BOUNDARY DETECTION BY MINIMIZING FUNCTIONALS, I

D. Mumford
J. Shah

IEEE COMPUTER SOCIETY REPRINT

BOUNDARY DETECTION BY MINIMIZING FUNCTIONALS, I

D. Mumford
Department of Mathematics
Harvard University
Cambridge, MA 02138

J. Shah
Department of Mathematics
Northeastern University
Boston, MA 02115

Abstract: One approach to segmenting an image into nearly homogeneous regions, separated by smooth boundaries, consists in finding the minimum of a functional which is formed by combining the length of the hypothetical boundary, the gradient of a smoothed version of the image and the difference of the true and smoothed images. We propose a method of finding a pseudo-minimum by solving the Euler-Lagrange equations for the boundaries, under the assumption that the boundaries are not too curved or too close and then using "hill-climbing". The method is worked out here in the simpler case of segmenting an interval into subintervals on which a 1-dimensional signal is roughly constant.

This paper studies a variational approach to locating boundaries in noisy signals. Our approach is a modification of one due to S. and D. Geman [GG] and subsequently developed by J. Marroquin [M]. All these algorithms are closely related to the Ising model of statistical mechanics (see for instance [G-J], §2.3). We may explain these ideas in general as follows. We are given an input signal

$$g: S \longrightarrow \mathbb{R}$$

where

a) S is either an open region R in Euclidean d-space $\mathbb{R}^d$ or the set of integral vectors $L \cap R$ in R (R is usually a large box or the whole space, L is the lattice of integral vectors).

b) the signal g may either take on any real value, or may be restricted to a finite set of values such as ± 1.

Thus if $d = 2$ and S = pixels in an image (or the cones in the fovea), then g is the image itself. The goal is to construct 2 things: a smoothed ideal signal*

$$f: S \longrightarrow \mathbb{R}$$

and a set of boundaries

---

*Yet another variant is where g is given on integral vectors and f is to be reconstructed on all of R: cf. [PVY]

$$B \subset R$$

Here,

c) the ideal signal f may have any real value or be restricted to a finite set of values such as ± 1,

d) if S = R, then f is assumed differentiable on R-B, but may have discontinuities across B,

e) if S = R, then B is a codimension 1, piecewise smooth (hyper)surface in R; if $S = L \cap R$, then B is a union of small rectangular hyperplane pieces running parallel to the coordinates and halfway between the lattice points.

Example: $d = 2$
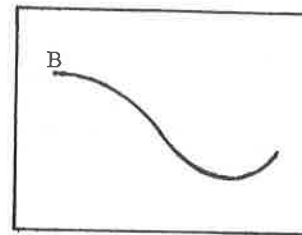
$S = R$
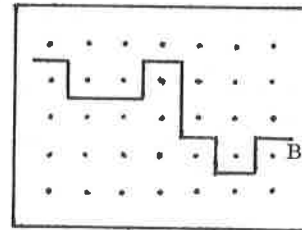
$S = L \cap R$

In the application to image processing, f is a reconstructed clean signal and B is the set of edges in the image.

In order to construct f and B, the procedure is to minimize the energy functional E. If S = R, then E is:

$$E(f,B) = \mu^2 \int_R (f-g)^2 + \int_{R-B} \|\nabla f\|^2 + \mu\nu(\text{volume of } B)$$

and if S = integral vectors in R, then

$$E(f,B) = \mu^2 \sum_{i \in S} (f_i - g_i)^2 +$$

$$\left\{ \begin{array}{c} \sum_{\substack{i,i' \in S \\ i,i' \text{ adjacent} \\ \overline{ii'} \text{ not separated by } B}} (f_i - f_{i'})^2 \end{array} \right\} + \mu\nu(\text{volume of } B)$$

Here μ and ν are parameters, where $\frac{1}{\mu}$ has the dimension of distance and ν has the same dimension as f^2. The volume of B means the $(d-1)$-dimensional volume of B, i.e. if $d = 1$, the number of points in B, if $d = 2$, the length of B and if $d = 3$, the area of B. If we specialize this to the case where f has values $+1$ and assume that B is empty, then up to a constant $\overline{C}$ depending only on g and S, E reduces to:

$$E(f,\phi) = -2\left[\mu^2 \sum_{i \in S} f_i g_i + \sum_{\{i,i' \in S \text{ adjacent}\}} f_i f_{i'} \right] + C$$

