restricting the contour in any way.

Geodesics are computed on the discrete grid using Sethian's fast marching method. Minimal closed geodesics are located efficiently using a best-first

branch and bound search tree adapted from previous work on circular shortest paths. The metric typically used in geodesic active contours is altered to obtain a scale invariant energy functional suitable for global minimisation. The optimal closed contour partitions the image into object and background such that the total similarity across the partition border is minimised. The algorithm proposed here is shown to efficiently obtain the globally optimal geodesic active contour in planar images. While this method cannot be directly extended to the computation of globally minimal surfaces, the authors are currently investigating an alternate approach to the general N-dimensional segmentation problem.

The proposed algorithm has been applied to the segmentation of microscope, x-ray, magnetic resonance and cDNA microarray images. The resulting contours have been shown to be isotropic and demonstrate robustness to gaps in object boundaries as well as low sensitivity to the placement of the interior point. They have been successfully applied to concave and convoluted boundaries, demonstrating the flexibility of the framework developed here. The new approach compares quite favourably with the classic curve evolution approach of Caselles et al. [10], achieving more reliable and accurate segmentations with reduced computational effort.

Appendix

Here we analyse the representation, introduced in Section 4.2, of an admissible closed curve in $\mathbb{R}^2$ as an admissible open curve in the spiral space S. A curve in $\mathbb{R}^2$ is defined as *admissible* if it is simple, closed, and contains the origin p_{int}. A curve in S is *admissible* if it is a simple open curve or path with one endpoint on the cut at the zero level and the other endpoint exactly one layer above. Furthermore it must lie entirely in the positive half-space $S' \times \mathbb{Z}^+$.

All admissible open curves in S may be projected one-to-one to their corresponding admissible closed curves C in $\mathbb{R}^2$ via Eq. (5). However it may not be clear that to each admissible closed curve in $\mathbb{R}^2$ is associated a unique admissible open curve in S, particularly in the case of a cut-concave curve. The choice of such a representation relies upon the selection of the point p_{cut} at which to cut the closed curve C to form the corresponding open curve lying entirely in $S' \times \mathbb{Z}^+$. Here we give a constructive proof that such a point exists uniquely for all simple closed curves C containing the origin. It makes use of elementary knot theory, which is introduced in [2].

A *knot* is a simple closed curve typically embedded in $\mathbb{R}^3$. A knot may be deformed isotopically to yield an equivalent knot by smoothly deforming the curve, not allowing the curve to pass through itself. Two knots form a non-trivial *link* if they are entangled and may not be separated isotopically. The knots considered here are the curve C and the z-axis of S identified with p_{int}.

Knots may be represented by a projection to the plane $\mathbb{R}^2$. To retain the structure of the original knot we make note of intersections in the planar projection, known as *crossings*. A knot may be deformed infinitesimally to ensure that crossings only occur pairwise between strands at single points. Crossings designate the orders of pairs of strands as *upper* or *lower* in the lost dimension of the projection. They may be labelled by the signum of the cross-product of the planar tangent of the upper strand with respect to the lower strand. The sum of the signs of the crossings of a link is an isotopic invariant equal to twice the *linking number*, a measure of the degree to which two knots are entangled. The knots C and p_{int} have unit linking number, indicating that C passes around p_{int} once in total. Here we orient C with respect to p_{int} so that their linking number is positive.

A planar representation of a knot or link may be manipulated isotopically using *Reidemeister* moves [2]. Of particular use in this proof is the second type of Reidemeister move, depicted in Fig. 11. It removes a pair of sequential crossings along C of alternating sign. In Fig. 11 each *segment* of the knot C (a portion of C which does not pass behind the z-axis) is labelled by the common z-value of that segment, the number of times that the curve has passed around the z-axis from p_{cut}. The strand labels are the running sum of crossing

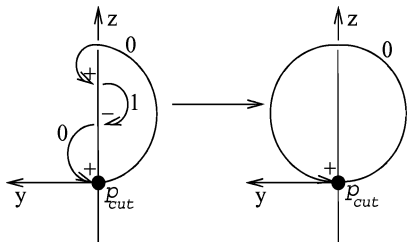


Figure 11. A type-II Reidemeister move elides a $(+1, -1)$ crossing pair. Depicted before and after are crossing signs and strand labels for the desired choice of p_{cut}.

numbers for understrands of C, where we set the label of the first strand after p_{cut} to zero. The desired cut point is that which results in no negative strand labels.

