tion of the discontinuity sets of interest.
Finally we should comment on the assumption (7.7). It ought to be justified on its own merits rather than simply being a necessary condition for our main conclusions to hold. Indeed, smooth surface patches in the scene generically intersect along smooth curves, whose projections onto the image plane generically are smooth. Hence the edges in the "true" image should be expected to be piecewise smooth for just about the same reasons as the "true" image function, which we are trying to estimate, is expected to be piecewise smooth. This calls for mechanisms analog to those governed by the smoothness and edge cost terms in the cost (7.1). In this context condition (7.7) is merely a straight forward analog of the edge cost. The action of the smoothness term is roughly taken care of by imposing the bound, ($\rho < \infty$,) on the Hölder modulus of the derivatives of the components of the image segmentation components γ.

APPROXIMATION, COMPUTATION, AND DISTORTION IN THE VARIATIONAL FORMULATION

Thomas Richardson
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

and

Sanjoy Mitter [1]

Center for Intelligent Control Systems
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

1. Introduction

Mumford and Shah [263][264] suggested performing edge detection by minimizing functionals of the form

$$E(f,K) = \beta \int_{\Omega} (f-g)^2 dx + \int_{\Omega - K} |\nabla f|^2 \, dx + \alpha \, |K| \, ,$$

where Ω is the image domain (a rectangle), dx denotes Lebesgue measure, g is the observed grey level image, i.e., a real valued function, f approximates g, K denotes the set of edges (a closed set), $|K|$ is the total length[2] of K, and β and α are real positive scalars. This approach is a modification of one due to Geman and Geman [132] that uses Markov random fields, which

[1] Research supported by the US Army Research Office under grant ARO DAAL03-92-G-0115 (Center for Intelligent Control Systems).

[2] In general K cannot be represented as a union of curves so one-dimensional Hausdorff measure is used to define 'total length.' See the contribution of Leaci and Solimini in this book.

was developed by Marroquin [249] and by Blake and Zisserman [34]. It is referred to as the variational formulation of edge detection.

The three terms of E 'compete' to determine the set K and the function f. The first term penalizes infidelity of f to the data while the second term forces smoothness of the approximation f, except on the edge set K. Thus, f will be a piecewise smooth approximation to g. The third term forces some conservativeness in the use of edges by penalizing their total length.

The primary difficulty in constructing an efficient algorithm to minimize E is appropriately representing the edges. One approach is to absorb the edges into the interaction between neighboring pixels. This idea appears in the anisotropic diffusion approach [293] [273], GNC (Graduated non-convexity) type algorithms [34], and in mean field annealing [33][129]. Another approach has emerged from the cross fertilization between computer vision and mathematical physics. The variational formulation, besides being a model for edge detection, now serves as a prototypical example of a 'free-discontinuity' problem. (See the contribution of Leaci and Solimini for an elaboration of this.) Within the framework developed for such problems the theory of Γ-convergence, approximation of one functional by another, has also been developed. Ambrosio and Tortorelli [12, 13] applied that theory to E. The approximation is achieved primarily by replacing the edge set with a function that modulates the smoothing of the image. Additional terms in the functional force this function to behave as if it were a smeared version of the corresponding edge set. The degree of smearing, i.e., the width of the effective edges, is controlled by a parameter. Convergence of the approximation to E occurs by taking the parameter to the appropriate limit, causing the effective edge width to tend to zero.

A benefit of replacing the edge set with a smooth function is that an obvious algorithmic approach suggests itself: discretize the functions and minimize the functional using descent methods. The product is an edge detector that can be represented as a coupled pair of non-linear partial differential equations (see Section 3). The idea of using coupled partial differential equations of this type is now being applied to many problems in computer vision. The chapter in this book by Proesmans, Pauwels, and Van Gool contains several examples. March [244] applied this approach to the stereo matching problem. In Section 2, we will outline the application of Γ–convergence to the variational formulation of edge detection and show how it leads to a coupled pair of partial differential equations.

The variational formulation was motivated in part by the desire to combine the processes of edge placement and image smoothing. Earlier edge detection techniques such as the Marr-Hildreth edge detector, the Canny edge detector, and their variants separated these processes; the image is first smoothed to suppress noise and control the scale, and edges are detected

subsequently, as gradient maxima, for example. One consequence of this two step approach is pronounced distortion of the edges, especially at high curvature locations. Corners tend to retract and be smoothed out; the connectedness of the edges at T-junctions is lost. By introducing interaction between the edge placement and the smoothing, it was expected that this effect could be abated. There is evidence, both theoretical, in one dimension (see [34]), and experimental, in two dimensions, that this is indeed the case. However, the model is known to place undesirable restrictions on admissible edge geometries (see Section 5). Nevertheless, certain limit theorems that have been proven by one of us in [306] and are discused in Section 5 indicate that the restrictions may not be too serious and that, asymptotically, any edge geometry is possible. Moreover, heuristic arguments suggest that the approximate formulation indicated above behaves better with respect to these distortions/restrictions than the original model. The relaxation of the distortion is achieved at the cost of the smearing of the edges. Hence, there is trade-off here between the resolution of the edges and the systematic distortion of the model.

Localization of edges cannot be reasonably discussed without also making reference to scale. The notion of 'scale', scale of features and scale of representation, is widely held to be of fundamental importance in vision. (This book contains two substantial chapters devoted to 'scale-space.') One reason for this is that hierarchal descriptions of scenes offer potential reductions in complexity of various visual processing tasks. Coarse scale segmentation of an image, for example, can be used to identify regions of interest for further processing, thereby reducing the computational load. It is important, therefore, that coarse scale descriptions retain those features of the data that are required for effective decision making. In the case of edge detection, T-junctions and corners play important roles in estimating the depth and shape of objects in a scene [126]. It is desirable, therefore, to accurately represent these features even at coarse scales. The 'finger-print' images of gradient maxima of one dimensional images in scale-space [386][395] are well known; the localization of edges degrades badly as scale increases. Many two dimensional examples can be found in the literature.