which is the energy in the Ising model with external magnetic field g. In the study of the Ising model, one often introduces the boundary B_f consisting of all hyperplane patches separating adjacent vectors i,i' where $f_i \neq f_{i'}$. Then with a constant C' depending only on g and S, the last equation may be rewritten as:

$$E(f,\phi) = -2\mu^2 \sum_{i \in S} f_i g_i + 4(\text{volume of } B_f) + C'$$

The idea of D. and S. Geman was to introduce f and B independently into the functional and minimize, for suitable parameters, with respect to both at once. They studied the case $d = 2$, $S = L \cap R$, R a rectangle, f with a small number (e.g. 4) of values and they also added extra terms for endpoints, corners, "T" and "+" crossings in B. Their method of solution was by a modified Monte Carlo method, the annealing method of Kirkpatrick [KGV]: it involves putting a probability distribution

$$P(f,B) = \frac{1}{Z} e^{-E(f,B)/T} \quad \text{where} \quad Z = \sum_{f,B} e^{-E(f,B)/T}$$

on all choices of f and B, taking long random walks in the space of (f,B)'s at each temperature T and slowly letting $T \to 0$. The Gemans prove that with a sufficiently slow "annealing schedule", this method gives the minimizing f and B with probability 1! J. Marroquin modified their set-up to allow f to have real values. Then for each fixed B, E is a nice quadratic

functional in f and it has a unique solution given by an elliptic boundary value problem (differential equation if $S = R$, difference equation if $S = L \cap R$). He could then solve by a deterministic algorithm for f and do his random walk purely in the space of B's.

Our idea has been to study the case $S = R$ in order to be able to apply infinitesimal variations to B as well as f and derive the corresponding Euler-Lagrange equations. It seems to us that the discretization of B in a lattice approximation makes it hard to choose a B-term in E with the right effect, e.g. for a sufficiently fine lattice, the minimizing f,B should be approximately rotationally invariant. Although the new Euler-Lagrange equations together with the old are now not possible to solve explicitly, there is a limiting case where they can be solved: μ^{-1} very small compared to the scale on which B "doubles back" on itself. Our idea is then to approximately solve for minimizing f,B by (i) solving for B for μ very large, (ii) eliminating parts of B corresponding to local extrema of E which are not local minima, and then (iii) solving for nearly local minima by straight "hill climbing", i.e. small modifications until a local minimum of E is found. In addition to being faster, the method has the advantage that it tells you something about the boundary B and allows you to compare this type of edge detection scheme to others.

This paper presents the case $d = 1$, which works out cleanly and has been implemented by us. On Vax 11/750, it works very fast for an S consisting of 256 lattice points. Work is in progress on the much more interesting case $d = 2$ which applies to images.

One of the motivations of this research is the failure of existing local differential edge operators to give perceptually reasonable edges in some quite simple cases (see [D]). Thus if

$$g(x,y) = \frac{1}{1+(ax)^2+(by)^2}$$

we get an image which, perceptually, is a blurry elliptical white blob on a black field. But if $\frac{a}{b} = 2$, the zero-crossings of the Laplacian do not give a closed contour surrounding the white blob. If $\frac{a}{b} = 4$, the zeroes of the second directional derivative along the gradient (see [TP]), also do not give a closed contour surrounding the white blob. This makes it seem very worthwhile to investigate whether global approaches, such as the present minimization formulation, will give perceptually reasonable contours in more cases.

§1. The Variational Problem

Fix parameters μ and ν. Let $g(x)$ be a given continuous function defined for $a \leq x \leq b$. For every k and every sequence of points

$$a = a_0 < a_1 < \dots < a_k < a_{k+1} = b$$

and every function $f(x)$ continuously differentiable on

each interval $a_i \leq x \leq a_{i+1}$, let

$$E(f,\{a_i\}) = \mu^2 \int_a^b (f-g)^2 dx + \sum_{i=0}^k \int_{a_i}^{a_{i+1}} (\frac{df}{dx})^2 dx + \mu\nu k$$

We assume that $f,\{a_i\}$ minimize E and see what follows. Fixing the a_i, look at the first variation of E in f. Replacing f by $f+\delta f$, where $\delta f(a_i) = 0$, we deduce

$$0 = \frac{1}{2} \delta E = \mu^2 \int_a^b (f-g)\delta f dx - \sum_{i=0}^k \int_{a_i}^{a_{i+1}} f'' \delta f dx$$

Hence, on each interval $[a_i, a_{i+1}]$:

(1) $f'' = \mu^2(f-g)$

If δf has limits $\delta f(a_i^+)$ and $\delta f(a_i^-)$ as $x \to a_i$ from above and below, integration by parts gives us also

$$\delta f(a_i^\pm)f'(a_i^\pm) = 0$$

Hence

(2) $f'(a_i^\pm) = 0$

(1) and (2) are an elliptic boundary value problem. The Green's function for equation (1) on $(-\infty,+\infty)$ is

(3) $K_\mu(x) = \frac{\mu}{2} e^{-\mu|x|}$

Hence

(4) $f_\mu(x) = \int_a^b K_\mu(x-y)g(y)dy$

solves (1) on $[a,b]$. f_μ is a twice differentiable version of the input function g, smoothed on a scale $\frac{1}{\mu}$. The solution f to (1) and (2) is given on each interval $[a_i, a_{i+1}]$ by

(5) $f(x) = f_\mu(x) + C_{1,i}K_\mu(x-a_i) + C_{2,i}K_\mu(a_{i+1}-x)$
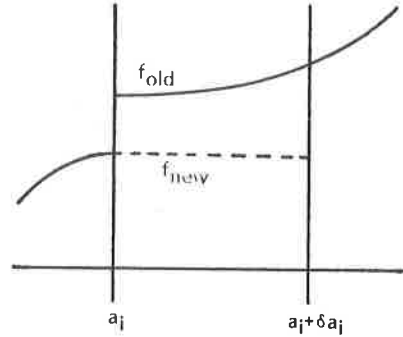
where $C_{1,i}$ and $C_{2,i}$ must be set so that (2) holds. This gives

(6) $C_{1,i} = \frac{2}{\mu^2(1-\alpha_i^2)} (f_\mu'(a_i)-\alpha_i f_\mu'(a_{i+1}))$

$C_{2,i} = \frac{2}{\mu^2(1-\alpha_i^2)} (-f_\mu'(a_{i+1})+\alpha_i f_\mu'(a_i))$

where $\alpha_i = e^{-\mu(a_{i+1}-a_i)}$.

Next vary one of the points a_i. As a_i varies, we imagine f extended so as to be differentiable function on each side of $a_i+\delta a_i$:

Then because $f'(a_i^\pm) = 0$, one sees immediately that

$$\delta E(f,\{a_i\})$$

$$= \int_{a_i}^{a_i+\delta a_i} [(f_{new}-g)^2-(f_{old}-g)^2]dx$$

$$= [(f(a_i^+)-g(a_i))^2-(f(a_i^-)-g(a_i))^2]\delta a_i + O(\delta a_i^2)$$

Thus

(7) $\{f(a_i^+)-f(a_i^-)\}\{f(a_i^+)+f(a_i^-)-2g(a_i)\} = 0$.

If $f(a_i^+)=f(a_i^-)$, then we might as well omit the breakpoint a_i and reduce the value of E. Therefore the variation in a_i gives:

(7') $\frac{1}{2}\{f(a_i^+)+f(a_i^-)\} = g(a_i)$

This condition can be rewritten in a surprising way. Calculating $f(a_i^\pm)$ from (5) and using (6), we get

$$\frac{1}{2}\{f(a_i^+)+f(a_i^-)\} - g(a_i)$$

$$= f_\mu(a_i)-g(a_i) +$$
$$\frac{1}{\mu} [(\frac{\alpha_i^2}{1-\alpha_i^2} + \frac{\alpha_{i-1}^2}{1-\alpha_{i-1}^2})f_\mu'(a_i)$$
$$- \frac{\alpha_i}{1-\alpha_i^2}f_\mu'(a_{i+1}) + \frac{\alpha_{i-1}}{1-\alpha_{i-1}^2}f_\mu'(a_{i-1})]$$

$$= \frac{1}{\mu^2}f_\mu''(a_i) + \frac{1}{\mu} O(\alpha_i,\alpha_{i-1})\sup|f_\mu'(x)|$$

Therefore as $\mu \to \infty$ with fixed $\{a_i\}$, $\mu\alpha_i = \mu e^{\mu(a_i-a_{i+1})} \to 0$ and condition (7') reads:

(7") $f_\mu''(a_i) \sim 0$

Thus the boundaries defined by minimizing E, for μ sufficiently large, will be the zeroes of the second derivative of the smoothed version f_μ of g. These are the same as the edges in most other treatments of 1-dimensional signals.

