Constrained Restoration and the Recovery of Discontinuities

Donald Geman, *Member, IEEE,* and George Reynolds

Abstract—The linear image restoration problem is to recover an original brightness distribution X^O given the blurred and noisy observations $Y = \mathcal{K}X^O + B$, where $\mathcal{K}$ and B represent the point spread function and measurement error, respectively. This problem is typical of ill-conditioned inverse problems that frequently arise in low-level computer vision. A conventional method to stabilize the problem is to introduce *a priori* constraints on X^O and design a cost functional $\mathcal{H}(X)$ over images X, which is a weighted average of the prior constraints (regularization term) and posterior constraints (data term); the reconstruction is then the image X, which minimizes $\mathcal{H}$.

A prominent weakness in this approach, especially with quadratic-type stabilizers, is the difficulty in recovering discontinuities. One seeks an estimate of X^O, which not only recovers the shape of the original image over smooth patches, for example, those that are planar or quadric, but also recovers sharp transitions between these components.

We therefore examine prior smoothness constraints of the form $\phi(D^k X)$, where $\phi(u) = -(1 + |u|)^{-1}$, and D^k denotes a kth order derivative $k = 1, 2$, or 3. The important attributes of ϕ are its *concavity* on $(0, \infty)$ and its *finite* asymptotic behavior $(\lim_{u \to \infty} \phi(u) < \infty)$. Such constraints permit the recovery of discontinuities without introducing auxiliary variables for marking the location of jumps and suspending the constraints in their vicinity. (In fact, our optimization problem is equivalent to one involving a *noninteracting* "line process.") In this sense, discontinuities are addressed *implicitly* rather than *explicitly*.

Selecting the parameters, especially the relative weight λ between the prior and posterior terms (the "smoothing parameter"), is also problematical. Moreover, in our view, there is a conspicuous absence of theoretical results on model validation, even for idealized X^O. By exploiting the concavity of ϕ and assuming that X^O is an ideal (but prototypical) pattern, we calculate λ by requiring that $Pr\{X^O \in \mathcal{W}\} \approx 1$, where $\mathcal{W}$ is the set of *coordinate-wise minima* of $\mathcal{H}$. This procedure then yields λ (actually an upper bound) as a function of the other model parameters, such as the noise variance and blur coefficients.

Index Terms— Concave stabilizers, discontinuity recovery, "dual" energy, higher order constraints, image deblurring, model validation, nonlinear restoration, parameter selection.

I. Introduction

A. Image Blurring

THE IMAGE restoration problem is to recover a 2-D brightness (or other source) distribution X^0 defined over a continuous domain from discrete energy measurements actually recorded by a sensor. For example, visible light is sensed by a video or CCD camera, and the continuous distribution X^0 is converted into discrete samples Y_s, where s ranges over some 2-D rectangular lattice. For electro-optical and other devices, the transformation from X^0 to Y involves the degradation of the signal by the transport medium, noise, and sensor imperfections, as well as by the process of discretization itself, namely, digitization and quantization. The most accurate (but seldom used) model would therefore account for a variety of degradation factors: blur, through the composition of X^0 with a (possibly space-variant) point spread function (PSF); quantum noise, i.e., random fluctuations in the number of photons striking the photoactive material; noise in the scanning electronics; radiometric distortion; and other effects (see, e.g., Andrews and Hunt [1]).

In many situations, the dominant effect is blurring, which is the principal concern in this work. This distortion and loss of spatial resolution may be due to defocusing or to other factors such as motion, scattering, and atmospheric turbulence. The simplest model accounting for both blur and (signal-independent) noise is the familiar linear degradation model

$$Y_s = \left(\mathcal{K}X^0\right)_s + \eta_s, \qquad s \in S$$

where S is an $N \times N$ lattice, η is taken as white noise, and $\mathcal{K}$ is an operator representing the PSF. If we represent the domain of X^0 by an $N_0 \times N_0$ lattice $S_0 \supseteq S$, then in matrix notation

$$Y = \mathcal{K}X^0 + \eta \tag{1.1}$$

in which case, relabeling the sites, we regard X^0 as an $N_0^2 \times 1$ vector, Y and η as $N^2 \times 1$ vectors, and $\mathcal{K}$ as an $N^2 \times N_0^2$ matrix. In general, $N < N_0$, due to the nature of optical blurring and data acquisition (see Fig. 1).

Two exceptional cases in which one may assume $N = N_0$ are toroidal blurring (hence, actual convolution) and given boundary values (i.e., X^0 is known on $S_0 \backslash S$). However, these are generally unrealistic assumptions, and consequently, we shall restrict our attention to the case of an underdetermined system. Linear inverse problems of the same nature appear frequently in such related areas as radio astronomy, microscopy, computed tomography, and low-level computer vision.

The determination of the (distribution of) data Y given the true brightness pattern X^0 is the so-called "direct problem," and we will assume this mechanism is specified. In particular, we assume that the PSF is known. This is often a reasonable assumption; for example, the blur induced by the Vidicon camera has been determined. Finally, we assume throughout

Manuscript received May 17, 1990; revised April 29, 1991. This work was supported by the National Science Foundation under grant DMS-8813699 and by the Office of Naval Research under Contract N00014-88-K-0289.

D. Geman is with the Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003.

G. Reynolds is with VI Corporation, Northampton, MA, 01060.

IEEE Log Number 9104527.

0162-8828/92$03.00 © 1991 IEEE

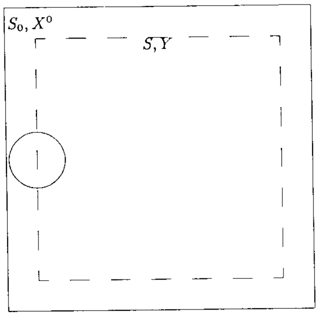


Fig. 1. Inner square S is where the data Y is given, and the outer square S_0 is where we wish to reconstruct the image X^0. The support of the point spread function determines the size of S_0.

that the noise process η consists of independent and identically distributed Gaussian random variables with zero means and known variance σ^2. This choice of the noise statistics is primarily for convenience, and the methodology supports other choices.

There is usually a severe loss of information in the transformation from X^0 to Y. In particular, the system is underdetermined, and $\mathcal{K}^{-1}$ is obviously not well defined. Moreover, even if $\mathcal{K}$ were invertible, the "inverse problem" is usually ill conditioned because the matrix $\mathcal{K}$ is nearly singular, and hence, there is little control over the propagation of measurement errors from the data to the solution. These observations can be made more precise in operator-theoretic terms (see, e.g., Bertero [2]), but the basic dilemma is clear: given $\mathcal{K}$ and Y, the solution space of (1.1) is typically very large, and two images with blurred values very close to Y can be far apart both visually and as vectors.

Finally, concerning the choice of blurs, most of our experiments involve (approximations to) Gaussian PSF's, which are known to model various phenomena, such as the distortion due to atmospheric turbulence. We also include an experiment with a 2D uniform PSF (which results, e.g., from a defocused, circular lens) and one with a 1D uniform PSF (motion blur). In general, as mathematical inverse problems, the Gaussian blur is the most difficult, and the motion blur is the least difficult; at least this has been our experience and can be partially substantiated by analyzing the degree of attenuation of high-frequency components. (Obviously, some adjustment must be made for the blur extent.) However, visually, the reverse appears true: The motion blur presents the greatest challenge, then the 2D uniform blur, and the Gaussian blur appears the easiest to "invert." It should also be emphasized that all these problems are substantially harder in the presence of noise, which imposes fundamental limitations on the degree of accuracy that can be obtained.

B. Summary of Results and Methodology

Most image restoration methods employ prior constraints in addition to those derived from modeling the image formation process. Our method belongs to this category; we formulate deblurring as a (nonlinear) optimization problem using a cost functional designed to accommodate assumptions about X^0 as well as the degradation model.

Specifically, we exploit the common observation that most real scenes are locally smooth, i.e., the variations in intensity are well behaved away from visual boundaries and textured areas. The cost functional is therefore constructed to emphasize images that are locally smooth and consistent with the data, i.e., whose blurred values are close to Y. This converts the ill-posed inverse problem (1.1) into a well-formulated (and hopefully well-conditioned) optimization problem. (A simpler constraint is simply to enforce *positivity* on the solution, in which case, the reconstructions can be obtained with quadratic programming. Generally, however, the problem is still badly formulated.)

The reconstruction is defined as any (global) minimum of a function

$$\mathcal{H}^k(X) = \Phi^k(X) + \lambda\|Y - \mathcal{K}\|^2 \tag{1.2}$$

$$= \sum_{C} \phi\big(D_C^k(X)/\Delta\big) + \lambda \sum_{s \in S} (Y_s - (\mathcal{K})_s)^2. \tag{1.3}$$

Here, Y and $\mathcal{K}$ are as above, X is a positive, integer-valued function on S_0, and Δ and λ are positive parameters. (Notice that λ multiplies the data term and, hence, is just the inverse of the usual "smoothing parameter.") We are primarily interested in the planar and quadric cases ($k = 2, 3$), although all three functionals $\mathcal{H}^k$, $k = 1, 2, 3$, will sometimes be utilized during the optimization procedure. The "regularization term" $\Phi^k(X)$ imposes a smoothness condition X of order k. For example, for $k = 1$, the first summation in (1.3) ranges over all horizontal and vertical nearest neighbor pairs of pixels, and $D_C^1(X)$ is simply the intensity difference $X_s - X_t$ for $C = \{s, t\}$. For $k = 2, 3$ the terms D_C^k correspond to discrete (linear) approximations to the differences between elements of the gradient vector ($k = 2$) and Hessian matrix ($k = 3$) at adjacent pixels. This will be amplified in Section II.

We have selected one from among the family of functions $\phi(u) = -(1 + |u|^\gamma)^{-1}$ suggested by Geman and McClure [13] for constructing "prior distributions" (see Section I-C) for image reconstruction and related problems. Notice that each such ϕ is even, increasing for $u \geq 0$, and $\lim_{u \to \infty} \phi(u) = 0$. The motivation for the latter property was to allow sharp transitions between distinct regions. In this sense, we say that $\mathcal{H}^k$ addresses discontinuities *implicitly.*

In this paper, we take

$$\phi(u) = \frac{-1}{1 + |u|}. \tag{1.4}$$

This function is concave on $(0, \infty)$ and therefore strictly noninterpolating in regard to image transitions in a sense that will be shortly explained. Other concave functions might do equally well, but this choice has yielded consistently good results. The motivations are the results on coordinate-wise minima (see Section IV) and the following noninterpolating property. Consider just a 1D discrete signal and the class J_δ of real-valued functions defined on the integers from 0 to I that have the property that $X_0 = 0$ and $X_I = \delta$. Then, provided ϕ is even, and concave, increasing on $[0, \infty)$, and for any choice of I and δ, the function

$$\Phi(X) = \sum_{i=0}^{I-1} \phi(X_{i+1} - X_i)$$

is minimized over J_δ by those functions in J_δ with a single jump. Moreover, it is not difficult to extend this property of minimizing the number of discontinuities to higher order derivatives. For example, again in 1D, functionals of the form $\sum \phi(X_{i+1} + X_{i-1} - 2X_i)$ are minimized by curves displaying the fewest number of linear segments subject to boundary conditions on X_0, X_1, X_{I-1}, and X_I.

The traditional choice in "regularized" least-squares restoration is the quadratic function $\phi(u) = u^2$. Despite the computational advantages (notice that $\mathcal{H}^k$ is then quadratic, resulting in a *linear* estimate $\hat{X} = \hat{X}(Y)$), we find it ill suited to image deblurring because the rapid growth as $u \to \infty$ inhibits the recovery of large intensity gradients, and the slow growth as $u \to 0$ promotes excessive smoothing and interpolation: Small intensity differences (or higher order analogs) incur a relatively small "penalty," and transitions are optimally represented as gradual changes. For example, if a 1D signal is constrained as above at two endpoints of an interval, then the minimum energy solution for the intervening points distributes the total jump in equal parts, i.e., linearly interpolates (see also the discussion in Shulman and Herve [27]). In contrast, for the function ϕ in (1.4), the jump is absorbed in one step. Moreover, as mentioned above, the finite asymptotic behavior ($\phi(\infty) = 0$) does not introduce a bias against large transitions. The result of these properties is that reconstructions are more accurate in the vicinity of discontinuities.