By embedding the ideas implicit in the limit theorems mentioned above into the approximation scheme also mentioned above, one can develop an edge focusing scheme that essentially removes the restrictions on the edge geometry present in the original model and, at the same time, circumvents the smearing/geometry trade-off to produce well localized, sharp edges. We indicate how this is done in Section 6. A complete description can be found in [307]. The resulting algorithm is described by a coupled set of non-linear second order parabolic partial differential equations (eqns. (8.5)–(8.8)) with explicit parameters β and c which are adjusted in an appropriate way (see

eqns. (8.9)–(8.11)). The adjustment induces focusing of the edges. The global coarse scale nature of the edges is retained by introducing scale stabilizing feedback mechanisms. The adjustment process commences after the non-linear parabolic equations have nearly converged to their equilibrium. The set of equations (8.5)–(8.8) and (8.9)–(8.11) should be viewed as an adaptive non-linear filter which performs edge detection via focusing. Indeed, the equations are the fundamental objects in this theory and, apparently, are far more well behaved (for example, convergence to global minima) than the original variational problem. It is apparent that the form and the properties of the Γ-convergent approximation mesh well with the parameter adjustment proposed for the edge focusing algorithm. In particular, the relaxation of the edge geometries due to the smearing of the edges is retained, by adjusting the parameters, while the edges are sharpened

2. Approximation via Γ–Convergence

To compute minimizers of E the critical question is how to represent the set K. This issue was raised in the section of this book by D. Mumford. A natural approach is to discretize K into "edge elements" and treat them combinatorially, adding or removing elements in an attempt to minimize E. Appending a stochastic component, based on a Markov random field model, leads to the simulated annealing approach first suggested in [132]. This tends to produce computationally impractical algorithms. Modifications which incorporate the edge elements into the interaction between image pixels have been proposed. One of these is based on mean field approximations of the Markov random field [128] [33] and another, GNC [34], is based on a homotopy of the interactions. Both these approaches have their strong points, and are in fact quite similar [33] [129]. A novel and powerful approach has appeared from the mathematical theory of approximation of functionals via Γ-convergence.

The concept of Γ-convergence is due, independently, to E. De Giorgi [84] and H. Attouch [23]. The idea is to approximate one functional, E for example, by more regular ones, E_c, so that minimizers of E_c approximate minimizers of E while enjoying greater regularity. For the variational formulation of edge detection, one would like to replace the edges, which are singular objects (in the context of 2-dimensional measure), with something more manageable. In this section, we provide a definition of Γ-convergence, state some of its basic properties, and present the application to the variational formulation of edge detection.

Let (S, d) be a separable metric space and let $F_n : S \to [0, +\infty], n = 1, 2, ...$ be a sequence of functions. We say this sequence $\Gamma(S)$ – converges to F :

$S \to [0, +\infty]$ if the following two conditions hold for all $f \in S$,

$$\begin{aligned} \forall f_n \to f \quad & \liminf_{n\to\infty} F_n(f_n) \geq F(f) \\ \text{and } \exists f_n \to f \quad & \liminf_{n\to\infty} F_n(f_n) \leq F(f)\,. \end{aligned}$$

The limit F, when it exists, is unique and lower–semicontinuous. The following theorem characterizes the main properties of Γ–convergence.

Theorem (Γ–Convergence.) Assume that $\{F_n\}$ $\Gamma(S)$–converges to F. Then the following statements hold.
(i) Let $t_n \downarrow 0$. Then, every cluster point of the sequence of sets

$$\{f \in S : F_n(f) \leq \inf_S F_n + t_n\}$$

minimizes F.
(ii) Assume that the functions F_n are lower semicontinuous and, for every $t \in [0, \infty)$, there exists a compact set $C_t \subset S$ such that for all n $\{f \in S : F_n(f) \leq t\} \subset C_t$. Then the functions F_n have minimizers in S, and any sequence of minimizers of F_n admits subsequences converging to some minimizer of F.

The point of (i) is that approximate minimizers of F_n approximate minimizers of F. Condition (ii) is useful for proving that minimizers of F_n exist and are well behaved.

To properly formulate the Γ–convergence results it is convenient to define E in terms of f alone, letting K be implicitly defined as the closure of the discontinuity set of f. This implicit definition is described in the chapter by Leaci and Solimini in this book.

We consider functionals of the form

$$E_c = \int_\Omega (\beta(f-g)^2 + \Phi(v) \mid \nabla f \mid^2 + \alpha(c\Psi(v) \mid \nabla v \mid^2 + (1-v)^2/4c))\ dx\,. \tag{8.1}$$

Here $\Phi(v)$ is playing the role of the K in E, i.e., it is modulating the smoothness constraint on f. The other terms involving v force $\Phi(v)$ to simulate the effect that K has in E. Implicitly, we have $0 \leq v \leq 1$. The algorithmic intention is to minimize E_c with respect to f and v. An obvious advantage the approximation offers over the original formulation is that v, since it is a function on Ω, can be discretized in a straightforward way and (local) minimizers of E_c can be computed using descent methods. Ambrosio and Tortorelli [13] proved that if one sets

$$\Phi(v) = v^2 \quad \text{and} \quad \Psi(v) = 1\,, \tag{8.2}$$

then E_c Γ-converges to E as $c \to 0$, i.e., for any sequence $c_n \to 0$. Some computational results based on this functional have already appeared in March [245], see also Shah [341].

The choice for Φ and Ψ given above may be one of the simplest possible, but it is far from unique. (See [12] for an example which is not of the form 8.1.) When one considers algorithms based on these functionals there are trade-offs to be made between speed and performance. For example, the choice reflected in equation (8.2) leads to simple equations and fast computation. However, other choices may produce sharper singularities in Φ and hence less smearing of f near the edges. With slight modifications, the proof of Γ-convergence found in [12] and [13] can be made to go through for a large class of Ψ and Φ. In particular, one can choose Ψ to be any C^1 function satisfying

$$\Psi(x) > 0 \text{ for } x \in (0,1],$$
$$2\int_0^1 (1-u)\Psi^{1/2}(u)du = 1.$$

Note that any C^1 function satisfying the first property can be made to satisfy the second property by suitable normalization. Given such a Ψ, one can choose Φ to be any C^1 function satisfying

$$\Phi(1) = 1,$$
$$\Phi(0) = 0,$$
$$\Phi(x) \in (0,1) \text{ for } x \in (0,1).$$

Although the conditions given above are sufficient for the proof of Γ-convergence, for algorithms based on 'gradient' descent on E_c one should also impose the condition that Ψ be monotonically non-decreasing and Φ be monotonically increasing on $(0,1)$. Furthermore, for our implementation, which is discussed in Section 7, the condition $\lim_{x\to 0} \dot{\Phi}(x)/x < \infty$ should be imposed.

Even more general Ψ and Φ than defined above are possible. For example, setting

$$\Psi(v) = \Phi(v) = \frac{1}{2}e^{-(1-v)^2} \tag{8.3}$$

also produces a Γ-convergent set of functionals. Examples in the class defined above are

$$\Phi(v) = v^{2n} \quad \text{and} \quad \Psi(v) = \frac{(m+1)^2(m+2)^2}{4}v^{2m}, \tag{8.4}$$

where $m \geq 0$ and $n > 0$. Equation (8.2) is a special case of this with $(n,m) = (1,0)$.