24

§2. Method of Computation

There are 2 ways to proceed if you want to find minima of E numerically. One is to discretize E itself, i.e. study the functional E on a lattice in the interval [a,b], replacing the integrals by sums and the derivatives by differences as in the Introduction. The other is to use the above analytic formulae for the minimizing function f, but evaluate them by numerical integration. The latter procedure seems to be easier in this case.

The function $f_\mu(x)$ is independent of the breakpoints $\{a_i\}$ and particularly easy to calculate in the 1-dimensional case. Let

$$f_{\mu-}(x) = 2\int_a^x K_\mu(x-y)g(y)dy$$

and

$$f_{\mu+}(x) = 2\int_x^b K_\mu(x-y)g(y)dy$$

so that

$$f_\mu(x) = \frac{1}{2}[f_{\mu+}(x) + f_{\mu-}(x)]$$

and

$$f_\mu'(x) = \frac{\mu}{2}[f_{\mu+}(x) - f_{\mu-}(x)]$$

$f_{\mu-}(x)$ satisfies the differential equation

$$f_{\mu-}'(x) = \mu[g(x)-f_{\mu-}(x)]$$

with the boundary condition $f_{\mu-}(a) = 0$ while $f_{\mu+}(x)$ satisfies the equation

$$f_{\mu+}'(x) = -\mu[g(x)-f_{\mu+}(x)]$$

with the boundary condition $f_{\mu+}(b) = 0$.
These first order differential equations may be solved very rapidly by step-by-step integration.

We divide the interval [a,b] into n equal subintervals. (In our examples, n = 255.) This gives us a lattice of n+1 points $\{x_j\}$ defined as $x_j = a + jh$ where $0 \leq j \leq n$ and $h = (b-a)/n$. The breaks $\{a_i\}$ are assumed to occur between lattice points and are treated as the entire open interval between 2 adjacent lattice points rather than as specific points somewhere in between. Thus a set of breaks is defined by integers $0 \leq j_1 < \ldots < j_k < n$; the corresponding breaks occur between the lattice point pairs $\{x_{j_1}, x_{j_1+1}\}$, $\ldots$,$\{x_{j_k}, x_{j_k+1}\}$. For each interval $[x_{j_i+1}, x_{j_{i+1}}]$ we calculate

(i) α_i, $C_{1,i}$, $C_{2,i}$, and $f(x_j)$ for $j_i < j \leq j_{i+1}$ by setting $a_i = x_{j_i+1}$ and $a_{i+1} = x_{j_{i+1}}$ in equations (5) and (6) and

(ii) the value of the functional E restricted to the interval using the trapezoidal rule.

We solve for minimizing breaks by the following interative procedure. Start with a set of breaks as initial values, usually chosen from the zero-crossings of $f_\mu''(x)$ or which are the same, zero-crossings of $f_\mu-g$. During each iteration, calculate new values of j_i's by varying each j_i in turn while keeping the others fixed at their values from the last iteration and choose a value for j_i which gives maximal reduction in the functional E. We fix the size of maximum allowable change in the integers $\{j_i\}$ during a single iteration. We found this method more stable than solving equation (7') directly by some iterative procedure.

§3. Examples:

We give two examples of the above algorithm. In both cases R is an interval with 256 pixels in it and μ^{-1} is taken to be 6 pixels in length. We have not taken any specific value of ν, but have estimated the minimum with various numbers of breakpoints.

The first example is a Gaussian curve with 4 breakpoints. Figure 1 shows the Gaussian g and the minimizing f. This figure seems to give a good general idea of what the minimizing f's tend to look like with very smooth input g.

The second example is a horizontal line through a digitized image of a person. The image is shown in figure 2 together with the chosen line which passes through her eyes. Figure 3 shows the graph of the actual image along this line (this is the function g) and the smoothed version f_μ of g. Figure 4 shows our program's estimate of the minimizing f with 10 breakpoints. The breakpoints separate her hair from the grey background and from her cheek, bracket her eyeballs and bracket the specular highlight on her nose. Because of the noise in the original image, especially in the background, f_μ'' has a great many zeroes in this case. We first pruned these eliminating zeroes where $|f_\mu'|$ did not have a strong maximum and then choosing as initial breakpoints random sets of 10 from those that remained. The iterative scheme described above usually settles down to one of a small number of configurations all of which have energy near each other. However, these near-optimal breakpoint configurations can look quite different from each other: for instance, 2 breakpoints may settle down on the 2 edges of a steep hill or valley in f, but in different configurations, they may settle down on different hills or valleys. The possibility of several very distinct local minima with almost the same near minimal energy is reminiscent of the non-uniqueness of the ground state in the original Ising model.