We may consider the set of signs of undercrossings of C to be a cyclic sequence λ with terms taken from $\{+1, -1\}$. The total sum of this sequence is the linking number $+1$. Consider a representation of λ as the linear sequence with end point corresponding to p_{cut}. Then the strand labels are given by the running sum $\Lambda^i = \sum_{j=1}^{i-1} \lambda^j$, the minimum of which must be non-negative. A type-II Reidemeister move is equivalent to eliding a sub-sequence $(\pm 1, \mp 1)$. Observe that eliding the 'positive' sub-sequence $(+1, -1)$ cannot change the minimum of Λ, whereas eliding the 'negative' sub-sequence $(-1, +1)$ may increase it.

Repeated application of the 'positive' elision to the cyclic sequence λ is sufficient to reduce the sequence to $\lambda' = (+1)$. As a simple proof by contradiction, consider an irreducible cyclic sequence of elements $\{\pm 1\}$ with sum $+1$ which does not contain a 'positive' pair $(+1, -1)$. As the sum is positive there must exist a positive element, which cannot be followed by a negative element. By induction the sequence λ' is entirely positive, and with sum $+1$ must consist solely of the element $+1$. The minimum (and indeed only) value of the corresponding label sequence Λ' is 0, and this must also be the minimum of Λ as noted previously. Therefore this sole remaining element corresponds to the unique choice of cutting point p_{cut} giving a non-negative label sequence Λ.

This choice of p_{cut} defines the unique representation of an admissible closed curve C in ^{2} as an admissible open curve in S.

Acknowledgments

Figure 7 was taken from the ADIAC public data web page:

http://www.ualg.pt/adiac/pubdat/pubdat.html (CEC contract MAS3-CT97-0122).

We would like to acknowledge Dr. Stephen Rose, Centre for Magnetic Resonance, University of Queensland and the Centre for Advanced MRI, The Wesley Hospital, Brisbane for making available the cardiac MR image of Fig. 4. We would also like to acknowledge Dr. Michael Buckley, CSIRO Mathematical and Information Sciences, for permission to use the microarray image of Fig. 10, acquired with an Axon GenePix microarray scanner. This image may be obtained from [8]. The first author would like to acknowledge Jenna Appleton for interesting discussions on the proof in the Appendix. We thank the anonymous reviewers for their helpful comments.

References

1. D. Adalsteinsson and J.A. Sethian, "A fast level set method for propagating interfaces," *Journal of Computational Physics*, Vol. 118, No. 2, pp. 269–277, 1995.
2. C.C. Adams, *The Knot Book*, W. H. Freeman and Company, 1994.
3. R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, pp. 641–647, 1994.
4. B. Appleton and H. Talbot, "Globally optimal surfaces by continuous maximal flows," in C. Sun, H. Talbot, S. Ourselin, and T. Adriaansen (Ed.), *Digital Image Computing: Techniques and Applications, Proc. VIIth APRS conference*, Vol. 2, pp. 987–996. CSIRO publishing: Sydney, 2003.
5. B. Appleton and C. Sun, "Circular shortest paths by branch and bound," *Pattern Recognition*, Vol. 36, No. 11, pp. 2513–2520, 2003.
6. P. Bamford, Performance evaluation in image segmentation. http://www.cssip.uq.edu.au/staff/bamford/Segmentation Evaluation/index.htm.
7. P. Bamford and B. Lovell, "Unsupervised cell nucleus segmentation with active contours," *Signal Processing (Special Issue: Deformable Models and Techniques for Image and Signal Processing)*, Vol. 71, No. 2, pp. 203–213, 1998.
8. M. Buckley, "Spot image analysis software," http://experimental.act.cmis.csiro.au/Spot/index.php.
9. V. Caselles, F. Catte, T. Coll, and F. Dibos, "A geometric model for active contours in image processing," *Numerische Mathematik*, Vol. 66, pp. 1–31, 1992.
10. V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *IJCV*, Vol. 22, No. 1, pp. 61–79, 1997.
11. V. Caselles, R. Kimmel, G. Sapiro, and C. Sbert, "Minimal surfaces based object segmentation," *IEEE Trans. on PAMI*, Vol. 1, No. 9, pp. 394–398, 1997.
12. T. Chan and L. Vese, "An active contour model without edges," in *Scale-Space Theories in Computer Vision*, 1999, pp. 141–151.
13. Y. Chen, T. Huang, and Y. Rui, "Optimal radial contour tracking by dynamic programming," in *Proc. of IEEE ICIP*, Thessaloniki, Greece, October 2001.
14. L.D. Cohen, "On active contour models and balloons," *Computer Vision, Graphics, and Image Processing. Image Understanding*, Vol. 53, No. 2, pp. 211–218, 1991.
15. L.D. Cohen and I. Cohen, "Finite-element methods for active contour models and balloons for 2-D and 3-D images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 15, pp. 1131–1147, 1993.
16. L.D. Cohen, "Multiple contour finding and perceptual grouping using minimal paths," *Journal of Mathematical Imaging and Vision*, Vol. 14, No. 3, pp. 225–236, 2001.
17. L.D. Cohen and Ron Kimmel, "Global minimum for active contour models: A minimal path approach,"