To formulate the Γ-convergence of E_c to E, one must find an appropriate metric space for f and v and extend the definition of E and E_c to this space. In [13], the choice of S is

$$(f,v) \in L^\infty(\Omega) \times \{v \in L^\infty(\Omega) : 0 \leq v \leq 1\},$$

and the metric is that induced by the $L^2(\Omega)$ norm. (In [12], slightly different choices were made.) First, it is possible to mathematically recast E (in a weak setting with $f \in \mathrm{SBV}(\Omega)$, see the chapter by Leaci and Solimini), whereby one defines K in terms of the discontinuities in f. In this way one can write $E(f,K) = E(f)$. The functional E can then be extended to S by setting $E(f,v) = E(f)$ if $v = 1$ (in $L^2(\Omega)$) and $f \in \mathrm{SBV}(\Omega)$ and $E(f,v) = \infty$ otherwise. Some additional care must be taken to define E_c on all of S. When ∇v and ∇f are not appropriately well defined, one should set $E_c(f,v) = \infty$; we refer the reader to [12] or [13] for technical details. Once the metric space is specified, the proof of Γ-convergence involves basically two steps. The first is to show that for any sequence $\{f_{c_i}, v_{c_i}\}$ where $c_i \to 0$, $f_{c_i} \to f$, and $v_{c_i} \to 1$ (in the appropriate sense, not pointwise) that $\liminf_{i\to\infty} E_{c_i}(f_{c_i}, v_{c_i}) \geq E(f,K)$. The second is to construct a sequence such that $\limsup_{i\to\infty} E_{c_i}(f_{c_i}, v_{c_i}) \leq E(f,K)$. If (f, K) minimizes E, then this second step requires constructing near minimizers of E_c. If

$$\liminf_{i\to\infty} E_{c_i}(f_{c_i}, v_{c_i}) < \infty,$$

then, roughly speaking, $x \in K$ implies $\lim_{i\to\infty} \Phi(v_{c_i}(x)) = 0$. (Thus at these points we do not have $v_{c_i}(x) \to 1$; however, K is a set of Lebesgue measure 0.) On the other hand, the last term in equation (8.1) forces $v_c(x)$ to converge to 1 for almost all $x \in \Omega$ (in the sense of Lebesgue measure), hence one has $\lim_{c\to 0} \Phi(v_c(x)) = 1$ for almost all $x \in \Omega$. The near minimizers of E_c are constructed by setting $\Phi(v_c(x)) \simeq 0$ on K and $\Phi(v_c(x)) \simeq 1$ outside some neighborhood of K, with a smooth transition in between. The approximations indicated here become equalities in the limit as $c \to 0$. The width of the transition depends on Ψ and on c. We give a brief heuristic description of how this occurs.

In the transition region of $\Phi \circ v$, we expect f to be relatively smooth, so only the terms not involving f in E_c will have a significant effect on the form of v there. In the following inequality,

$$c\Psi(v) \mid \nabla v \mid^2 + \frac{(1-v)^2}{4c} \geq \Psi^{1/2}(v) \mid \nabla v \mid (1-v),$$

equality holds only if $\mid \nabla v \mid = \Psi^{-1/2}(v)\frac{1-v}{2c}$. This suggests that (in one dimension) if $u_c(t)$ satisfies $\frac{\partial u_c(t)}{\partial t} = \frac{1-u_c(t)}{2c}\Psi^{-1/2}(u_c(t))$ with $\Phi(u(0)) \simeq 0$,

then setting $v(x) = u_c(\text{dist}\,(x,\ K))$, for dist $(x,\ K) \leq \tau_c$ where $\Phi(u_c(\tau_c)) \simeq 1$ (with $u_c(\tau_c) \to 1$ as $c \to 0$), will produce near optimal transitions. This is how the near optimal v_c are constructed in [12] and [13]. Note that assuming that $u_c(0)$ does not depend on c, we obtain $u_c(t) = u_1(t/c)$. Thus the edge width is proportional to c. Let $G(s) = \int_0^s (1-r)\Psi^{1/2}(r)dr$. We now compute

$$\begin{aligned}\int_0^{\tau_c} \left(c\Psi(u(t)) \mid \frac{\partial u_c(t)}{\partial t} \mid^2 + \frac{(1-u(t))^2}{4c} \right) dt &= \\ \int_0^{\tau_c} \left(\Psi^{1/2}(u(t)) \mid \frac{\partial u_c(t)}{\partial t} \mid (1-u(t)) \right) dt &= \\ \mid \int_0^{\tau_c} \frac{\partial}{\partial t} G(u(t))dt \mid = G(\tau_c) &\simeq \frac{1}{2}\end{aligned}$$

with the last approximate equality becoming equality in the limit as $c \to 0$. In the one dimensional case, we now see that the last term in 8.1 will contribute approximately α times the number of discontinuity points of f. In two dimensions, one obtains approximately α times the length of K. Thus, we see that E_c approximates E.

3. Minimizing E_c

Local minimers of E_c can be found by gradient descent. Simple, practical algorithms can be obtained from finite element discretizations. In this section, we formulate the main ideas in the continuum setting.
The Euler-Lagrange equations associated with E_c are given by $\partial_v E_c = 0$ and $\partial_f E_c = 0$ with Neumann boundary conditions, where

$$\partial_f E_c \triangleq \beta(f-g) - \nabla \cdot (\Phi(v)\nabla f)\,, \tag{8.5}$$

$$\partial_v E_c \triangleq \dot{\Phi}(v)\,\alpha^{-1} \mid \nabla f \mid^2 - c\nabla \cdot (\Psi(v)\nabla v) + 2c\dot{\Psi}(v) \mid \nabla v \mid^2 + (1-v)/2c \tag{8.6}$$

are the functional derivatives of E_c.
Allowing f and v to depend on t, we can write a 'gradient' descent on E_c in the form

$$\frac{\partial}{\partial t} f(x,t) = -c_f \partial_f E \tag{8.7}$$

$$\frac{\partial}{\partial t} v(x,t) = -c_v \partial_f E \tag{8.8}$$

with Neumann boundary conditions, where c_f and c_v control the rates of descent; they would be constant for a strict gradient descent, but may not be in general. It turns out that better implementations (faster and with more stable convergence) can be obtained with c_v not held constant (see Section 7).
Since the functional E_c is not jointly convex in v and f, we do not expect to always reach a global minimum by a descent method. Thus the solution obtained will depend on the initial conditions and also on the parameters c_f and c_v.
Equation (8.7) (after substituting from equation (8.5)) strongly resembles the anisotropic diffusion scheme for image enhancement introduced by Perona and Malik [293] and, even more strongly resembles, the 'biased' anisotropic diffusion scheme of Norström [273]. The solution to the diffusion equation

$$\frac{\partial}{\partial t} f(x,t) = \Delta_x f(x,t), \quad f(x,0) = g(x)$$

(with Neumann boundary conditions) is identified with 'scale-space' smoothings of g, parametrized by t. The solution $f(x,t)$ is obtained by convolving $g(x)$ with a Gaussian kernel whose variance is linear in t. Perona and Malik [293] suggested performing image enhancement by controlling the diffusion coefficient to prevent smoothing across edges. Thus they were led to consider equations of the form

$$\frac{\partial}{\partial t} f(x,t) = \nabla_x \cdot (h(\mid \nabla_x f(x,t) \mid)\nabla_x f(x,t)), \quad f(x,0) = g(x)\,.$$