There is another way to interpret the role of discontinuities in this approach. Discontinuities may be addressed explicitly by the addition of a line process as originally done in [12]. If the line variables do not interact with each other (i.e., there is no term for "organizing" the boundaries), then we shall show in Section III that the two approaches are in fact equivalent in the following sense. Suppose $\mathcal{H}$ is of the form above, and ϕ is selected such that $\phi(\sqrt{u})$ is concave for $u > 0$; in particular, this holds whenever ϕ itself is concave for $u > 0$. Then, one can define a coupled functional $\mathcal{H}^*(X, B)$, where B denotes an auxiliary (and continuous-valued) array, and $\mathcal{H}^*$ is quadratic in X for each B, such that $\mathcal{H}$ and $\mathcal{H}^*$ have the same global minima in X.

Note: Since the function $\phi(u)$ in (1.4) is concave, there is a corresponding coupled functional; see Example 2 in Section III. However, as noted above, the equivalence persists for a wider class, for instance, for the functions $\phi(u) = -(1 + |u|^\gamma)^{-1}$ with $\gamma \leq 2$.

Usually, a particular method is validated by heuristics and by displaying successful experiments. In addition, however, it might be worthwhile to *guarantee* the photometric accuracy of restorations relative to a collection of templates, i.e., a prototype class of original images. In our case, this would mean that $\mathcal{H}^k$ is minimized by X^0, at least for elementary patterns, for instance, those composed of piecewise constant regions separated by very simple boundaries. However, due to the noise, the set of minima is, in fact, a *random* set. An appropriate optimality criterion must then be formulated, for instance, in terms of the probability of the event that $\mathcal{H}^k$ is minimized by the true image, but results on *global minima* appear elusive in any generality; see Section IV.

Instead, we develop a surrogate criterion in terms of *coordinate-wise minima.* A value X is a coordinate-wise minimum for a function $\mathcal{H}$ if any change in X at a single coordinate (i.e., pixel) increases the value of $\mathcal{H}$. (Notice that since ϕ is not differentiable, neither is $\mathcal{H}$; if it were, a coordinate-wise minimum X would necessarily be a stationary point of $\mathcal{H}$, that is, $\nabla\mathcal{H}(X) = 0$.) Let $\mathcal{H}^k$ and ϕ be as in (1.3) and (1.4); again, the important property of ϕ that is used is *concavity.* Then, for the class of images X^0 mentioned above, one can show that with probability arbitrarily close to one, X^0 is at least a coordinate-wise minimum of the function $\mathcal{H}^k$ for each k, provided the parameters λ and Δ are suitably chosen. This is not surprising in view of the following observation: Consider any site s, whether interior to a smooth patch or in the vicinity of a discontinuity of (an idealized) X^0; then, a small perturbation $X_s^0 + \delta u$ of X_s^0 "toward the data" will incur a decrease in the data component of $\mathcal{H}^k$ of order δu but an increase in the regularization term of the *same* order because $\phi'(0+) > 0$. The magnitudes of these increments will depend on all the model parameters, including k, Δ, σ^2, and the blur coefficients, as well as the particular noise realization. However, one can select λ to make it very likely that the combined effect is a *net increase* in energy, i.e., X^0 is a stationary point. On the other hand, if $\phi'(0) = 0$, there will always be interpolation, and in general, X^0 will be neither a coordinate-wise nor local minimum.

In this way, we specify λ as a function of the other parameters; see (4.3). Actually, what we obtain is an *upper bound* $\overline{\lambda}$; any $\lambda \leq \overline{\lambda}$ will suffice, although we just use the upper bound in our experiments. The dependence on $\mathcal{K}$ is only through the sum of squares β of the blur coefficients. Moreover, as we shall see in Section IV, $\overline{\lambda}$ decreases as σ and β increase, as we might expect. Finally, since $\overline{\lambda}$ also depends on Δ, we may also interpret the results as a $\Delta - \lambda$ curve for selecting these *two* parameters. However, since appropriate values for Δ are more or less evident (but not at all for λ), we have chosen to specify Δ by hand and use the corresponding upper bound for λ.

Of course, the experimental images, both synthetic and real, may only *locally* resemble the ideal patterns. Nonetheless, our persistent experience has been that the value of λ so derived yields results comparable with those obtained by extensive trial and error, i.e., by trying many values of λ over several orders of magnitude and selecting the one that yields the most faithful reconstruction according to, say, the L_1 norm, which in turn is usually the one that appears the best *visually.* (This is reminiscent of "scale space" methods (see Section IV); see also the results and experiments of Yang [30], which corroborate our findings and extend the formal results to the case of signal-dependent noise.)

The optimization problem involved in minimizing $\mathcal{H}^k$ is formidable, especially for the higher order models. We have used stochastic relaxation with annealing. This algorithm involves repeatedly sampling from the univariate conditional distributions of the probability measure

$$\Pi_\tau(X) = e^{-\mathcal{H}^k(X)/\tau} \Big/ \sum_X e^{-\mathcal{H}^k(X)/\tau}$$

for increasingly smaller values of the "temperature" τ. The effect is to progressively concentrate the mass around the minimum. The *asymptotic* properties of the annealing algorithm are fairly well understood; in particular, the sequence of states generated converges to the global minimum of $\mathcal{H}$ in an appropriate sense (see Section V). Still, what is important is the *finite-time* behavior. Nonetheless, we have found this procedure at least comparable with other optimization methods, although it is computationally expensive. Some of the practical implementation issues are addressed in Section V. In particular, we use a computational trick to gain about an order-of-magnitude speedup over proper stochastic relaxation; recently, this has been understood theoretically [11], [30].

The "energy surfaces" $(X, \mathcal{H}^k(X))$ are complex. In particular, for the cases $k = 2, 3$, there are many states X with nearly the same energy as the lowest energy states but visually quite distinct. (This becomes evident by sampling from the measures Π_τ for small τ using an original image with sharp boundaries as the starting point.) For example, whereas the lowest state $\hat{X}$ may display a perfect "step edge," there are numerous interpolated transitions, built from small planar or quadric ramps, that assume nearly the same energy value as $\hat{X}$. This makes the recovery of discontinuities quite difficult for the higher order models by *starting at the data Y;* since the data are in fact the blurred image (with noise), the discontinuities are already interpolated. In contrast, the first-order model is very effective in locating and recovering jump discontinuities but inadequate for recovering the basic geometric structure of regions, such as planar or quadric patches, for which the higher order functions are obviously better suited; indeed, the corresponding distributions Π_τ are concentrated on a richer and more plausible set of interpretations. These observations suggest a *coarse-to-fine analysis in the order of the model,* and this is the strategy we have adopted in some of our experiments. Starting at the data, we use the first-order model to generate a starting point for the second-order model, which in turn provides the starting point for the third-order model. Moreover, the results on parameter selection yield values for which the higher order models preserve the discontinuities recovered by the first-order model. This procedure is useful to the extent that sharp transitions constitute the information content of the image; if these features are not important (or present), then it is sometimes possible to obtain comparable (or even better) results beginning directly with the planar or quadric model initialized with the data Y; see Section VI.

C. Related Approaches

In the classical version of constrained least-squares, the linear inverse problem is formulated as a quadratic, constrained minimization problem:

$$\text{minimize}_X \left\{ \|QX\|^2 \mid \|Y - \mathcal{K}X\|^2 = c \right\} \tag{1.5}$$

where Q is a matrix representing first- or second-order differences or perhaps the Laplacian. The parameter c should be chosen in accordance with the noise variance σ^2; for the true image X^0, we know that $\|Y - \mathcal{K}X^0\|^2 = \|\eta\|^2 \approx N^2\sigma^2$. For instance, in the first-order case

$$\|QX\|^2 = \sum (X_s - X_t)^2$$

where the sum extends over all pairs $\langle s, t\rangle$ of adjacent vertical or horizontal pixels. The effect is to emphasize reconstructions that are locally constant. The Lagrangian formulation is then

$$\text{minimize}_X \left\{ \|QX\|^2 + \lambda \|Y - \mathcal{K}X\|^2 \right\} \tag{1.6}$$

where $\lambda > 0$ must be adjusted to satisfy the constraint $\|Y - \mathcal{K}X\|^2 = c$. Notice that since λ is positive, any solution $\hat{X}$ to (1.6) is then a solution to (1.5) with $c = \|Y - \mathcal{K}\hat{X}\|^2$. The solution is found by solving the linear system

$$(\lambda^{-1}Q^tQ + \mathcal{K}^t\mathcal{K})X = \mathcal{K}^tY \tag{1.7}$$

which is usually inverted in the Fourier domain after a convolution has been arranged by approximating $\lambda^{-1}Q^tQ + \mathcal{K}^t\mathcal{K}$ by a circulant matrix. The case $Q = 0$ corresponds to ordinary least-squares (which is badly ill conditioned), and the classical Wiener filter is also a special case of (1.7) with Q constructed from the covariance matrices for X and η. Of course, the resulting estimate is a *linear* function of the data, and neither these constraints nor those involving positivity or entropy address the problem of discontinuities.

The two main components of the Bayesian approach are a "prior distribution" on images X, which encodes *a priori* knowledge or assumptions about the true image, and a degradation model, which is the conditional distribution of the data Y given X. Estimates are based on properties of the "posterior" distribution of X given Y. If this prior distribution is chosen with log likelihood proportional to $-\Phi^k(X)$ in (1.2), then under the assumption of Gaussian white noise, the posterior distribution is $\Pi_\tau(X)$ with $\tau = 1$ and $\lambda = 1/2\sigma^2$. In particular, our estimator is then the mode of the posterior distribution, which is often called the *maximum a posteriori* (MAP) estimator. More generally, the MAP estimator is of the form

$$\hat{X} = \arg\ \min_X(\Phi(X) + \lambda\Psi(X, Y))$$

where Φ and Ψ correspond, respectively, to the (negative) log likelihoods of the prior distribution on X and the conditional distribution of Y given X. Whereas the MAP estimator may be formulated independently of the distribution itself, other estimators, for instance, the posterior mean, are genuinely distributional attributes. The model parameters, e.g., λ and Δ, are construed as unknown parameters of the posterior distribution and are often estimated from the data using conventional methods such as maximum likelihood and method of moments. In contrast, our choice is *independent of the data.*

Recent applications of this methodology to problems in image restoration, reconstruction, and segmentation appear in Besag [3], [4], Chellappa, *et al.* [6], Derin and Elliott [7], Geiger and Girosi [8], Geiger and Yuille [9], Geman and McClure [13], Gidas [15], Green [16], Jeng and Woods [18], Marroquin [21], Marroquin *et al.* [22], Molina and Ripley [23], Rangarajan and Chellappa [25], Terzopoulos [28], and elsewhere. In particular, we share the emphasis in [5], [22], [24], [25], [28] on detecting and preserving discontinuities, although the problems there differ from ours; in particular, they do not involve blur. Closer to ours is the work of Molina and Ripley [23] on deconvolving astronomical images using log-Gaussian priors.

Previous applications utilizing members of the class of functions $\phi(u) = -(1+|u|^{\gamma})^{-1}$ include, in the first-order case, work on boundary detection and noise suppression for infrared images [10] with $\gamma = 3/2$ (including a "line process"), the work of Geman and McClure [13] on computed tomography, in which the prior distribution is over isotope surfaces ($\gamma = 2$) and the data term Ψ involves an attenuated Radon transform, recent work of Geman, *et al.* [14] on film restoration using higher order constraints with Φ^k, and unpublished work of the same authors on ultrasonic and infrared image enhancement, also using higher order constraints.

Convex regularization terms, with k=1, appear in the work of Besag [4] ($\phi(u) = |u|$) and Green [16] ($\phi(u) = \log \cosh u$). In Shulman and Herve [27], the quadratic is extended linearly rather than truncated as in [5]; see below. This is motivated by the theory of "influence functions" in robust statistics [17]. The convexity simplifies the computational problem, but these choices of ϕ lack the properties we seek, and the applications differ from ours; for example, [27] concerns optical flow.

