A Generalized Gaussian Image Model for Edge-Preserving MAP Estimation †

Charles Bouman
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907-0501
(317) 494-0340

Ken Sauer
Laboratory for Image and Signal Analysis
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
(219) 239-6999

Abstract

We present a Markov random field model which allows realistic edge modeling while providing stable maximum *a posteriori* (MAP) solutions. The proposed model, which we refer to as a generalized Gaussian Markov random field (GGMRF), is named for its similarity to the generalized Gaussian distribution used in robust detection and estimation. The model satisfies several desirable analytical and computational properties for MAP estimation, including continuous dependence of the estimate on the data, invariance of the character of solutions to scaling of data, and a solution which lies at the unique global minimum of the *a posteriori* log likelihood function. The GGMRF is demonstrated to be useful for image reconstruction in low dosage transmission tomography.

† *IEEE Trans. on Image Processing*, vol. 2, no. 3, pp. 296-310, July 1993.

1 Introduction

Many important problems in image processing and computer vision require the estimation of an image or other 2D field, X, from noisy data Y. For example, tomographic reconstruction and 2D depth estimation are two seemingly dissimilar problems which fit into this structure. When the data is of good quality and sufficient quantity, these problems may be solved well by straightforward deterministic inverse formulae. However, when data are sparse or noisy, direct inversion is usually excessively sensitive to noise. If the data is sufficiently sparse, the inverse problem will be underdetermined or ill-posed. In such cases, the result can be significantly improved by exploiting prior information about the behavior of X.

Bayesian estimation is a statistical approach for incorporating prior information through the choice of an *a priori* distribution for the random field X. While many Bayesian estimation techniques exist, a common choice for image estimation problems is the maximum *a posteriori* (MAP) estimator. The MAP estimate has the appealing attribute that it yields the most likely image given the observed data. In addition, it results in an optimization problem which may be approached using a variety of established techniques.

The specific choice of prior distribution for X is, of course, a critical component in MAP estimation. The Markov random field (MRF) has been applied widely during the recent past[1, 2, 3, 4], due to its power to usefully represent many image sources, and the local nature of the resulting estimation operations. A variety of distinct models exist within the class of MRFs, depending on the choice of the *potential functions*. Each potential function characterizes the interactions among a local group of pixels by assigning a larger cost to configurations of pixels which are less likely to occur. We will restrict our attention to potential functions $\rho(\lambda(x_i - x_j))$, which act on pairs of pixels. The shape of $\rho(\Delta)$, where Δ is the difference between pixel values scaled by λ, then indicates the attributes of our model for X.

One of the more troublesome elements of applying MRFs to image estimation is coping with edges. Because most potential functions penalize large differences in neighboring pixels, sharp edges are often discouraged. This is especially true for the Gaussian MRF, which penalizes the square of local pixel differences. Many approaches to ameliorate this effect have been introduced. Geman and Geman[2], incorporated a "line process" into their MRF to describe sharp discontinuities. Others limited the penalty of any local difference at some prescribed threshold[5, 6], or created other potential functions which become flat at large magnitudes of their arguments[7, 8, 9]. Since such functions are non-convex, the entire cost function may be nonconvex, unless relatively very little weight is applied to the prior distribution portion of the cost. If the cost is nonconvex, the global optimization required in MAP estimation cannot be exactly computed, and an approximate MAP estimate must be used. We also show a second important liability to using MRFs with non-convex potential functions: the MAP estimate may not be a continuous function of the input data. This means that the position of the $\hat{X}$ with globally minimal cost may undergo a large shift due to a small perturbation in Y. Therefore, the MAP estimator with a nonconvex potential function can be an unstable and ill-posed inverse operation.

Several researchers have proposed the use of convex potential functions. Stevenson and Delp[10] used the convex Huber function[11], which is quadratic for small values of Δ, and linear for large values. The point of transition between the quadratic and linear regions of the function is a predetermined threshold, T. Green[12] and Lange[13] included the strict convexity criterion, also for the sake of computational tractability. Green's choice of $\log\cosh(\Delta)$ has a shape similar to that of the Huber function. Lange also derived several other potential functions in [13], each satisfying convexity and several other desired properties.