They experimented with $h(s) = \frac{J}{1+(\frac{s}{K})^2}$ and $h(s) = e^{-(\frac{s}{K})^2}$, where J and K are constants. Equation (8.7) resembles this equation in that it is a diffusion with controlled conductivity. The control of the conductivity depends on $\mid \nabla f \mid$ indirectly through equation (8.8). The term $\beta(f-g)$ in equation (8.7) stabilizes the solution at some particular scale. Such a term also appears in the 'biased' anisotropic diffusion scheme studied in [273]. In [293], the authors analyze their scheme to show that the maximum principle holds, i.e., that the solution's extrema never exceed those of the original image[3]. They argue that this implies that no new 'features' (blobs) are introduced into the solution. Here, as in [273], this property is a trivial consequence of the formulation. The functionals E_c would *increase* if such new features appeared. (Truncating such a new feature would decrease E_c.)
An advantage of the scheme represented by equations (8.7) and (8.8) is that it yields an explicit representation of the edges (via the function $\Phi(v)$). The resulting system of equations admits a particularly simple implementation in digital mesh connected parallel machines with simple processors or,

[3] It is unclear that the maximum principle can be invoked since an appropriate existence theorem for the Perona-Malik equation has not been proved.

potentially, in an analog network, such as discussed in [149]. Although a direct study of the existence, uniqueness, and well-posedness of (8.7) and (8.8) has not been carried out, local stability of these equations can be proved [305].

4. Remarks on Energy Functionals, Associated Non-Linear Diffusions and Stochastic Quantization

The methodology of minimizing energy functionals involving an "elliptic" form and a "geometric" term (leading to free discontinuity problems) for filtering and boundary reconstruction of images is quite general. For example, a modified energy functional which does not have some of the disadvantages of the functional considered in this chapter is

$$\begin{aligned} E(f,K) &= \int_\Omega (f-g)^2 dx + \int_{\Omega/K} |\nabla f|^2 dx + \int_K |H^2| d\mathcal{H}^1 \\ &\quad + \text{No. of singular points in } K \text{ with multiplicities,} \end{aligned}$$

where H is the curvature of K (appropriately defined), and $\mathcal{H}^1$-represents the one-dimensional Hausdorff measure. This class of problems has been considered in the recent work of Ambrosio and Mantegazza (cf. thesis of Mantegazza at the University of Pisa in 1993). If an elliptic approximation to this class of free-discontinuity problems can be found, then the method of computation of minimizers via non-linear parabolic equations can in principle be constructed.

There is a connection between the ideas presented in this chapter and the ideas of stochastic quantization and the Bayesian formulation of Image Analysis (cf. Mitter [255] and the chapter of Mumford in this book). For this purpose consider

$$\begin{aligned} \exp(-E_c) &= \exp(-\beta \int_\Omega (f-g)^2 dx) \cdot \\ &\quad \exp(-\int_\Omega [\Phi(v)|\nabla f|^2 + \alpha(c\psi(v)|\nabla v|^2 + \frac{(1-v)^2}{4c})]\, dx) \\ &\triangleq \exp(-\beta L(f,g)) \exp(-\Lambda_c(f,v)) . \end{aligned}$$

Then it is necessary to give meaning to the expression

$$d\mu_c = \exp(-\Lambda_c(f,v)) \prod_{x \in \mathbb{R}^2} d(f(x), v(x))$$

as a probability measure on an appropriate distribution space. This probability measure is formally interpreted as a prior measure on (f,v)

while the expression $\exp(-\beta L(f,g))$ is formally interpretated as a likelihood function.

The minimization of the energy functional E_c corresponds to computing the maximum a posteriori probability estimate of (f,v), but the probabilistic interpretation via the measure μ_c and the likelihood function allows one to compute other estimates such as the conditional mean estimate. Finally, the connection to stochastic quantization can be made by considering the coupled infinite dimensional stochastic differential equations

$$df(t,x) = -\partial_f E_c dt + d\xi(t,x)$$

$$dv(t,x) = -\partial_v E_c dt + d\eta(t,x) ,$$

where

$$\partial_f E_c \text{ and } \partial_v E_c$$

denote functional derivatives as above and ξ and η are infinite dimensional Brownian motions. Formally, μ_c is the invariant measure of this coupled pair of stochastic differential equations.

5. Scale, Noise, and Accuracy

The notion of scale-space representation of an image is a central one in this book. With regard to edge detection one of the central problems arising from the scale-space concept is the correspondence problem: which of the fine scale edges correspond to coarse scale edges? The problem is aggravated by the fact that the distortion of the edges, mentioned earlier, depends on scale. Typically 'coarse scale' implies more smoothing and, hence, more distortion. This is undesirable in many situations because salient features, corners and T-junctions for example, tend to be obscured. The correspondence problem is therefore of great importance.

Since the variational approach combats the distortion caused by smoothing, one hopes that the correspondence problem will be alleviated by using it. Although this appears to be the case, problems remain; there still are distortions, depending on scale, and the model intrinsically restricts the geometry of possible edge sets in an unnatural way. The analysis of Mumford and Shah [263] showed that edge sets produced by the variational approach have the following properties, which are illustrated in figure 8.1.

If K is composed of $C^{1,1}$ arcs, then

- at most three arcs can meet at a single point and they do so at 120°,
- they meet $\partial\Omega$ only at an angle of 90°,
- it never occurs that exactly two arcs meet at a point (other than the degenerate case of two arcs meeting at 180°), i.e., there are no corners.