There are several papers on image segmentation and surface interpolation from sparse data that do not involve blur (i.e., $\mathcal{K} = I$) but also focus on what we are calling "implicit discontinuities." For example, Blake and Zisserman [5] experiment with the truncated quadratic $\phi(u) = (u^2 - 1)^-$ (where $v^- = 0$ if $v > 0$ and otherwise equals v) and observe the duality between implicit and explicit discontinuities for the special case of a binary line process. Indeed, this is the justification in [5] for the elimination of the line process. We have found this choice unstable for deconvolution because in the critical early stages, the smoothing term is suspended over much of the image (specifically at each clique C for which $D_C^k(X) \geq \Delta$), and the method then exhibits the instability of *unconstrained* least squares. In addition, boundaries are interpolated due to the quadratic behavior near the origin, as explained above. Along similar lines, the segmentation procedure in the recent paper of Leclerc [20] roughly corresponds to setting $\Delta = 0$; consequently, $\Phi^k(X)$ simply counts the number of kth order discontinuities. In addition, a penalty is added for the order k of the model, and the minimum description-length (MDL) principle (Rissanen [26]) is employed for *a priori* parameter selection. The removal of the scaling parameter Δ is certainly appealing, but we have found this procedure also unstable in the context of deconvolution.

Finally, the recent work of Geiger and Girosi [8] involves computing the marginal distribution on the intensity process from the joint distribution on intensities and lines of a coupled Markov random field. The authors suggest the MAP estimator of the marginal distribution for surface reconstruction and discuss the extent to which discontinuities are accommodated by this procedure. Note that the marginal distribution is of the form (1.3), where $\phi'(0) = 0$. (It should be noted that the methodology in [5], [8], and [20] actually involves a *sequence* of ϕ functions.) Other results on "scalar line processes" and on the equivalence between functions with and without line variables have recently appeared in Geiger and Yuille [9] and Rangarajan and Chellappa [25].

II. Locally Constant, Planar, and Quadric Models

Let us say that X is planar on a subregion $T \subset S_0$ if there are constants A, B, C such that $X_{i,j} = Ai + Bj + C$ for all $(i,j) \in T$ and that X is quadric if there are constants A, B, C, D, E, F such that $X(i,j) = Ai^2 + Bj^2 + Cij + Di + Ej + F$ for all $(i,j) \in T$. Then, X is locally planar (resp. quadric) if the pixels in S_0 can be partitioned into regions on each of which X is planar (resp. quadric). Obviously, since the domain is discrete, these definitions apply to *any* image unless some assumptions are made about the *size* of the regions. Still, the basic idea is clear.

We now describe the functions D_C^k, $k = 1, 2, 3$. For the first-order case, we define a clique C as any pair of horizontal or vertical and adjacent pixels (s,t), which we can visualize as

$$(1)\ \begin{array}{c} s\bullet \\ t\bullet \end{array} \qquad (2)\ \begin{array}{cc} s\bullet & t\bullet \end{array}.$$

The first-order model is then

$$\mathcal{H}^1(X) = \sum_C \phi(D_C^1(X)/\Delta) + \lambda \sum_{s \in S} (Y_s - (\mathcal{K}X)_s)^2$$

where

$$D_C^1(X) = X_s - X_t \quad \text{and} \quad C = (s,t).$$

Note: The nearest-neighbor model does introduce a bias towards vertical and horizontal edges, especially in the first-order case; this can be ameliorated by including diagonal adjacencies (i.e., an eight-neighbor system), although we have not done so.

For the planar case, looking at second differences, i.e., differences between components of the gradient at adjacent pixels, yields cliques of three types, each involving three or four pixels:

$$(1)\ \begin{array}{c} s\bullet \\ t\bullet \\ u\bullet \end{array} \qquad (2)\ \begin{array}{cc} s\bullet & t\bullet \\ u\bullet & v\bullet \end{array} \qquad (3)\ \begin{array}{ccc} s\bullet & t\bullet & u\bullet \end{array}.$$

Now define

$$D_C^2(X) = \begin{cases} X_s - 2X_t + X_u & \text{if } C \text{ is of type 1 or 3} \\ X_s - X_t - X_u + X_v & \text{if } C \text{ is of type 2} \end{cases}.$$

Finally, for the quadric case, looking at third differences, i.e., differences between components of the (discrete) Hessian matrix at adjacent pixels, yields cliques of four types involving between four and six pixels:

$$(1)\ \begin{array}{c} s\bullet \\ t\bullet \\ u\bullet \\ v\bullet \end{array} \qquad (2)\ \begin{array}{ccc} p\bullet & q\bullet & r\bullet \\ s\bullet & t\bullet & u\bullet \end{array}$$

$$(3)\ \begin{array}{cc} p\bullet & s\bullet \\ q\bullet & t\bullet \\ r\bullet & u\bullet \end{array} \qquad (4)\ \begin{array}{cccc} s\bullet & t\bullet & u\bullet & v\bullet \end{array}.$$

In this case, we define $D_C^3(X)$ at the bottom of this page.

It is then easy to see that X is constant, planar, or quadric on S_o if and only if $D_C^k(X) = 0$ for every C for $k = 1, 2$ or 3.

III. Discontinuities

In view of the shortcomings of standard regularization theory for image restoration, in particular oversmoothing, a coupled Markov random field model was introduced in [12] in which image recovery and boundary detection are performed simultaneously. The basic idea was to include a "line process" $L = (L_{s,t})$ that is indexed by the dual lattice and suspends the continuity constraint associated with the pixel pair s, t when $L_{s,t} = 1$ (the line is "on") but preserves the constraint when $L_{s,t} = 0$ (the line is "off"). Moreover, the state $L_{s,t} = 1$ should be more likely (resp. less likely) than the state $L_{s,t} = 0$ if there is a large (resp. small) intensity gradient across $\langle s, t\rangle$, and the entire configuration L should reflect our prior expectations about boundary structure; for example, boundaries are usually sparse and connected. The prior model chosen in [12] was a Gibbs measure with "energy"

$$\Phi^*(X, L) = \sum_{\langle s,t\rangle} \phi^*(X_s - X_t)(1 - L_{s,t}) + V(L) \qquad (3.1)$$

where V is constructed to organize the line process, and $\phi^*(u) = 2\delta_u - 1 = -1$ if $u = 0$ and $= 1$ if $u \neq 0$. (This choice of ϕ^* is appropriate for a small number of grey levels.) If the degradation consists of blur and independent, additive white Gaussian noise, then the MAP estimator of the intensity-line pair (X, L) is the minimum of the (posterior) energy

$$\mathcal{H}^*(X, L) = \Phi^*(X, L) + \lambda\|Y - \mathcal{K}X\|^2. \qquad (3.2)$$

If we eliminate the interactions among the line variables by removing V and if we substitute the quadratic stabilizer $\phi^*(u) = (u/\Delta)^2 - 1$ for $\phi^*(u) = 2\delta_u - 1$ in (3.1), then Blake and Zisserman [5] observed that

$$\inf_L \Phi^*(X, L) = \sum_{\langle s,t\rangle} \phi((X_s - X_t)/\Delta)$$

where ϕ is the *truncated* quadratic: $\phi(u) = \left(u^2 - 1\right)^-$. It follows immediately that minimizing $\mathcal{H}^*$ in (3.2) (with $V = 0$ in Φ^*) with the quadratic ϕ^* and minimizing

$$\mathcal{H}(X) = \sum_{\langle s,t\rangle} \left(((X_s - X_t)/\Delta)^2 - 1\right)^- + \lambda\|Y - \mathcal{K}X\|^2$$

are equivalent problems in the sense that the set of $\hat{X}$ for which $(\hat{X}, \hat{L})$ minimizes $\mathcal{H}^*(X, L)$ for some $\hat{L}$ is identical to the set of $\hat{X}$, which minimizes $\mathcal{H}(X)$. It should be noted that we delete V solely to explore this connection and that allowing interactions among the line variables is useful for many problems.

In order to pursue this correspondence for general ϕ's and higher order constraints, we consider a process $B = (B_C)$ indexed by the appropriate cliques C (depending on the order of the model; see Section II) where each B_C assumes *continuous* nonnegative values (possibly $+\infty$) and represents the strength of the constraint associated with C. In the first-order, binary case, the relationship between B and L is simply $B_{s,t} = 1 - L_{s,t}$.

Now, given a model of the form we are using, namely

$$\mathcal{H}(X) = \sum_C \phi(D_C(X)/\Delta) + \lambda \sum_{s\in S} (Y_s - (\mathcal{K}X)_s)^2$$

one can ask for conditions on ϕ such that there exists a "dual energy"

$$\mathcal{H}^*(X, B) = \sum_C \left(B_C(D_C(X)/\Delta)^2 + \psi(B_C)\right) + \lambda \sum_{s\in S} (Y_s - (\mathcal{K}X)_s)^2$$

such that

$$\mathcal{H}(X) = \inf{}_B \mathcal{H}^*(X, B).$$

In this case, the problems of minimizing $\mathcal{H}$ and $\mathcal{H}^*$ are equivalent. Since there are no interactions among the B variables, this is equivalent to finding conditions on ϕ for which there exists a function ψ with

$$\phi(u) = \inf_{0\leq b} \left(bu^2 + \psi(b)\right).$$

This has the simple geometric interpretation that ϕ is the infimum of a family of quadratic functions; see Fig. 2.

Our motivation for this inquiry is twofold: first, to explain exactly how our model corresponds to one in which discontinuities are explicitly marked and, second, to explore how the computational difficulties we have encountered in minimizing $\mathcal{H}$ might be reduced by reformulating the optimization problem using $\mathcal{H}^*$. Notice that the term $(D_C(X)/\Delta)^2$ is quadratic in X because $D_C(X)$ is linear in X. It follows that under the joint probability law $e^{-\mathcal{H}^*(X,B)}/Z$, Z a constant, the process X is conditionally Gaussian given B, and the variables B_C are conditionally independent given X, with the same density up to a single parameter depending on $D_C(X)$. Consequently, stochastic relaxation with the dual process is easier (although perhaps not more efficient) than with the original process; see Example 3 below.

$$D_C^3(X) = \begin{cases} X_s - 3X_t + 3X_u - X_v & \text{if } C \text{ is of type 1 or 4} \\ X_p - 2X_q + X_r - X_s + 2X_t - X_u & \text{if } C \text{ is of type 2 or 3} \end{cases}$$

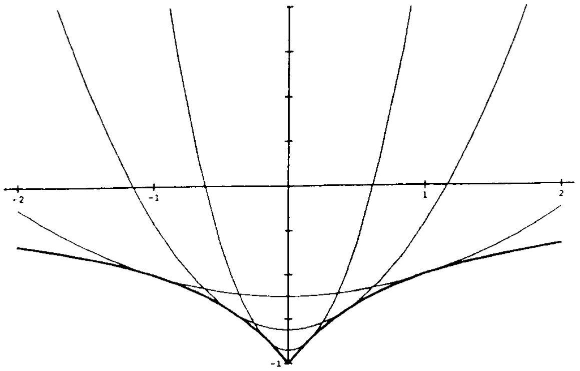


Fig. 2. Function $\phi(x)$ satisfying the conditions of Theorem 1 can be realized as the infimum of a family of quadratic functions. Here, $\phi(x) = -(1 + |x|)^{-1}$; see Example 2.

Theorem 1 (Existence of a Dual): Suppose $\phi(u)$ has the following properties on $[0, \infty)$:

1) $\phi(o) = -1$
2) $\phi(\sqrt{u})$ is concave
3) $\lim_{u \to +\infty} \phi(u) = 0$.

Then, there exists a function $\psi(b)$ defined on an interval $[0, M]$ such that

$$\phi(u) = \inf_{0 \le b \le M} \left(bu^2 + \psi(b)\right) \tag{3.3}$$

and such that $\psi(b)$ has the properties

a) $\psi(0) = 0$
b) $\psi(b)$ is strictly decreasing
c) $\psi(M) = -1$.

Note: The function $f(u) = \phi(\sqrt{u})$ is *necessarily* concave given the stated conditions on ψ. The reason is that f defines the lower envelope of a one-parameter affine family (cf. Kendall [19]).