To analyze the dependence on the number of breakpoints, we repeated the above with the same image but all breakpoint numbers between 1 and 10. Figure 5 shows the graph of minimum energy found against breakpoint number. The vertices of the convex hull of this graph represent configurations which, for a suitable choice of ν, are optimal over all sets of breakpoints of any cardinality.

References

[GG] S. and D. Geman, Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, Preprint, Brown Dept. of Appl. Math.,1983.

[M] J. Marroquin, Surface Reconstruction Preserving Discontinuities, MIT AI Lab Memo 792, 1984.

[D] J.G. Daugman, Principles of Visual Neuronal Receptive Field Organization, IEEE Trans. on Systems, Man and Cybernetics, 1984.

[G] J. Glimm and A. Jaffe, Quantum Physics, Springer-Verlag, 1981.

[KGV] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by Simulated Annealing, preprint, IBM Thomas Watson Research Lab., Yorktown Heights, NY, 1982.

[PVY] T. Poggio, H. Vorhees and A. Yuille, Regularizing Edge Detection, MIT AI Lab Memo 776, 1984.

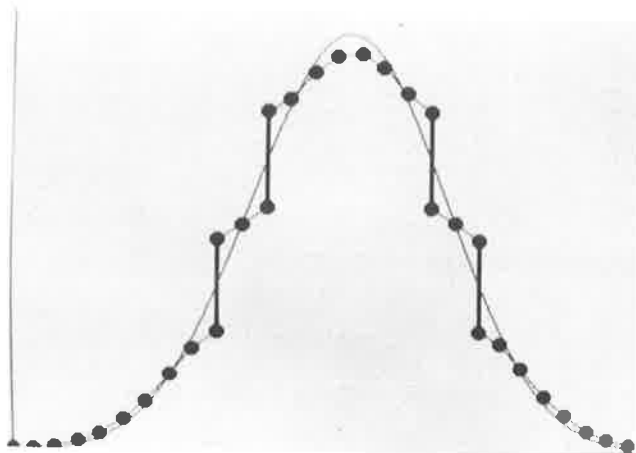[TP] V. Torre and T. Poggio, On Edge Detection, MIT AI Lab Memo 768, 1984.
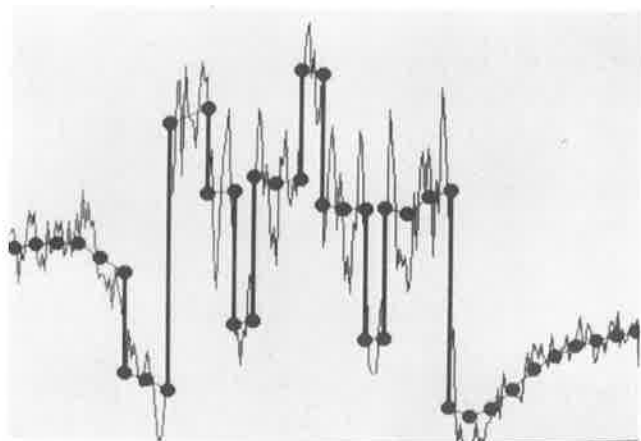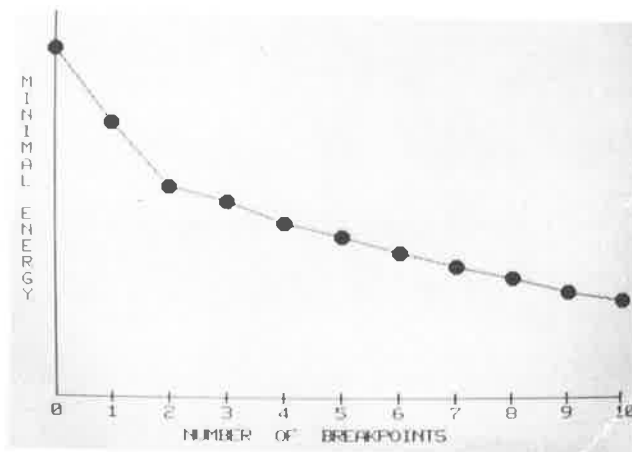
Figure 1

Figure 2

Figure 3

Figure 4

Figure 5