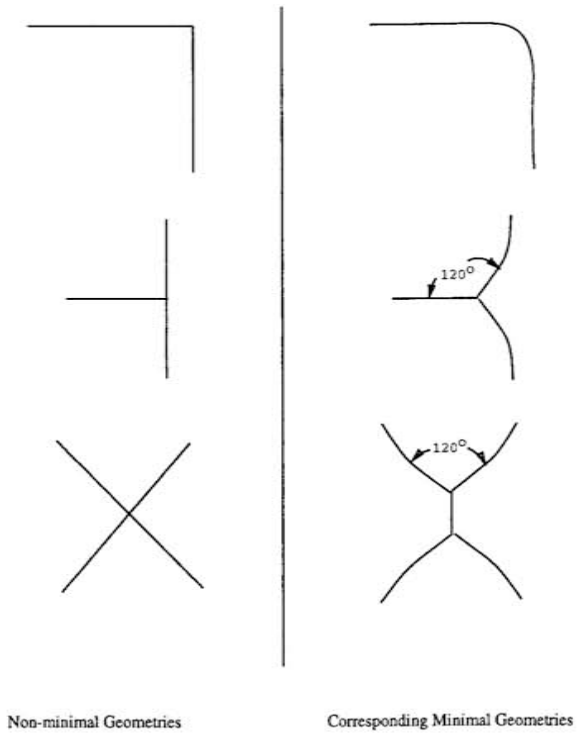

Figure 8.1. Calculus of Variations Results

These results are a consequence of the fact that the term $|K|$ in E locally dominates the behavior of singularities in K. Hence, the types of singularities observed are identical to those of minimal surfaces. Of course 'real' edges are not restricted to these geometries. The dependence on scale derives from the interaction between the singularities of f and those of K. Roughly speaking, smaller values of β produce greater distortion. This becomes more clear in light of the theorem quoted below and the analysis of Mumford and Shah [264]. It turns out that under some mild assumptions on the data, it is possible to prove that the edge set produced by the variational formulation can approximate arbitrarily well, depending on the parameters, any edge geometry.

To quantify the disparity between one edge set and another, we introduce the Hausdorff metric. For $A \subset \mathbb{R}^2$, the ϵ-neighborhood of A will be denoted by $[A]_\epsilon$ and is defined by $[A]_\epsilon = \{x \in \mathbb{R}^2 : \inf_{y \in A} |x - y| < \epsilon\}$ where $|\cdot|$ denotes the Euclidean norm. Denoted by $d_H(\cdot,\cdot)$, the Hausdorff metric is evaluated by

$$d_H(A,B) = \inf\{\epsilon : A \subset [B]_\epsilon \text{ and } B \subset [A]_\epsilon\} .$$

Elementary analysis shows that d_H is a metric on the space of non-empty compact sets in R^2.

Suppose for the moment that we have ideal data: g is a piecewise smooth function. To make this clear we denote it by g_I. We assume that there exists a set K_g, a union of curves, satisfying length $(K_g) < \infty$ such that g_I is discontinuous on K_g and smooth elsewhere. More precisely, we require that $\int_{\Omega - K_g} | \nabla g_I |^2 \, d\mu < \infty$, that there exists a constant L such that g_I, restricted to any straight line segment lying in $\Omega - K_g$, is a Lipschitz function with Lipschitz constant L, and that g actually has a discontinuity everywhere on K_g except, possibly, for a set having zero total length. Under these conditions it can be proved that minimizers of E have the property that the edge sets converge to Γ_g in Hausdorff metric as $\beta \to \infty$. This means, in particular, that edge sets can assume arbitrary geometries asymptotically. This result can be interpreted as an asymptotic fidelity result for the variatio- nal approach. It implies that the distortions resulting from ad hoc functions of this type are local, small scale effects.

From a practical point of view this convergence is inadequate because $\beta \to \infty$ forces f to match g_I exactly; noise in g_I will result in the appearance of many spurious boundaries. However, noise and smearing effects can be incorporated into the result if they are scaled appropriately. Roughly speaking, if the admissible noise magnitude scales as $o(\beta^{-\frac{1}{2}})$ and the admissible smearing acts over a radius of $o(\beta^{-1})$, then their presence can be tolerated and we have the following,

Theorem (Approximation.) *For any fixed $\alpha > 0$ and $\epsilon > 0$ there exists $\beta^\star < \infty$ such that if $\beta \geq \beta^\star$ and K_β is minimal for E for some $g \in \Psi(\beta)$, then*

$$d_H(K_g, K_\beta) < \epsilon \, .$$

The proof may be found in [306] and [305]. The relevant mathematical framework is outlined in [11]. This theorem indicates how noise and localization defects should scale with the parameter β to maintain fidelity.

6. Edge Focusing via Scaling

In [307] the theory outlined above is applied to the Γ-convergent approximation to the variational formulation to produce an edge-focusing algorithm. The basic idea of edge-focusing is to start with coarse scale edges and then adjust them to get better localization without introducing finer scale edges. It is an attempt to circumvent the noise/accuracy/scale tradeoffs inherent in the underlying edge detection model.

Bergholm [31] developed an edge focusing algorithm based on the Marr-Hildreth edge detector. Since the variational model has inherently better localization, the correction required by the focusing algorithm is smaller. Prior to the existence of the theory outlined in previous sections, the main barrier to edge focusing based on the variational formulation was the computational difficulty of implementing such a scheme.

The edge focusing algorithm developed in [307] is based on the scaling conditions of the Approximation Theorem and on the Γ-convergent approximation to the variational formulation. Two problems immediately suggest themselves with regard to applying the Approximation Theorem. First, a real image has fixed noise which cannot be scaled since it cannot be identified, and second, smearing is fixed and cannot (in general) be removed. The algorithm resolves these objections by making a heuristic identification of noise with error and of distortion of the edges with smearing. A minimizer of the variational formulation provides a piecewise smooth approximation to the data and a nominal set of edges. If we assume that the edges are essentially correct but imprecisely localized, then, according to the Approximation Theorem, if we increase β, then the localization should improve. To prevent the introduction of smaller scale edges, we smooth the data g. This is accomplished by introducing feedback from f into g. This smoothing is suppressed in a neighborhood of the coarse scale edges to prevent smearing of the true edges in the data. These various mechanisms are balanced in accordance with the scaling conditions of the Approximation Theorem to achieve a stable, convergent algorithm. The edge focusing algorithm adheres to this paradigm with the additional feature that

the variational formulation is replace by Γ-convergent approximation, where the degree of approximation is refined as the algorithm proceeds.