Viewed geometrically, the idea of the proof is simple, although writing down the details is awkward and will only be done, in the Appendix, for the case in which ϕ is continuously differentiable. Basically, the values assumed by ψ are the y intercepts of the tangent lines to the graph of $f(u) = \phi(\sqrt{u})$. (Imagine Fig. 2 with the quadratics replaced by linear segments.) More specifically, the conditions on ϕ imply it is differentiable a.e. Let $A = \{u \ge 0 | f'(u) \text{ exists}\}$; then ψ can first be defined on $E = f'(A)$ by

$$\psi(b) = f(u) - ub, \qquad b = f'(u) \in E.$$

Then, ψ is decreasing on E, satisfies (3.3) for $u \in A$ (with the infimum over E), and can be extended from E to an interval $[0, M]$ to satisfy the stated conditions; M is the right-hand derivative of f at the origin.

Here are some examples; the given formulae are verified in the Appendix.

Example 1: If

$$\phi(u) = \frac{-1}{1 + u^2}$$

then

$$\phi(\sqrt{u}) = \frac{-1}{1 + |u|}$$

is strictly concave, and

$$\psi(b) = b - 2\sqrt{b}, \quad \text{with } 0 \le b \le 1.$$

In some cases, it can be difficult to compute ψ directly, and it is easier to reparametrize so that

$$\phi(u) = \inf_b \left(\xi(b)u^2 + \psi(b)\right) \tag{8}$$

for some function ξ for which $\xi(0) = 0$ and ξ is increasing.

Example 2: If

$$\phi(u) = \frac{-1}{1 + |u|}$$

then the conditions of Theorem 1 are satisfied, and we may choose

$$\xi(b) = \frac{b^{3/2}}{2(1 - b^{1/2})},$$

$$\psi(b) = \frac{b - 3b^{1/2}}{2} \quad \text{with } 0 \le b \le 1.$$

Note: In this example (see Fig. 2), the fact that $\phi'(0+) > 0$ is reflected in the potentially *infinite* bonding strength ($\xi(1) = +\infty$) associated with regions for which $D_C = 0$. Indeed, the value of B for which the minimum of $\mathcal{H}^*$ is achieved will be such that $B_C = +\infty$ for regions containing C, which are perfectly constant ($k = 1$), planar ($k = 2$), or quadric ($k = 3$).

Example 3: For the truncated quadratic $\phi(u) = (u^2 - 1)^-$, (3.3) is satisfied with

$$\psi(b) = -b, \qquad 0 \le b \le 1.$$

It the easy to check that the infimum in (3.3) is degenerate, and achieved at $b = 1$ whenever $u < 1$ and at $b = 0$ whenever $u \ge 1$. In the first-order case, the dual energy is then

$$\mathcal{H}^*(X, B) = \sum_{\langle s,t \rangle} B_{s,t}\Big(((X_s - X_t)/\Delta)^2 - 1\Big) + \lambda \sum_{s \in S} (Y_s - (\mathcal{K}X)_s)^2.$$

Coordinate-wise descent on the bond variables amounts to setting $B_{s,t} = 1$ if $|X_s - X_t| < \Delta$ and setting $B_{s,t} = 0$ otherwise; thus, $1 - B_{s,t}$ mimics the role of the binary line process in earlier work. Moreover, in this case, the distribution of the process $\{B_{s,t}\}$ under the joint law $e^{-\mathcal{H}^*(X,B)}/Z$ can be explicitly computed. As already noted, the variables $B_{s,t}$ are conditionally independent given X. Fix $\langle s, t \rangle$, and let $A = ((X_s - X_t)/\Delta)^2 - 1$. Then, an easy calculation shows that $B_{s,t}$ is uniformly distributed on $[0, 1]$ if $A = 0$ and otherwise as

$$-\frac{1}{A} \ln\left(1 - (1 - e^{-A})U\right)$$

where U is uniform on $[0, 1]$. It follows that stochastic relaxation with the dual process is reduced to choosing only uniform random numbers.

Remark: We have not assessed the complexity of the coupled energy surface associated with $\mathcal{H}^*(X, B)$; maybe it is inherently more resistent to global optimization than the original one. Consequently, we do not know if the balance between rapid simulation and overcoming local minima will result in a more efficient optimization procedure than ours, namely, stochastic relaxation based directly on $\mathcal{H}(X)$. Finally, other work along similar lines appears in [8], [9], [22], and [25].

IV. Parameter Selection

The free parameters are Δ (a scaling parameter) and λ, which balances fidelity to the prior constraints and fidelity to the data. (Recall that we are assuming the noise variance σ^2 and the point spread matrix $\mathcal{K}$ are given.) In our view, one very attractive method of selecting Δ and λ (putting aside computational difficulties) would be to determine those pairs for which the true image X^0 is the global minimum of $\mathcal{H}^k$ with high probability, provided that X^0 belongs to some (idealized) image class C_k. For instance, C_3 might consist of all images which are "locally" quadric, with components of a minimal size relative to the blur support. Another possibility is to select Δ on an ad hoc basis (regard it as a "knob") and then to choose λ to satisfy the constraint $\|Y - \mathcal{K}\hat{X}\|^2 \approx N^2\sigma^2$, where $\hat{X}$ minimizes $\mathcal{H}^k$. Of course, this is motivated by the simple observation that by the law of large numbers, this constraint is satisfied by X^0. Still another possibility is to adopt a Bayesian viewpoint, regard one or both of Δ and λ as unknown (hyper-)parameters of the probability distribution $\Pi_T(X)$ and attempt to *estimate* them from the data using standard statistical procedures; see e.g., [13]. (See also the survey of Thompson *et al.* [29] for a comprehensive analysis of several widespread methods for the choice of λ, especially in the quadratic case.) Finally, other techniques have been employed, such as applying the MDL principle [20] and "tracking" the reconstructions for varying λ [9].

Instead, we have chosen to seek conditions on Δ and λ such that under certain assumptions on the true intensity surfaces, X^0 is a *coordinate-wise* minimum. In particular, any coordinate-wise descent algorithm will remain at X^0 if it arrives there. This criterion is mathematically more tractable than those based on global minima, constraint satisfaction, or statistical estimation theory and seems particularly suited to our optimization procedure, namely low-temperature, single-site stochastic relaxation. All our experiments employ the parameters derived by this analysis. As it turns out, the constraint on $\| Y - \mathcal{K}\hat{X} \|$ is approximately satisfied in all cases.

First, here is some notation used in the statement of Theorem 2. We assume that $\eta = (\eta_s)_{s \in S}$ consists of independent and identically distributed Gaussian random variables with mean 0 and variance σ^2. The blur operator is

$$(\mathcal{K}X)_s = \sum_t \gamma_{s-t} X_t$$

and we define $\beta = \sum_t \gamma_t^2$. No assumptions are made about the PSF (except space invariance); thus, there are no constraints on the weights γ_t other than summing to 1. Fix the model order k, write $\mathcal{H}$ for $\mathcal{H}^k$, and let u^s denote an image that is zero in every coordinate except s, where it assumes the value u.

Definition 1: X is a coordinate-wise minimum of $\mathcal{H}$ if

$$\mathcal{H}(X + u^s) > \mathcal{H}(X)$$

for all u^s, $s \in S_0$, $u \neq 0$.

The definition depends on the particular noise realization η, in addition to the parameters λ, Δ, $\mathcal{K}$, and X^0 itself. Let Ω be the underlying probability space. The distribution of the process η is given by the product measure

$$Pr(\eta \in dv) = (2\pi\sigma^2)^{-|S|/2} \prod_{s \in S} e^{-v_s^2/2\sigma^2}\, dv_s.$$

For convenience, let us identify Ω with the set of noise realizations:

$$\Omega = \{\eta | \eta : S \to R\}.$$

The random nature of various functions will then be indicated by including the argument η; for example, let

$$\delta_{s,u}\mathcal{H}(\eta) = \mathcal{H}(X^0 + u^s, \eta) - \mathcal{H}(X^0, \eta)$$

where

$$\mathcal{H}(X, \eta) = \sum_C \phi(D_C(X)/\Delta) + \lambda \|\mathcal{K}X^0 + \eta - \mathcal{K}X\|^2.$$

Consider the event

$$\begin{aligned}\Lambda &= \{\eta | X^0 \text{ is a coordinate-wise minimum of } \mathcal{H}(X, \eta)\} \\ &= \{\eta | \delta_{s,u}\mathcal{H}(\eta) > 0 \; \forall\, s \in S_0, \quad \forall\, u \neq 0\}.\end{aligned}$$

Our goal now is to find conditions on X^0, and on λ and Δ in terms of $\mathcal{K}$, σ, and the model order k, such that

$$Pr(\Lambda) \approx 1. \tag{4.1}$$

Specifically, we want to determine an upper bound on λ as a function of Δ, $\mathcal{K}$, σ so that (4.1) is true. (Actually, we finally obtain a $\Delta - \lambda$ curve for which (4.1) holds; however, we regard Δ as easier to select "by hand" because of its interpretation as a scaling parameter.) The local geometry of X^0 (corner, step edge, ramp, etc.) is a crucial factor, and we want to examine cases that exhibit generic difficulties. One natural case, the only one we will analyze in full detail, is a simple step edge.

As we shall see in the Appendix

$$\delta_{s,u}\mathcal{H}(\eta) = f_s(u) + \lambda(\beta u^2 - 2u Z_s(\eta))$$

where the function $f_s(u)$ depends on ϕ and Δ as well as the local geometry of X^0 at pixel s, and where

$$Z_s(\eta) = \sum_t \gamma_{t-s}\eta_t.$$

The collection of random variables $\{Z_s : s \in S_0\}$ is a Gaussian process with means 0 and covariance function given by

$$\int Z_s Z_t \, dP = \sigma^2 \sum_r \gamma_r \gamma_{s-t+r}.$$

In particular, each $Z_s(\eta)$ is normal with mean 0 and standard deviation $\sqrt{\beta}\,\sigma$. It is easy to check (see the Appendix) that the condition $\delta_{s,u}\mathcal{H}(\eta) > 0$ is implied by

$$F_s(u) = \frac{1}{2\lambda u}\left(f_s(u) + \lambda\beta u^2\right) > |Z_s(\eta)|, \qquad u \neq 0$$

where F is defined by the first equality. Thus, Λ contains an event expressed in terms of the process $\{Z_s\}$ and the values $L_s = \inf_u F_s(u)$. For idealized images (see below), the dependence of L_s on s is simplified, and L_s assumes a few characteristic values corresponding to interior points, edges, etc. Consequently, the corresponding probability $Pr(\Lambda)$ is related to the distribution of $\sup_s Z_s$ over subsets of S_0.

The calculation is still complicated because the distribution of the supremum is inaccessible (which necessitates estimating $Pr(\Lambda)$) and because the computation of L_s is nontrival. In the first-order case, one can in fact obtain a general result for a "binary" image involving a step of size J, that is, an image such as:

$$\begin{array}{ccccccc}
\circ & \circ & \circ & \circ & \circ & \circ & \circ \\
\circ & \circ & \bullet & \bullet & \circ & \circ & \circ \\
\circ & \bullet & \bullet & \bullet & \circ & \circ & \circ \\
\circ & \bullet & \bullet & \bullet & \bullet & \circ & \circ \\
\circ & \circ & \bullet & \bullet & \bullet & \bullet & \circ \\
\circ & \circ & \circ & \circ & \bullet & \bullet & \circ \\
\circ & \circ & \circ & \circ & \circ & \circ & \circ
\end{array}$$

where each pixel labeled $\bullet$ has value $J > 0$, all others have value 0, and every site has at least two alike neighbors. It is a consequence of the concavity of ϕ that, in the first-order case, images such as these are coordinate-wise minima of the *regularization* term Φ. Thus, it is not difficult to imagine that if there is not too much noise, then there are settings of the parameters for which images such as these are also coordinate-wise minima of $\mathcal{H}$. However, the situation is more delicate for the higher order models, and we will not analyze the general (binary) situation for these. (The reader might want to consider the simple case of a "diagonal" edge for the second-order derivative and verify that in fact the image is *not* a coordinate-wise minimum of the regularization itself, although there are still values of λ for which it is a coordinate-wise minimum of $\mathcal{H}$.)