International Journal of Computer Vision, Vol. 24 No. 1, pp. 57–78, 1997.
18. J. Denzler and H. Niemann, "Active rays: Polar-transformed active contours for real–time contour tracking," *Journal on Real-Time Imaging*, Vol. 5, No. 3, pp. 202–213, 1999.
19. E. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, Vol. 1, pp. 269–271, 1959.
20. R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, "Fast geodesic active contours," *IEEE Trans. On Image Processing*, Vol. 10, No. 10, pp. 1467–1475, 2001.
21. M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, Vol. 1, No. 4, pp. 321–331, 1998.
22. R. Kimmel, "Numerical geometry of images: Theory, algorithms and applications," Technion, Israel Institute of Technology, Haifa 32000, Oct. 2000.
23. R. Kimmel and A.M. Bruckstein, "Regularized laplacian zero crossings as optimal edge integrators," *International Journal of Computer Vision*, Vol. 53, No. 3, pp. 225–243, 2003.
24. E. Kreyszig, *Advanced Engineering Mathematics*, 7th edn., W. Anderson, 1993.
25. R. Malladi, J.A. Sethian, and B.C. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 2, pp. 158–175, 1995.
26. S. Osher and J.A Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics*, Vol. 79, pp. 12–49, 1988.
27. J. Sethian, "A fast marching level set method for monotonically advancing fronts," *Proceedings of the National Academy of Sciences*, Vol. 93, No. 4, pp. 1591–1595, 1996.
28. J.A. Sethian, *Level Set Methods and Fast Marching Methods—Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, 1999.
29. C. Sun and S. Pallottino, "Circular shortest path on regular grids," in *Asian Conference on Computer Vision*, Melbourne, Australia, 2002, pp. 852–857.
30. C. Sun and S. Pallottino, "Circular shortest path in images," *Pattern Recognition*, Vol. 36, No. 3, pp. 709–719, 2003.
31. J. Weickert, B.M. ter Haar Romeny, and M.A. Viergever, "Efficient and reliable schemes for nonlinear diffusion filtering," *IEEE Transactions on Image Processing*, Vol. 7, No. 3, pp. 398–410, 1998.
32. C. Xu and J.L. Prince, "Gradient vector flow: A new external force for snakes," in *Proc. IEEE Conf. on Comp. Vis. Pat. Rec. (CVPR)*, Comp. Soc. Press: Los Alamitos, 1997, pp. 66–71.
33. C. Xu and J.L. Prince, "Snakes, shapes and gradient vector flow," *IEEE Transaction on Image Processing*, Vol. 7, No. 3, pp. 359–369, 1998.
34. W. Zhang and R. Korf, "An average-case analysis of branch-and-bound with applications: Summary of results," in *Proc. 10th National Conf. on Artificial Intelligence, AAAI-92*, San Jose, CA, 1992, pp. 545–550.

Ben Appleton received degrees in engineering and in science from the University of Queensland in 2001 and was awarded a university medal. In 2002 he began a Ph.D at the University of Queensland in the field of image analysis. He is supported by an Australian Postgraduate Award and the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Mathematical and Information Sciences. He has been a teaching assistant in image analysis at the University of Queensland since 2001. He has also contributed 10 research papers to international journals and conferences and was recently awarded the prize for Best Student Paper at Digital Image Computing: Techniques and Applications. His research interests include image segmentation, stereo vision and algorithms.

Hugues Talbot received the engineering degree from École Centrale de Paris in 1989, the D.E.A. (Masters) from University Paris VI in 1990 and the Ph.D from École des Mines de Paris in 1993, under the guidance of Dominique Jeulin and Jean Serra. He has been affiliated with the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Mathematical and Information Sciences since 1994. He has worked on numerous applied projects in relation with industry, he has contributed more than 30 research papers in international journals and conferences and he has co-edited two sets of international conference proceedings on image analysis. He now also teaches image processing at the University of Sydney, and his research interest include image segmentation, linear structure analysis, texture analysis and algorithms.