Since we intend to use an approximation to the variational formulation, it is prudent to consider whether the approximation deviates in a significant way from the original formulation with regard to distortion of edges. Although analysis is prohibitively difficult, we expect (and simulations have borne this out) that the spreading of the edges in the Γ-convergent approximation actually ameliorates some of the geometric distortion. We recall that the primary reason for the geometric distortion is that the term $|K|$ in E determines the structure of the singularities. Roughly speaking, this arises because the length term is one dimensional and scales linearly while the other terms are two dimensional integrals and hence scale quadratically in the size of the domain. (Actually this is not precisely true because singularities arise in f, but the dominance of the length term still occurs at singularities in K.) When the edges are smeared and length is replaced by a two-dimensional integral the concentration of cost in the length term is alleviated and, hence, we expect the distortion to be relaxed. The price paid for this is the lack of resolution of the edges. The edge focusing algorithm begins with thick edges, thus relaxing distortion, but ends by sharpening the edges while scaling the parameters in accordance with the Approximation Theorem. Thus, the edges are focussed as resolution increases.

Edge focusing is achieved by perturbing equations (8.7) and (8.8), introducing dynamics into β, c, and g. The additional dynamics take the following form,

$$\frac{\partial}{\partial t} g(x,t) = \epsilon \rho(v(x,t))(f(x,t) - g(x,t)) \, , \tag{8.9}$$

$$\frac{\partial}{\partial t} \beta(t) = \epsilon \beta(t) \, , \tag{8.10}$$

$$\frac{\partial}{\partial t} c(t) = -\epsilon c(t) \, , \tag{8.11}$$

where ϵ is a small positive constant included to reflect the fact that these equations are perturbations of equations (8.7) and (8.8).

The dynamics in β, c, and g are intended to come into effect only after the basic descent equations (8.7) and (8.8) have essentially converged. Thus, we assume $g(x,0)$ is the initial data and $f(x,0)$ and $v(x,0)$ satisfy their respective Euler-Lagrange equations with $\beta = \beta(0)$ and $g(x) = g(x,0)$. This implies the presence of a nominal set of edges, i.e., a function $v(x,0)$. We will be guided by the heuristic that the subsequent focusing should only focus the edges already found and not introduce new ones.

To understand the effect of these equations, it is best to first assume $\rho \equiv 1$ and eliminate equation (8.11), i.e., fix $c(t) = c(0)$. When this is done, one

obtains the solution

$$\begin{aligned} v(x,t) &= v(x,0)\,, \\ f(x,t) &= f(x,0)\,, \\ \beta(t) &= \beta(0)e^{\epsilon t}\,, \\ g(x,t) &= g(x,0)e^{-\epsilon t} + f(x,0)(1-e^{-\epsilon t})\,. \end{aligned}$$

It can be checked that $\partial_f E = 0$ and $\partial_v E = 0$ for all t and that $f(x,t)$ (locally) minimizes E for all t. From the perspective of the Approximation Theorem, if we define $g_I = f(x,0)$ then by appropriately interpreting $g(x,t)$ as a smeared, noisy version of g_I we see that these equations effectively implement a continuous version of the limit process described in the Approximation Theorem. However, with the simplifying assumptions made here, there is no change in the edge function with time, i.e., no edge focusing. This is the role played by the dynamics in c and the function ρ.
Consider now the full set of equations (8.9)-(8.11). Equation (8.11) reduces c. This heuristic mimics the Γ-convergence to the variational formulation. Thus, edges are sharpened as t increases. Equation (8.9) introduces feedback from f into g, effectively smoothing g. We choose ρ to suppress the smoothing of g in a neighborhood of the edges, i.e., we choose ρ so that $\rho(v(x,t))$ will be approximately zero inside some neighborhood of the edges and approximately one outside some larger neighborhood. We make a correspondence with the conditions of the Approximation Theorem by interpreting the effect of ρ on $f(x,t)$ as a smearing of the edges in the ideal data which is iterpreted as $\lim_{n\to\infty} f(x,t)$.[4] Hence, the width of the larger neighborhood should shrink as $\beta^{-1}(t)$. A simple and reasonable choice, for example, is $\rho = \Phi$ since in this case the neighborhood width is proportional to $c(t) = (c(0)/\beta(0))\beta^{-1}(t)$. Since the edges in the data are not smoothed and $\beta(t)$ is becoming large, the singularities of the edge function should converge to the 'true' edge locations.

7. Discretization and Parameter Choices

In this section, we address some of the issues which arise as a consequence of discretization of E_c. In particular, appropriate step sizes for the discrete versions of the descent algorithm are given, and the relative rates of the gradient descent and the scaling dynamics are considered. A more detailed version of this section can be found in [307].
We assume a discretization in which f and g are defined on a rectangular subset of $\mathbb{Z}^2$, i.e., the lattice generated by the vectors $(0,1),(1,0)$. The discrete version of v is defined on the inter-leaving subset of a square

[4] See [307] for a detailed description of the correspondences.

lattice which is twice as dense and rotated 45°, i.e., the lattice generated by $(1/\sqrt{2},1/\sqrt{2}),(1/\sqrt{2},-1/\sqrt{2})$ and translated by $(1/2,0)$. Presumably one can consider different lattice spacings, but it turns out that, by scaling, the lattice spacing can be absorbed into the other parameters.
Each pixel y in the discretization of v is uniquely associated with the two nearest pixels x and x' in the discretization of f. For each such y let $df(y)$ denote $(f(x)-f(x'))^2$. The derivatives of Ψ and Φ as real functions will be denoted $\dot{\Psi}$ and $\dot{\Phi}$ respectively. Note that in equations (8.7) and (8.8), time scaling can be absorbed into c_f and c_v. Thus, time is discretized simply by substituting

$$\begin{aligned} \frac{\partial}{\partial t} f(x,t) &\to f(x,t+1) - f(x,t)\,, \\ \frac{\partial}{\partial t} v(y,t) &\to v(y,t+1) - v(y,t)\,. \end{aligned}$$

We now address the question of choosing c_f and c_v. A standard gradient descent would have both c_f and c_v constant. If we try to set c_v constant then it must be chosen small since $|\nabla f|^2/\alpha$ can be quite large, hence convergence will be slow. A computationally efficient choice that gives much faster convergence is to approximate a Newton type descent. By appropriately choosing $c_v(y)$ and making certain mild approximations[5], we obtain

$$v(y,t+1) = \frac{1}{2}\left(v(y) + \frac{\frac{1}{4c} + 2c\Psi(v(y)) \sum_{y' \in \mathcal{N}_v(y)} v(y')}{\dot{\Phi}(v(y))df(y)/(\alpha v(y)) + \frac{1}{4c} + 8c\Psi(v(y))} \right)\,, \quad (8.12)$$

where all quantities on the right hand side are evaluated at time t. This update formula enjoys many desirable properties. Note, in particular, that $v(y,t+1)$ is an average of $v(y,t)$ and a well behaved quantity which lies between 0 and 1, so $v(y,t+1) \in [0,1]$ is guaranteed. Setting c_f constant is much less problematic, and for our simulations we have set it to $(2\beta+8)^{-1}$ since this gives a good rate of convergence without allowing overshoot in f. The initialization of f and v will effect which local minimum is reached by the initial gradient descent. We expect this will have little effect on the edge focusing part of the algorithm. We have experimented with $f(0) = g(0)$ and also letting $f(0)$ be the solution to $f = \beta(0)\Delta(f-g)$ with Neumann boundary conditions. The second choice is better when more smoothing is desirable, i.e., in noisy or textured images. In general, we set $v(0) = 1$. With these choices, we observe that the basic descent on f and v converges in about 30 iterations for the range of parameters we have experimented with. (Larger values of c and smaller values of β will reduce the rate of convergence.)