Instead, we focus on the case of a horizontal or vertical step edge. Specifically, we mean an image X^0 of the form

$$\begin{array}{ccccccc}
\circ & \circ & \circ & \circ & \bullet & \bullet & \bullet \\
\circ & \circ & \circ & \circ & \bullet & \bullet & \bullet \\
\circ & \circ & \circ & \circ & \bullet & \bullet & \bullet \\
\circ & \circ & \circ & \circ & \bullet & \bullet & \bullet \\
\circ & \circ & \circ & \circ & \bullet & \bullet & \bullet \\
\circ & \circ & \circ & \circ & \bullet & \bullet & \bullet \\
\circ & \circ & \circ & \circ & \bullet & \bullet & \bullet
\end{array}$$

where, as above, each pixel labeled $\bullet$ has value $J > 0$ and all others have value 0. The following theorem states that, in this case, for *any* σ, there is a value of λ (actually an upper bound) for which the probability that X^0 is a coordinate-wise minimum is arbitrarily close to one.

For any $\epsilon > 0$, let $d = d(\epsilon)$ be defined by the equation

$$\frac{2}{\sqrt{2\pi}}\int_0^d e^{-t^2/2}\,dt = 1 - \frac{\epsilon}{|S_0|}. \tag{4.2}$$

The proof of Theorem 2 appears in the Appendix, together with additional comments on the general (binary) case.

Theorem 2: Let X^0 be a "step edge" of size J, and let Λ denote the set of noise realizations for which X^0 is coordinate-wise minimum of $\mathcal{H}(X)$ in (1.3) with ϕ given by (1.4). Let $\epsilon > 0$. Then

$$Pr(\Lambda) \geq 1 - \epsilon$$

provided that

$$\lambda \leq \bar{\lambda} = \begin{cases} c\big/\left(2\Delta\sqrt{\beta}\,\sigma d\right) & \text{if } \sigma \leq \sqrt{\beta}\,\Delta\big/2d \\ c\big/\left(\sqrt{\beta}\,\Delta\,/\,2 + \sigma d\right)^2 & \text{if } \sigma \geq \sqrt{\beta}\,\Delta\big/2d \end{cases} \tag{4.3}$$

where c is a constant that depends only on the order of the model, and d is determined by (4.2). In the first-order case $c = 2$, in the second-order case $c = 5$, and in the third-order case $c = 14$.

Comments:

1) The analysis is based on a "worst-case" scenario in which *all* pixels are assumed to lie near boundaries. As a result, the number $|S_0|$ in (4.2) can be replaced by one substantially smaller, and we have found that choosing $d = 3$ is quite sufficient.
2) Notice that $\bar{\lambda}$ increases with the order of the model. This is consistent with the fact that the coefficients of D_C^k increase in magnitude with k (see Section II), and a change in u^s will therefore induce a relatively larger change in Φ^k as k increases. Notice also that $\bar{\lambda}$ decreases as σ and β increase, as we might expect.
3) The choice of λ is actually *independent* of the jump size J. A more careful analysis provides sharper bounds, but arbitrarily small jumps will be preserved with the result as stated.

V. Stochastic Relaxation

Our reconstruction is any $\hat{X}$ that minimizes $\mathcal{H}$:

$$\mathcal{H}\left(\hat{X}\right) = \min_X \mathcal{H}(X)$$

where $\mathcal{H} = \mathcal{H}^k$ is defined by (1.2) and (1.3) and D_C^k was described in Section II. A good approximation to the global minimum can be obtained by stochastic relaxation (specifically the Gibbs sampling algorithm) with annealing.

Stochastic relaxation is a Monte Carlo method designed for sampling from probability distributions of Markov random fields, such as those of the form

$$\Pi_\tau(X) = e^{-\mathcal{H}(X)/\tau}\Big/\sum_X e^{-\mathcal{H}(X)/\tau}.$$

When the goal is to minimize a function $\mathcal{H}$, typically (as here) nonconvex and defined over a very large but finite configuration space, stochastic relaxation is combined with annealing by introducing a control parameter (corresponding to temperature in a real physical system) during the sampling process, which increasingly concentrates the mass in the vicinity of $\hat{X}$. Notice that $\hat{X}$ is the mode of the distribution Π_τ for every τ. This is the basic optimization algorithm used in some earlier work with the line process as well as in much other related work; see e.g., [6] and [15]. The simulated annealing algorithm is computationally demanding but has the desirable feature of converging to a global minimum of $\mathcal{H}$. However, this is (by definition) an *asymptotic* statement and is usually difficult to realize in practice, partly because the theoretically correct annealing schedule τ_j requires a logarithmic decay of temperature to guarantee eventual escape from local minima. In particular, we have no guarantee of obtaining an actual minimum with a finite amount of computation; in fact, it is highly doubtful that we ever achieve the minimum energy, and indeed, the original image X^0 usually assumes a lower energy value than the estimate $\hat{X}$. Nonetheless, we have found this algorithm more reliable for our problem than other optimization methods.

Let us briefly review the ingredients of stochastic relaxation. One generates a Markov chain $X(j)$, $j = 0, 1, \cdots$, whose values are intensity images X representing, in our case, successive restorations. The initial value $X(0)$ is arbitrary in principle, although in practice, this choice can be pivotal; we shall return to this point later. It is well known that if the transition dynamics are suitably chosen, then the Markov chain converges to the uniform probability measure over the set $\Omega_\mathcal{H}$ of global minima of $\mathcal{H}$; in particular, $Pr(X(j) \in \Omega_\mathcal{H}) \to 1$ as $j \to \infty$. At each stage j of the algorithm, one updates the value of the pending restoration $X(j)$ at a single, predetermined, pixel s by computing a sample from the conditional probability distribution, with respect to Π_τ, $\tau = \tau_j$, of the random variable X_s, given the current values $X_t = X_t(j)$, $t \neq s$ at the other sites. This operation is repeated indefinitely, visiting the entire set of pixels in some predetermined fashion, usually just cyclically. Each cycle through the pixels is referred to as a "sweep." This is the single-site version. In theory, one can also update a group of sites using the multidimensional conditional distributions and thereby accelerate convergence; however, the amount of computation necessary is prohibitive. For example, to update a four by four array of pixels would necessitate sampling from a space of size G^{16}, where G is the number of grey levels.

Even single-site stochastic relaxation is computationally demanding, especially for a full dynamic range. However, the algorithm is highly parallelizable and given remarkable improvements in parallel hardware, single-site update algorithms present no significant time limitations on advanced machines. Moreover, we have found that a very simple approximation to the usual recipe yields a considerable speedup with little if any apparent degradation in the quality of the results. When updating the value of X at site s, instead of sampling from the actual (conditional) distribution of X_s, which puts positive weight on every intensity value, we reduce the support of the distribution to the values obtained by taking the union of small intervals about the current value at site s, the current values at the neighbors of s, and the data value Y_s. Specifically, in the case of a full dynamic range with 256 grey levels, the four nearest neighbors, and an interval of radius five about each of the resulting six values, the distribution that must be constructed for sampling has, on the average, 15 to 25 weights rather than 256. This yields an order of magnitude decrease in the number of operations performed with no apparent change since the true distribution places virtually zero mass on the complement of the reduced support.

Note: This procedure has recently been explored in a theoretical setting by Yang [30]. By slightly modifying the truncation procedure, the reduced support can be associated with local sections in a (slightly) restricted image space; moreover, the usual results on simulation and annealing remain intact.

Another computational problem is that the number of sweeps required to escape from very "wide" or "deep" local minima may be prohibitive, and there is a tendency to get trapped in local minima, especially when the algorithm is initialized with the data Y. As we noted earlier, the energy surfaces associated with the higher order models are extremely complex, and the energy values assumed by images with much smoother transitions than in the original image may in fact be only slightly larger than the minimum energy value. Consequently, because of the blur, the data provides a poor starting point. We have explored a stepwise analysis in the order of the model that overcomes some of these problems and by which, in many cases, we obtained better results (see Section VI). This approach has three steps:

1) Obtain a reconstruction with the first-order model starting at the data.
2) Use this image as the starting point for obtaining a reconstruction with the second-order model.
3) Use the second-order reconstruction as the starting point for obtaining a reconstruction with the third-order model.

This procedure is often more dependable than starting directly with the second- or third-order models. The reason is this: the first-order model is very efficient at detecting first-order discontinuities; moreover, if the parameters are suitably chosen (see Section IV), the second- and third-order models will preserve the discontinuities. The higher order models cannot detect jumps as effectively due to the existence of near-global minima, which display excessive ramping and interpolation and hence may appear visually quite distinct from the original image. Furthermore, escape from these local minima would necessitate large moves in the configuration space, which is virtually impossible with single-site updates. On the other hand, whereas the reconstruction generated with the first-order model will display jumps, the basic geometric structure of the original intensity surfaces is missing, even if it appears in the data. This is because the first-order distributions are highly concentrated on locally constant solutions to the reconstruction problem. For example, in the case of a linear ramp, the low energy states of the first-order model have a terraced quality. Significantly, these artifacts are overcome by the higher order models, and some of the fine geometric structures, particularly planar and quadric patches, of the original image are recovered.

Remark: In theory, simulated annealing is independent of the initial value, i.e., the starting point; at high temperatures, many changes occur, and the starting point is "forgotten." However, in practice, if one has a good starting point, such as the data Y for the first-order reconstruction, or especially the outcomes of previous reconstructions as discussed above, then one wants to "remember" this image and in fact preserve some of its attributes. For this reason, we initiate the annealing process at a relatively "low" temperature (see Section VI). This can be determined by examining the local conditional distributions at a few representative image locations; the idea is to keep the variance relatively small.

VI. Experiments

We present experiments on five images of varying difficulty, which demonstrate the merits (and demerits) of the algorithm. The first image (see Fig. 3) is a simple locally constant ("Mondrian") image and the second (see Fig. 5) is locally planar. Both of these are 64 × 64 synthetic images. (Figs. 4 and 6 show them as surface plots.) The third (Fig. 7) is the image of a building, of size 100 × 100, and digitized from a 35-mm slide. The fourth (Fig. 8, a face) and the fifth (Fig. 10, a soccer ball) are both obtained from a standard vidicon camera and are of sizes 100 × 100 and 128 × 128, respectively. All the images have 256 grey levels.

Each of the experiments involve both blur and noise. There are four types of blur. Two are generated by convolving the 3 × 3 mask

$$\frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

with itself, 2 and 3 times, respectively. Two iterations yields the 7 × 7 mask

$$\frac{1}{729}\begin{pmatrix} 1 & 3 & 6 & 7 & 6 & 3 & 1 \\ 3 & 9 & 18 & 21 & 18 & 9 & 3 \\ 6 & 18 & 36 & 42 & 36 & 18 & 6 \\ 7 & 21 & 42 & 49 & 42 & 21 & 7 \\ 6 & 18 & 36 & 42 & 36 & 18 & 6 \\ 3 & 9 & 18 & 21 & 18 & 9 & 3 \\ 1 & 3 & 6 & 7 & 6 & 3 & 1 \end{pmatrix}$$

and three iterations yields a 9 × 9 mask with an approximately Gaussian shape, referred to as the "Gaussian 9 × 9" blur. One can show that the standard deviation of the (marginal) distribution after n iterations of the 3 × 3 uniform mask is $v = \left(\frac{2}{3}(n+1)\right)^{1/2}$ in pixel units. Thus, $v = \sqrt{2}$ for the 7×7 mask and $v = \sqrt{8/3}$ for the 9 × 9 mask. The third bur is a 7 × 7 uniform blur, and the fourth is a 1 × 30 uniform blur.

White Gaussian noise was added with means zero and variance σ^2 determined by first specifying the decibel level. Recall that $\mathrm{dB} = 10\log_{10}(\mathrm{SNR})$ in which SNR denotes the signal-to-noise ratio, defined by

$$\begin{aligned} \mathrm{SNR} &= \hat{\sigma}^2(Y)/\sigma^2 \\ &= \frac{1}{|S|}\sum_s \left(Y_s - \overline{Y}\right)^2 / \sigma^2 \end{aligned}$$

where $\overline{Y}$ is the mean of the data (=signal). The value of SNR is essentially unchanged if $\hat{\sigma}^2(Y)$ is replaced by $\hat{\sigma}^2(\mathcal{K}X^0)$ since the difference between these is of much smaller order than $\hat{\sigma}^2(\mathcal{K}X^0)$ (unless σ^2 is absurdly large). Consequently, given dB, the data is obtained by adding noise with variance $\sigma^2 = \hat{\sigma}^2(\mathcal{K}X^0)/(10^{dB/10})$. Quantization error (rounding the grey levels to integers) may be regarded as uniform noise with $\sigma = 0.29$. We considered two nontrivial noise levels: 40 dB, which in a typical image with 256 grey levels corresponds roughly to $\sigma = 0.5$, and 25 dB, resulting in $\sigma \approx$ 3–4 grey levels in our images.

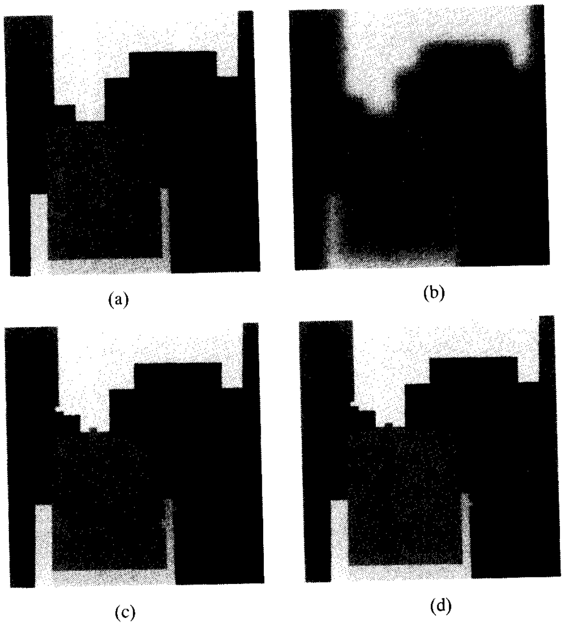


Fig. 3. Locally constant image: (a) Original image; (b) data: Gaussian 9 × 9 blur + 25-dB noise; (c) restoration: first-order model; (d) restoration: first → second-order model.

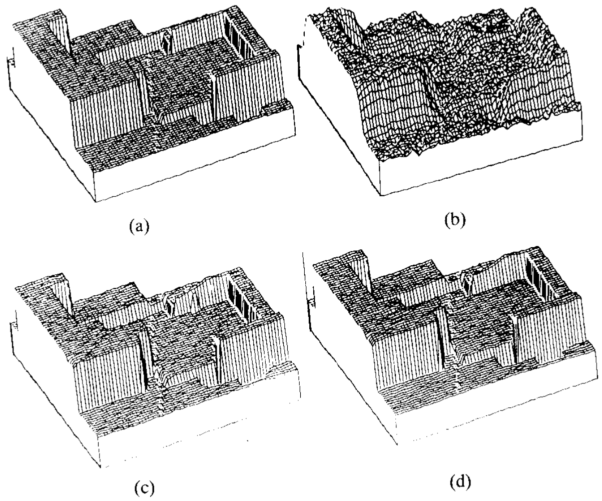


Fig. 4. Locally constant image; surface plots: (a) Original image; (b) data: Gaussian 9 × 9 blur + 25-dB noise; (c) restoration: first-order model; (d) restoration: first → second-order model.

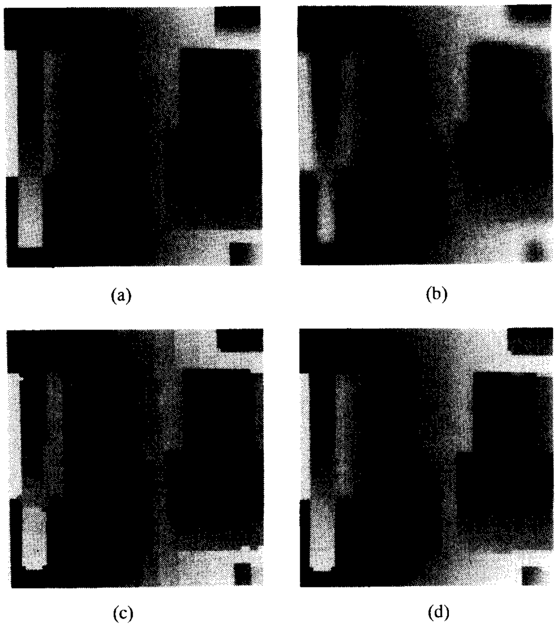


Fig. 5. Locally planar image: (a) Original image; (b) data: Gaussian 7 × 7 blur + 25-dB noise; (c) restoration: first-order model; (d) restoration: first → second-order model.

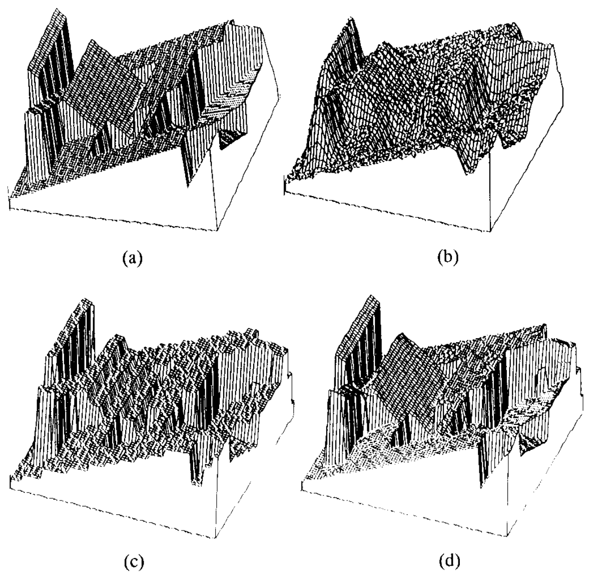


Fig. 6. Locally planar image; surface plots: (a) Original image; (b) data: Gaussian 7 × 7 blur + 25-dB noise; (c) restoration: first-order model; (d) restoration: first → second-order model.

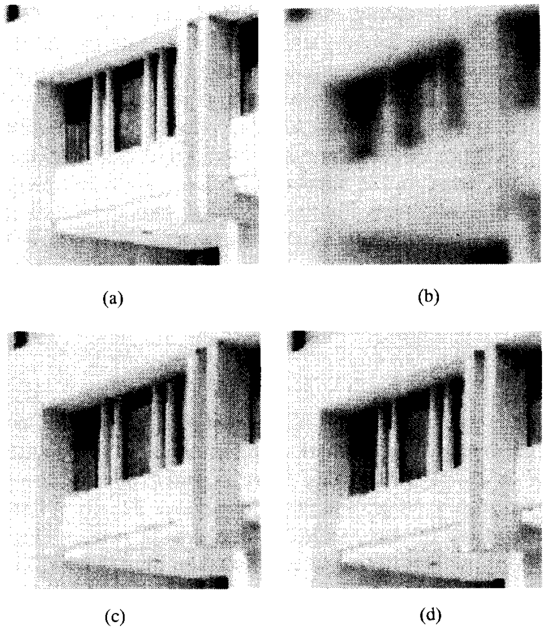


Fig. 7. Building image digitized from a 35-mm slide: (a) Original image; (b) data: Gaussian 7 × 7 blur + 40-dB noise; (c) restoration: second-order model; (d) restoration: first → second-order model.

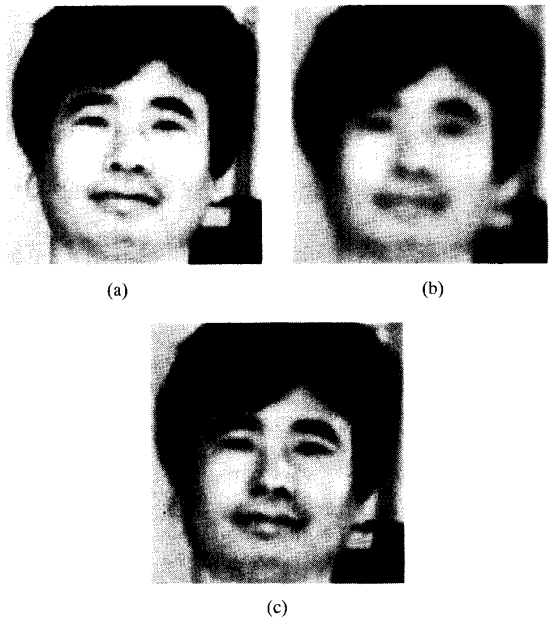


Fig. 8. Face image obtained with a standard vidicon camera: (a) Original image; (b) data: Gaussian 9×9 blur + 40-dB noise; (c) restoration: third-order model.

In every experiment, λ was chosen according to the value given the Theorem 2 in Section IV, with $d = 3$ and $\beta = 0.037$ (Gaussian 7 × 7 blur), $\beta = 0.028$ (Gaussian 9 × 9 blur), $\beta = 0.020$ (uniform 7 × 7 blur), and $\beta = 0.033$ (1 × 30 motion blur). The choice of Δ is ad hoc. In a standard image with 256 grey levels, it seems reasonable that an edge of 20 to 30 grey levels is significant. On the other hand, a change of 2 or 3 in the *slope* of a planar surface is visually significant. In fact, in a large number of experiments on these images, we have found that setting $\Delta \approx 10 - 20$ in the first-order models, and $\Delta \approx 3 - 10$ in the second- and third-order model, gives consistently good results.

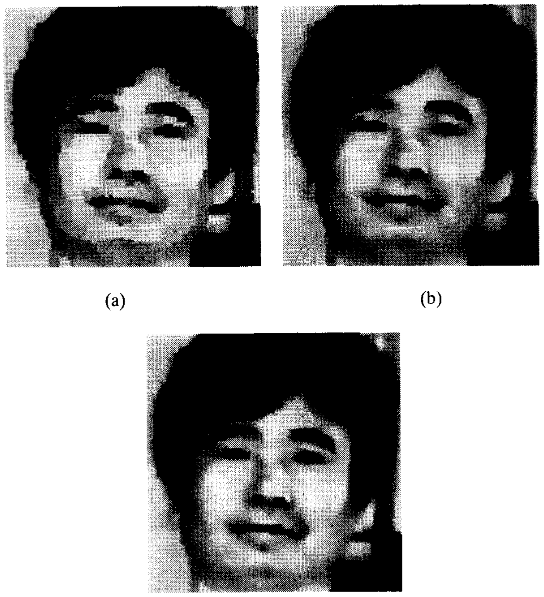


Fig. 9. Output of the hierarchical approach applied to the face image: (a) Restoration: first-order model; (b) restoration: first → second-order model; (c) restoration: first → second-order → third-order.

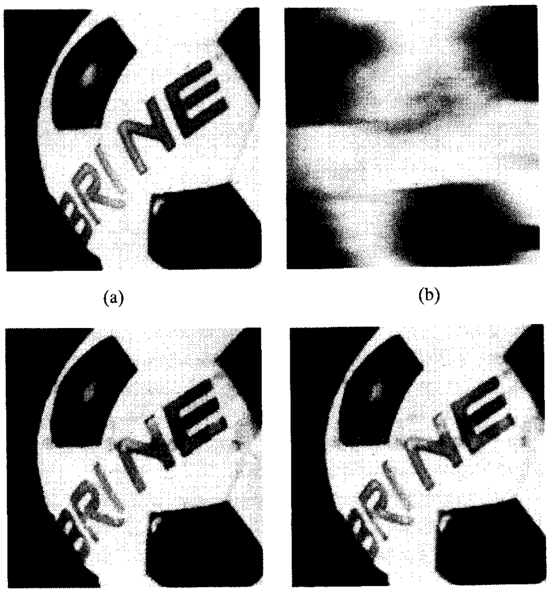


Fig. 10. Soccer ball image obtained with a standard vidicon camera: (a) Original image; (b) data: 1 x 30 motion blur + 60-dB noise: (c) restoration: first-order model; (d) restoration: second-order model.

Finally, all the experiments were run using 200 sweeps (cycles of the pixel lattice) for each order and dropping temperature linearly from an initial value τ_0 (0.3 in these experiments) to a final value $\tau \approx 0$.

A. Locally Constant Image

See Figs. 3 and 4. This result clearly indicates the utility of the first-order model when the original is indeed a Mondrian. The image was blurred with the Gaussian 9 x 9 mask, and 25 dB noise was added. In this case, the noise standard deviation is $\sigma = 3.9$ grey levels. For the first-order model, Δ was chosen to be 10 which implies $\lambda = 0.013$. The reconstruction of the first-order model was used to initialize the second-order model with $\Delta = 4$ and $\lambda = 0.33$.