[5] An approximation is only required when $\Psi \not\equiv 1$; the details may be found in [307].

The scaling dynamics associated with edge focusing can be discretized in a straightforward way.

$$\begin{aligned} g(x,t+1) &= g(x,t) + \epsilon\rho(x,t)(g(x,t) - f(x,t)) , \\ \beta(t+1) &= (1-\epsilon)^{-1}\beta(t) , \\ c(t+1) &= (1-\epsilon)c(t) , \end{aligned}$$

where ρ and ϵ are to be specified in each case.
Termination of the computation is best controlled through the value of $c(t)$. As $c(t)$ becomes very small, the discretization error becomes more significant. For the choices of Ψ and Φ used for the simulations presented in this chapter, we allowed $c(t)$ to become small enough so that the effective edge width is one pixel. (Effective edge width can be defined as the width of the set $\{\Phi(y) < 1/2\}$, for example.)

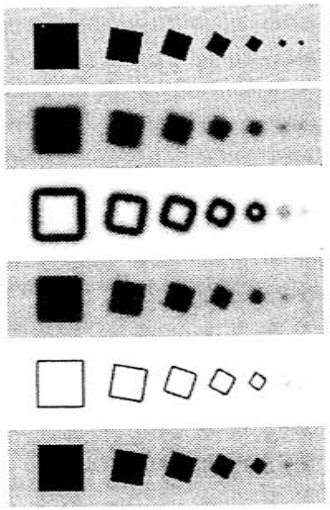

Figure 8.2. Square images, 120 × 490. From top to bottom: Data g, no edge smoothed data f_I, pre-focusing $\Phi(v)$, pre-focusing f, final $\Phi(v)$, and final f. Initial parameters are $\beta = 0.01, \alpha = 0.008, c(0) = 2.0$ and final parameters are $\beta = 0.1, c = 0.2$.

8. Simulation Results

In this section, we present the results of some simulations of the edge focusing algorithm. Image intensity is linearly scaled to lie in the range $[0, 1]$. In the two dimensional plots and in the images, we plotted $\Phi(v)$ on the same mesh as f, i.e., the plotting mesh corresponds to $\mathcal{L}_f$. For each $x \in \mathcal{L}_f$, we plot the minimum value of $\Phi(v)$ among the four nearest neighbors.
Figure 2 serves as a demonstration of the behavior of the edge focusing algorithm and its dependence on the parameters α and β. The image is a synthetic image of several squares of various sizes and rotation. The squares have side lengths of $60, 50, 40, 30, 20, 10$, and 6 pixels. We chose squares to illustrate the effects observed at high curvature edges, and rotated them to demonstrate the rotational invariance of the algorithm. The functions Φ and Ψ are as in equation (8.2) and we set $\rho(v) = \Phi(v)$. The initial value f_I is given by the no-edge smoothing of g associated with $\beta(0)$. We have carefully chosen the parameters to make the detection of some of the squares marginal. For Figure 2, the initial values are $\beta(0) = 0.01$, $\alpha = 0.008$, and $c(0) = 2.0$. This value of $\beta(0)$ corresponds to smoothing of the image over a radius of approximately 10 pixels, relatively large compared to the sizes of the squares. Such smoothing would not be required for good quality 'real' images. The final values of the parameters are $\beta(T) = 0.1$ and $c(T) = 0.2$ (where T is the time of termination), and ϵ was chosen so that 200 iterations with scaling are required to reach T. Figure 2 presents the data $g, f_I, \Phi(v(0)), f(0), \Phi(v(T))$, and $f(T)$ from top to bottom. Note the accurate localization of those edges detected. There is little visible distortion in the edges even of the smallest square whose edges were detected at all.
Figure 3a and 3b illustrate the effects of variation in α and β. Both are images of $\Phi(v(T))$. The image consists of the original with two more copies of the series of squares with intensities chosen so that the difference in intensity between the background and square has been reduced by a factor of 0.7 and 0.5 respectively. By scaling, one can see that this has the same effect as increasing α by a factor of 2 and 4, respectively. To generate Figure 3a, we used parameters $\beta(0) = 0.01, \alpha(0) = 0.003, c(0) = 2.0$ and $\beta(T) = 0.1, \alpha(T) = 0.003, c(T) = 0.2$. The parameters used to generate Figure 3b are the same except $\beta(0) = 0.006$ and $\beta(T) = 0.06$.
Figure 4 demonstrate the algorithm on 'real' images. Figure 4 is 512 × 512 pixels. In general, ϵ is chosen so that 200 iterations with scaling are required. The data is in Figure 4a. The image has been processed for two different sets of parameters to indicate the stability of the edges under a change in scale. In both cases, the displayed images are the following. Figures b,c, and d are $\Phi(v(0))$, $f(T)$, and $\Phi(v(T))$, respectively. Figures e,f, and g reiterate b,c, and d for the second set of parameters. The first set of parameters is

given by

$$\beta(0) = 0.01, \quad \alpha = 0.001, \quad c(0) = 2.0, \quad \beta(T) = 0.1, \quad c(T) = 0.2$$

and the second by

$$\beta(0) = 0.2, \quad \alpha = 0.005, \quad c(0) = 2.0, \quad \beta(T) = 2.0, \quad c(T) = 0.2 \ .$$

The value $\beta = 0.01$ roughly corresponds with smoothing by averaging over a 10×10 window and hence represents a high degree of smoothing. The value $\beta = 0.2$ roughly corresponds with smoothing by averaging over a 2×2 window. The degree of smoothing in the two examples is widely different and yet the edges found at the coarse scale are essentially a subset of those found at the finer scale. Virtually no scale dependent distortion is visible.