The second-order model produced little change; a slight improvement can be observed in a few places. Similarly, the third order-model was run on the output of the second-order model, resulting in virtually no change.

B. Locally Planar Image

See Figs. 5 and 6. The experiment performed on this image uses the Gaussian 7 x 7 blur at 25 dB ($\sigma = 3.52$). In this experiment, $\Delta = 15$ in the first-order model, and $\Delta = 10$ in the second-order model. Note that the first-order model nearly succeeds in restoring the discontinuities but terraces the linear ramps. The second-order model kept the jump discontinuities and restored the linear ramps. It is worthwhile noting that repeated attempts to get a similar reconstruction using only the second-order model (starting at the data) were never as successful.

C. Building Image

The image (see Fig. 7) was blurred with the 7 x 7 mask, and 40 dB noise was added ($\sigma = 0.74$). The value $\Delta = 20$ ($\lambda = 0.01$) was chosen for the first-order model, and $\Delta = 6$ ($\lambda = 0.038$) for the second-order model.

This image has the appearance of being locally planar, and close examination shows that this is approximately true. It is reasonable that the second-order model is the most appropriate. Indeed, the application of the third-order model yielded no significant improvement of the restoration. It might be noted that the window is difficult to reconstruct since the vertical structures are only 1–2 pixels wide. It is useful to compare the result of the second-order model alone with the hierarchical one. Clearly, the output of the hierarchical approach is "sharper"; however, the visual quality of the second-order reconstruction may be better in some areas, for instance, in the window detail.

D. Face Image

The image (see Fig. 8) was blurred with the Gaussian 9 x 9 mask, and 40 dB noise was added ($\sigma = 0.32$). The third-order model seems to be the most appropriate. The parameters are the same as for the building image. We show both the hierarchical reconstruction and the result of using only the third-order model; see Figs. 8 and 9.

E. Soccer Ball

The blur is 1 x 30 horizontal motion blur, and the noise is very small (just 60 dB ($\sigma = 0.07$)). The result is typical of those possible in high signal-to-noise situations. Fig. 10 compares the output of the first- and second-order models both starting at the data. In this case, $\Delta = 10$ was chosen for both models, and $\lambda = 1.0$ for the first-order model, and $\lambda = 2.5$ for the second-order model. The output of the second-order model is slightly more "realistic" in the sense that extended gradients are more faithful to the original (e.g., less terraced).

VII. Conclusion

We have considered the problem of image deblurring, which is a classical example of the type of ill-conditioned inverse problem that frequently arises in low-level computer vision and many other fields. There is a substantial loss of spatial resolution in passing from the original, continuous radiance pattern to the observed image values, due principally to defocusing and other effects that distort point sources, but also to sampling and measurement error.

A conventional method to stabilize the problem is to introduce *a priori* smoothness constraints on the true image X^0 and construct a cost functional $\mathcal{H}(X) = \Phi(X) + \lambda\|Y - \mathcal{K}X\|^2$, which is a weighted average of prior constraints (Φ) and posterior constraints. The reconstruction is the value $\hat{X}$, which minimizes $\mathcal{H}$. There are several well-known weakness in this approach: recovering discontinuities (which is especially difficult with quadratic stabilizers) and choosing the model parameters (especially λ). In addition, there is the problem of model validation: Do the extremal or near-extremal states of $\mathcal{H}$ necessarily resemble X^0, at least in very simple cases?

The approach we have taken involves constraints Φ that are locally composed of functions of the type $\phi(D^k(X))$, where D^k denotes a kth order derivative ($k = 1, 2$, or 3), and $\phi(u) = -(1 + |u|)^{-1}$. Due to the concavity of ϕ on $(0, \infty)$ and its finite asymptotic behavior ($\lim_{u\to\infty} \phi(u) < \infty$), discontinuities may then be recovered *without* the use of a "line process" or other device for explicitly marking their locations.

Perhaps the main contribution of this paper is an effort to determine the parameters and validate the model by the same mechanism. An explicit formula is given for choosing the smoothing parameter λ in terms of the noise variance, the blur coefficients, and the order k of the model. This exploits the concavity of ϕ and was derived by requiring that an idealized original image be at least a coordinate-wise minimum for the cost functional.

In addition, we provide an analysis of the relationship between these models and those that involve a cost functional $\mathcal{H}^*(X, B)$ (such as coupled Markov random fields), where B is an auxiliary process designed to suspend the smoothness constraints in the vicinity of discontinuities. Basically, if $\phi(\sqrt{u})$ is concave (which includes our choice), then there exists a mixed cost functional $\mathcal{H}^*$ with the same data term as $\mathcal{H}$ and quadratic in X for each B fixed such that the set of minima in X coincide with those of $\mathcal{H}$.

Finally, when first-order discontinuities are important to detect and preserve, we introduce a computational method based on simulated annealing (although other procedures could be utilized) in which the outcome of the first-order model is used as a starting point for the second-order model, etc. With the model parameters suitably chosen, planar and quadric structures may then be recovered with the higher-order models while existing jump discontinuities are maintained.

Appendix

Proof of Theorem 1: We sketch the proof in the special case that ϕ is continuously differentiable, and $f(u) = \phi(\sqrt{u})$ is strictly concave. Let $M = f'(0+)$; then $0 < M \leq \infty$. It suffices to prove

$$f(z) = \inf_{0<b\leq M} (bz + \psi(b))$$

for some ψ as described.

Since f is strictly concave, $f'(u)$ is strictly decreasing and has an inverse $h(b) = (f')^{-1}(b)$. Define

$$\psi(b) = f(h(b)) - bh(b), \qquad 0 < b \leq M$$

which is the y intercept of the tangent line to f with slope b. Then, for any $z = h(b) > 0$, the line $y = bu + \psi(b)$ is tangent to $f(u)$ at the point $(z, f(z))$. In particular $f(z) = bz + \psi(b)$ so that

$$\inf_{0<b\leq M} (bz + \psi(b)) \leq f(z).$$

On the other hand, for any value $b_1 \neq b$, since the line $y = b_1 u + \psi(b_1)$ is tangent to $f(u)$ at $(z_1, f(z_1))$, $z_1 = h(b_1)$, and since $f(u)$ is strictly concave, we must have

$$f(z) < b_1 z + \psi(b_1).$$

Thus

$$f(z) = \inf_{0<b\leq M} (bz + \psi(b)).$$

Another computation shows that

$$\frac{d\psi}{db} = -\frac{df^{-1}}{du} < 0$$

and this establishes conclusion b). Finally, since $\lim_{u\to\infty} f'(u) = \lim_{u\to\infty} f(u) = 0$, we see that $\lim_{b\to 0} \psi(b) = 0$, which is conclusion a). Conclusion (c) follows directly from Condition 1. This completes the proof.

Verification of Example 1: This example can be verified by following the recipe of the proof. Observe that

$$h(b) = \frac{1}{\sqrt{b}} - 1.$$

Thus

$$\begin{aligned}\psi(b) &= f(h(b)) - b(h(b)) \\ &= \frac{1}{1 + \left(\frac{1}{\sqrt{b}} - 1\right)} - b\left(\frac{1}{\sqrt{b}} - 1\right) \\ &= b - 2\sqrt{b}.\end{aligned}$$

Verification of Example 2: Set $f(x, b) = \xi(b)x^2 + \psi(b)$. It suffices to show that if $0 < b_0 < 1$ and $x_0 = b_0^{-1/2} - 1$, then

1) $f(x_0, b_0) = -(1 + x_0)^{-1}$
2) $\frac{\partial f}{\partial x}(x_0, b_0) = (1 + x_0)^{-2}$.

The functions $f(x, b_0)$ and $\phi(x)$ are then tangent at $(x_0, \phi(x_0))$. Straightforward computations yield $f(x_0, b_0) = -\sqrt{b_0}$ and $\frac{\partial f}{\partial x}(x_0, b_0) = b_0$.

Proof of Theorem 2: Recall from Section IV that

$$\mathcal{H}^k(X,\eta) = \sum_C \phi\big(D_C^k(X)/\Delta\big) + \lambda \sum_{s\in S} \left(\left(\mathcal{K}X^0 + \eta\right)_s - \left(\mathcal{K}X\right)_s\right)^2$$

where $(\mathcal{K}X)_s = \sum_t \gamma_{s-t} X_t$. The energy difference at $s \in S_0$ is

$$\delta_{s,u}\mathcal{H}^k(\eta) = \mathcal{H}^k\left(X^0 + u^s, \eta\right) - \mathcal{H}^k\left(X^0, \eta\right)$$

where u^s denotes a state that is zero in every coordinate except s, where it assumes the value u, and X^0 is the "step-edge" image described in Section IV of the text. Since nothing is changed by adding a constant to $\mathcal{H}^k$ or incorporating Δ into ϕ, the analysis is slightly simplified by replacing ϕ with

$$\phi_+(u) = \phi(u/\Delta) + 1 = \frac{-1}{1 + |u/\Delta|} + 1 = \frac{|u/\Delta|}{1 + |u/\Delta|}.$$

In addition, we shall only establish the result assuming all pixels lie in S; the extension to $S_0 \backslash S$ is straightforward. Set

$$f_s^k(u) = \sum_C \left(\phi_+\left(D_C^k\left(X^0 + u^s\right)\right) - \phi_+\left(D_C^k\left(X^0\right)\right)\right)$$

and observe that for each s

$$\begin{aligned}\left\|Y - \mathcal{K}\left(X^0 + u^s\right)\right\|^2 - \left\|Y - \mathcal{K}X^0\right\|^2 = \\ \|\mathcal{K}X^0 + \eta - \mathcal{K}\left(X^0 + u^s\right)\|^2 - \|\eta\|^2 \\ = \sum_t \left(\left(\eta_t - \gamma_{t-s}u\right)^2 - \eta_t^2\right) \\ = u^2 \sum_t \gamma_{t-s}^2 - 2u \sum_t \gamma_{t-s}\eta_t \\ = \beta u^2 - 2uZ_s(\eta)\end{aligned}$$

where

$$Z_s(\eta) = \sum_t \gamma_{t-s}\eta_t$$

and $\beta = \sum_t \gamma_t^2$. The random variable $Z_s(\eta)$ is normal, mean 0, and variance $\sqrt{\beta}\sigma$. Thus, for each $s \in S$ and $u \neq 0$

$$\begin{aligned}\delta_{s,u}\mathcal{H}^k(\eta) &= \sum_C \left(\phi_+\left(D_C^k\left(X^0 + u^s\right)\right) - \phi_+\left(D_C^k\left(X^0\right)\right)\right) \\ &\quad + \lambda\left(\left\|\mathcal{K}X^0 + \eta - \mathcal{K}\left(X^0 + u^s\right)\right\|^2 - \|\eta\|^2\right) \\ &= f_s^k(u) + \lambda\left(\beta u^2 - 2uZ_s(\eta)\right).\end{aligned}$$

We now establish several lemmas, the first of which is a simple consequence of the fact that ϕ_+ is even and concave and increasing on the positive half-line.

Lemma 1: Let $J > 0$ and define $\alpha(u) = \phi_+(u) + \phi_+(J - u) - \phi_+(J)$. Then, $\alpha(u) \geq 0$ for all u.

Proof: First observe that if $u \leq 0$ or $u \geq J$, then the result follows from that fact that ϕ_+ is increasing and positive for $u \geq 0$. Next, the concavity of ϕ_+ implies for any three numbers $0 < a \leq b < c$

$$\frac{\phi_+(a)}{a} \geq \frac{\phi_+(c) - \phi_+(b)}{c - b}.$$

Now, apply this inequality to the points $a = u$, $b = J - u$, $c = J$ if $0 < u \leq J - u < J$ and to $a = J - u$, $b = u$, $c = J$ if $J - u < u < J$.

Lemma 2: Let $c_1 = 2$, $c_2 = 5$, $c_3 = 14$. Then

$$f_s^k(u) \geq c_k\phi_+(u), \quad \forall\ u,\ s \in S, \qquad k = 1, 2, 3.$$

Proof: We give the proof for the first-order case; the others are similar. Let X^0 be a step edge as described in the text. If s is a site that is not adjacent to the edge, then it is easy to check that

$$f_s^1(u) = 4\phi_+(u).$$

If s is adjacent to the edge, then

$$f_s^1(u) = \begin{cases} 3\phi_+(u) + \phi_+(J - u) - \phi_+(J) & \text{if } X_s^0 = 0 \\ 3\phi_+(u) + \phi_+(J + u) - \phi_+(J) & \text{if } X_s^0 = J \end{cases}$$

Thus, in all cases, and using Lemma 1, we obtain $f_s^1(u) \geq 2\phi_+(u)$.

By Lemma 2

$$\delta_{s,u}\mathcal{H}^k \geq c_k\phi_+(u) + \lambda\left(\beta u^2 - 2uZ_s(\eta)\right), \qquad s \in S, \quad k = 1, 2, 3.$$

Define

$$F_k(u) = \frac{1}{2\lambda u}\left(\lambda\beta u^2 + c_k\phi_+(u)\right), \qquad u > 0. \tag{8.1}$$

It follows that for each $s \in S$, $k = 1, 2, 3$

$$|Z_s(\eta)| < \inf\nolimits_{u>0} F_k(u) \Rightarrow \delta_{s,u}\mathcal{H}^k > 0, \qquad \forall\ u \neq 0. \tag{8.2}$$

Lemma 3: Let $F_k(u)$ be defined by (8.1). Then

$$L_k = \inf\nolimits_{u>0} F_k(u) = \begin{cases} c_k/(2\lambda\Delta) & \text{if } \lambda \geq c_k/\beta\Delta^2 \\ \sqrt{\beta c_k/\lambda} - \beta\Delta/2 & \text{if } \lambda \leq c_k/\beta\Delta^2. \end{cases}$$

Proof: We outline the proof, which is a calculus exercise. It suffices to assume $u > 0$. Note that

$$F_k(u) = \frac{\beta u}{2} + \frac{c_k}{2\Delta\lambda(1 + u/\Delta)}$$

and that

$$F_k'(u) = \frac{\beta}{2} - \frac{c_k}{2\Delta^2\lambda}\frac{1}{(1 + u/\Delta)^2}.$$

Notice also that F_k' is increasing. If $F_k'(0) > 0$, which is equivalent to $\lambda > c_k/\beta\Delta^2$, the infimum occurs at $u = 0$ and $L_k = F_k(0) = c_k/2\lambda\Delta$. On the other hand, if $F_k'(0) \leq 0$, then the infimum occurs at $u = \sqrt{c_k/\lambda\beta} - \Delta$, and substituting this value of u into F completes the proof.

To finish the proof of Theorem 2, define

$$\Lambda_s = \left\{\eta | \delta_{s,u}\mathcal{H}^k(\eta) > 0, \quad \forall\ u \neq 0\right\}.$$

Observe that by (8.2), for any $s \in S$

$$Pr(\Lambda_s) > Pr(|Z_s(\eta)| < L_k) = \frac{2}{\sqrt{2\pi}} \int_0^{L_k/\sqrt{\beta}\sigma} e^{-t^2/2}\, dt.$$

Let $\epsilon > 0$. Since

$$\bigcap_{s \in S_0} \Lambda_s = \Lambda$$

it suffices to show that

$$Pr(\Lambda_s) > 1 - \frac{\epsilon}{|S_0|}$$

for each $s \in S_0$. This follows immediately on setting $L_k/\sqrt{\beta}\sigma = d$, where d is given by (4.2). If $L_k = c_k/2\Delta\lambda$, then $\lambda = c_k/\left(2\sqrt{\beta}\Delta\sigma d\right)$ and $\sigma \leq \sqrt{\beta}\Delta/2d$. If, on the other hand, $L_k = \sqrt{\beta c_k/\lambda} - \beta\Delta/2$, then $\lambda = c_k/\left(\sqrt{\beta}\Delta/2 + \sigma d\right)^2$, and $\sigma \geq \sqrt{\beta}\Delta/2d$. This completes the proof of Theorem 2.

Remark: The difference between the above analysis for a "step-edge" and the general case illustrated in Section IV is the existence of "corner" pixels, where two neighbors assume the value $u = 0$, and the other two assume the value $u = J$. The argument is similar, but more delicate, than the one above. For example, in the first-order case, in the lower bound on $\delta_{s,u}\mathcal{H}^k$, the function ϕ_+ must be replaced by $\alpha \leq \phi_+$, where α was defined in Lemma 1. The calculation of the infimum L_k is messy, and the corresponding value is smaller than the one given in Lemma 3, resulting in a smaller upper bound for λ. In fact, there is then an *upper bound* on noise variance for the result to hold: $\sigma \leq \sigma^* = \sqrt{\beta}J/2d$. This is not surprising if one notes that the corner configuration is unstable because the prior term gives *equal* weight to the two cases $X_s^0 = 0$ and $X_s^0 = J$. Consequently, even though the true value at s is, say, $u = 0$, it may happen, depending on the particular noise realization, that $\delta_{s,J}\mathcal{H} < \delta_{s,0}\mathcal{H}$. We then expect, and indeed observe, erosion at the corners for low SNR ratios. The corresponding formula for $\overline{\lambda}$ in the first-order case is

$$\overline{\lambda} = \begin{cases} (\theta_1/\theta_2)(\sigma/\sigma^*)^{-1} & \\ & \text{if } (\sigma/\sigma^*) \leq \theta_2 \\ \theta_1\left(1 - \left((1-\theta_2)(1-(\sigma/\sigma^*))^{1/2}\right)\right)^{-2} & \\ & \text{if } \theta_2 \leq (\sigma/\sigma^*) \leq 1 \end{cases}$$

where $\theta_1 = 2(2\Delta + J)/\beta(\Delta + J)^3$, $\theta_2 = \Delta/(\Delta + J)$.

Acknowledgment

The authors wish to thank the referees (one in particular) for useful suggestions and for bringing additional references to our attention.

The authors are also grateful to W. Kendall [19] for illuminating remarks about Theorem 1.

References

[1] H. C. Andrews and B. R. Hunt, *Digital Image Restoration.* Englewood Cliffs, NJ: Prentice-Hall, 1977.

[2] M. Bertero, "Regularization methods for linear inverse problems," in *Lecture Notes in Mathematics, vol. 1225, Inverse Problems* (G. Taleuti, Ed.). Berlin: Springer-Verlag, 1986, pp. 52–112.

[3] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Statist. Soc.,* Ser. B, vol. 48, pp. 259–302, 1986.

[4] ——, "Towards Bayesian image analysis," *J. Appl. Statistics,* vol. 16, 1989.

[5] A. Blake and A. Zisserman, *Visual Reconstruction,* Cambridge, MA: MIT Press, 1987.

[6] R. Chellappa, T. Simchony, and H. Jinchi, "Relaxation algorithms for MAP restoration of gray level images with multiplicative noise," Tech. Rep., Signal Image Processing Inst., Univ. S. Calif., 1988.

[7] H. Derin and H. Elliott, "Modelling and segmentation of noisy and textured images using Gibbs random fields," *IEEE Trans. Patt. Anal. Machine Intell.,* vol. PAMI-9, pp. 39–55, 1987.

[8] D. Geiger and F. Girosi, "Mean field theory for surface reconstruction and visual integration," A.I. Memo 1114, Artificial Intell. Lab., Mass. Inst. Technol., 1989.

[9] D. Geiger and A. Yuille, "A common framework for image segmentation," in *Proc. Int. Conf. Patt. Recog. ICPR-90,* (Atlantic City), 1990, pp. 502–507.

[10] D. Geman and S. Geman, "Bayesian image analysis" in *Disordered Systems and Biological Organization.* (E. Bienenstock, F. Fogelman, and G. Weisbuch, Eds.). Berlin: Springer-Verlag, 1986, vol. F20.

[11] D. Geman, G. Reynolds, and C. Yang, "Stochastic algorithms for restricted image spaces and experiments in deblurring," to be published in *Markov Random Fields: Theory and Applications.* New York: Academic.

[12] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Machine Intell.,* vol. PAMI-7, no. 6, pp. 721–741, 1986.

[13] S. Geman and D. E. McClure, "Statistical methods for tomographic image reconstruction," in *Proc. 46th Sess. Int. Stat. Inst. Bulletin ISI,* 1987, vol. 52.

[14] S. Geman, D. E. McClure, and D. Geman, "A nonlinear filter for film restoration and other problems in image processing," Tech. Rep., Div. Appl. Math., Brown Univ. Providence, RI, 1990.

[15] B. Gidas, "A renormalization group approach to image processing problems," *IEEE Trans. Patt. Anal. Machine Intell.,* vol. 11, pp. 164–180, 1989.

[16] P. J. Green, "Bayesian reconstruction from emission tomography data using a modified EM algorithm," *IEEE Trans. Med. Imaging,* vol. 9, pp. 84–93, 1990.

[17] F. R. Hampel, E. M. Ronchetty, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions.* New York: Wiley, 1986.

[18] Jeng and J. Woods, "Compound Gauss-Markov random fields for image estimation," Tech. Rep., Rensselaer Polytech. Inst., 1988.

[19] W. Kendall, personal communication, July 1989.

[20] Y. G. Leclerc, "Constructing simple stable descriptions for image partitioning," Int. J. Comput. Vision, vol. 3, pp. 73–102, 1989.

[21] J. L. Marroquin, "Surface reconstruction preserving discontinuities," Artificial Intell. Lab. Memo 792, Mass. Inst. Technol., 1984.

[22] J. L. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Stat. Assoc.,* vol. 82, pp. 76–89, 1987.

[23] R. Molina and B. D. Ripley, "Using spatial models as priors in astronomical image analysis," *J. Appl. Stat.,* vol. 16, no. 2, pp. 193–206, 1989.

[24] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature,* vol. 317, pp. 314–319, 1989.

[25] A. Rangarajan and R. Chellappa, "Generalized graduated nonconvexity algorithm for maximum *a posteriori* image estimation," in *Proc. Int. Conf. Patt. Recog. ICPR-990,* (Atlantic City, NJ), 1990.

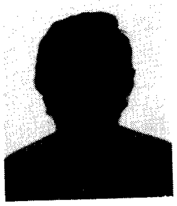
[26] J. Rissanen, "Minimum-description-length principle," in *Encylclopedia of Statistics.* New York: Wiley, 1987, pp. 523–527, vol. 5.

[27] D. Shulman and J. Y. Herve, "Regularization of discontinuous flow fields," in *Proc. IEEE Comp. Soc. Workshop Visual Motion '89,* 1989, pp. 81–86.

[28] D. Terzopoulis, "Regularization of inverse visual problems involving discontinuities," *IEEE Trans. Patt. Anal. Machine Intell.,* vol. PAMI-8, pp. 413–424, 1986.

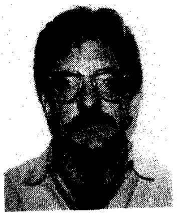
[29] A. M. Thompson, J. C. Brown, J. W. Kay, and M. Titterington, "A study of methods of choosing the smoothing parameter in image restoration by regularization," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 13, pp. 326–339, 1991.

[30] C. Yang, "Stochastic methods for image restoration," Ph.D. thesis, Dept. math. Stat., Univ. Mass., Amherst, MA, 1991.



Donald Geman (M'84) received the B.A. degree in english literature from the University of Illinois, Urbana, in 1965 and the Ph.D. degree in mathematics from Northwestern University, Evanston, IL, in 1970.

Currently, he is Professor of Mathematics in the Department of Mathematics and Statistics at the University of Massachusetts, Amherst, where he has been since 1970. Visiting appointments include those at University of North Carolina (1976–77), Brown University (1984), Universite de Paris-Sud (1986), and INRIA-Rocquencourt (1990, 1991). His research interests are in image reconstruction, object recognition, and stochastic processes.



George Reynolds received the Ph.D. degree in mathematics from Wesleyan University, Middletown, CT, in 1974.

He was an assistant professor at Union College until 1981 and was visiting associate professor at the University of Massachusetts at Amherst until 1991. He now works for VI Corporation, Northampton, MA. His interests include image processing, graphical user-interface development environments, and the violin.