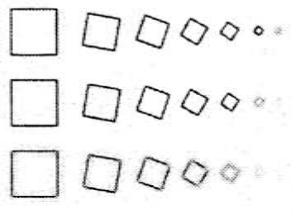

Fig. 3a

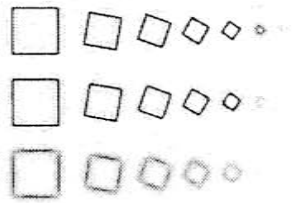

Fig. 3b

Figure 8.3. Square images of varying intensities, 340×490. Final $\Phi(v)$. For Figure 3a the initial parameters are $\beta = 0.01, \alpha = 0.003, c(0) = 2.0$ and the final parameters are $\beta = 0.1, c = 0.2$. For Figure 3b the initial parameters are $\beta = 0.006, \alpha = 0.003, c(0) = 2.0$ and the final parameters are $\beta = 0.06, c = 0.2$.

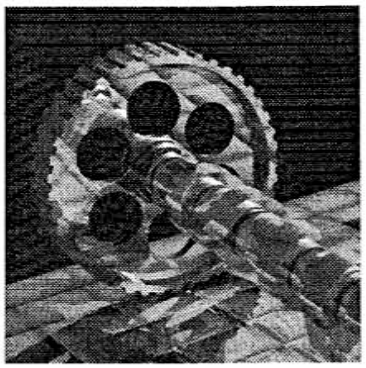

Fig. 4a

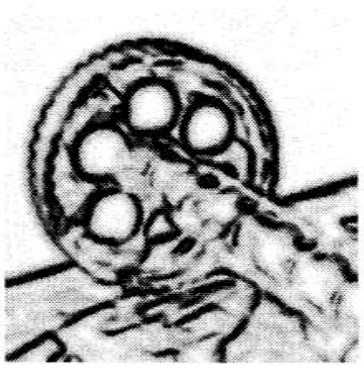

Fig. 4b

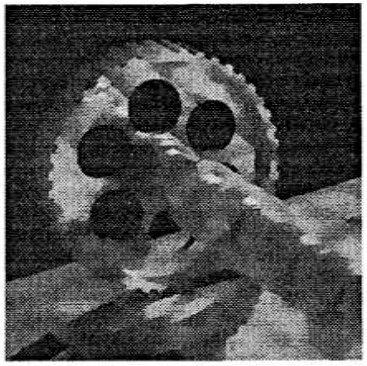

Fig. 4c

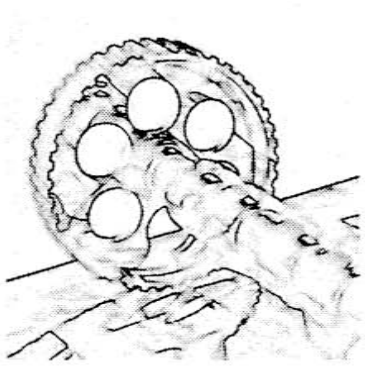

Fig. 4d

Figure 8.4a, b, c, d. Cam image 510×512. Original g, Prescaling $\Phi(v)$, final f, and final $\Phi(v)$, respectively, with initial parameters $\beta = 0.01$, $\alpha = 0.001$, $c(0) = 2.0$ and final parameters $\beta = 0.1$, $c = 0.2$.

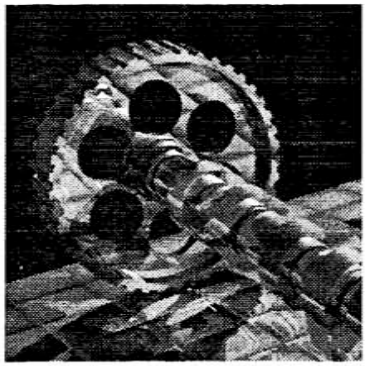

Fig. 4a

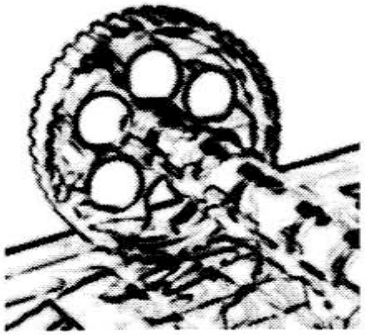

Fig. 4e

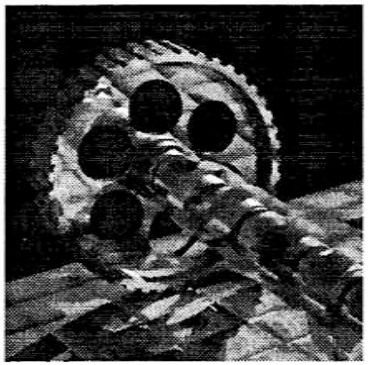

Fig. 4f

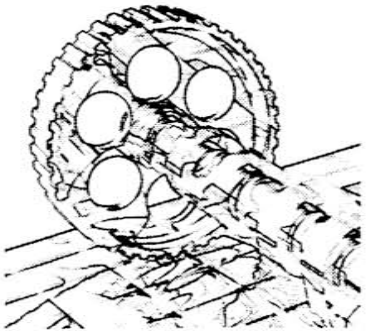

Fig. 4g

Figure 8.4a, e, f, g. Original g, Prescaling $\Phi(v)$, final f, and final $\Phi(v)$, respectively, with initial parameters $\beta = 0.2$, $\alpha = 0.005$, $c(0) = 2.0$ and final parameters $\beta = 2.0$, $c = 0.2$.

Acknowledgements: We are indepted to Pietro Perona and to Stefano Casadei for help with images and image processing software.

COUPLED GEOMETRY-DRIVEN DIFFUSION EQUATIONS FOR LOW-LEVEL VISION

Marc Proesmans

and

Eric Pauwels[1]

and

Luc van Gool

ESAT-MI2, Catholic University Leuven
Mercierlaan 94, B-3001 Leuven, Belgium

Abstract. This chapter introduces a number of systems of coupled, nonlinear diffusion equations and investigates their role in noise suppression and edge-preserving smoothing. The basic idea is that several maps describing the image, undergo coupled development towards an equilibrium state, repre- senting the enhanced image. These maps could e.g. contain intensity, local edge strength, range, or another quantity. All these maps, including the edge map, contain continuous rather than all-or-nothing information, following a strategy of least commitment. Each of the approaches has been developed and tested on a parallel transputer network.

1. Introduction and basic philosophy

1.1. Energy minimization and systems of coupled diffusion equations

Optimization of energy-functionals provides a clear-cut and, from a conceptual point of view, very attractive framework for the regularisation and processing of images. It is not surprising therefore that it has been the inspiration and point of departure for many investigators (cfr. [362, 34, 264, 273] to name just a few). The underlying basic idea is that a given input-signal g is transformed into an output-signal f in such a way that the result *minimizes* a predefined cost- or energy-functional. In its simplest form both

[1] Post-doctoral Research Fellow of the Belgian National Fund for Scientific Research (NFWO).