The restriction to convex potential functions makes the computation of the exact MAP estimate feasible, but the effect of the above approaches in MAP estimation is dependent on the *scaling* of X and Y. The transition threshold for the Huber function, for example,

should be related to the magnitude of edges expected in X. If this magnitude is unknown, or edges of widely varying magnitudes are expected, then the smoothing of these edges may be inconsistent. Similar difficulties hold for the other non-quadratic functions mentioned.

In this paper, we introduce an MRF model for Bayesian estimation, which is intended to ameliorate both of the problems discussed above. The general form of the potential function is $|\Delta|^p$, with $1 \leq p \leq 2$. The resulting form of the probability density function for X is similar to the generalized Gaussian distribution commonly used as a noise model in robust detection and estimation[14]. Due to this similarity, we use the name generalized Gaussian Markov random field (GGMRF) to describe these images. The parameter p controls the cost of abrupt edges. When $p = 1$ sharp edges are no more costly than smooth edges, and when $p = 2$ the familiar Gaussian assumption holds.

The log of the GGMRF has two important properties. It is convex, and it scales with the data. Convexity makes minimization efficient, and is sufficient to guarantee stable MAP estimation. The scaling property leads to a homogeneous MAP estimator when the observation noise has the generalized Gaussian distribution with a corresponding form. We also give the canonical form for all distributions which have the convexity and scaling properties.

We briefly explore the connection between median filtering and MAP estimation using the GGMRF prior together with the generalized Gaussian noise model. An operation very similar to the recursive weighted median filter results as the local update for computation of MAP estimate when $p = 1$. However, it is shown that the local median filter updates do *not* converge to the global MAP estimate. This connection is of interest since median filters are a useful class of homogeneous edge preserving nonlinear filters for image processing.

In the experimental section of this paper, we first present examples illustrating some of the properties above in one-dimensional functions. We then apply the GGMRF to the problem of image reconstruction from integral projections. Bayesian techniques have been applied to similar problems, but most previous assumptions for prior distributions have

been Gaussian[15, 16, 17]. We consider the transmission tomographic case, with low X-ray dosage, and attendant high photon counting noise. Both a synthetic phantom, and actual data from nondestructive testing experiments are included. Photon noise is especially problematic in projection rays passing through highly absorptive regions; in the limit these regions are effectively radio-opaque, and present the equivalent of the hollow projection (a.k.a. "bagel") problem. Reconstructions using the convolution backprojection algorithm suffer from a trade-off between excessive blurring and noise artifacts. A similar trade-off, with better results, can be made in using a Gaussian MRF as a prior density on X. The GGMRF, however, with smaller values of p, allows the formation of sharp edges, while more effectively suppressing the photon counting noise in the estimate. The success of the GGMRF in regularization of the tomographic reconstruction offers hope that it will be useful in many other image restoration and reconstruction problems.

2 Statistical Framework

We first define some basic notation. We will use upper case letters for random quantities and lower case letters for their deterministic realizations. A random field X will be defined on the set of N points S, and each pixel, X_s for $s \in S$, takes values in $\mathbb{R}$. The neighbors of X_s will be denoted by $X_{\partial s}$ where $\partial s \subset S$. Further, the neighbors of each point must be chosen so that they have the property that $\forall s, r \in S \; s \notin \partial s$ and $r \in \partial s \; \Leftrightarrow s \in \partial r$.

The maximum likelihood (ML) estimate of the image X from data Y is given by

$$\hat{x} = \arg\max_x p(y|x)$$

where $p(y|x)$ is the conditional density of Y given X. While the ML estimate accurately fits the data, it does not incorporate reasonable prior information about the image. In practice, this can produce excessive noise or nonuniqueness[18] of the result. Similar problems of underdetermined or ill-posed solutions occur in a wide variety of problems in motion estimation[19], surface reconstruction[20] and edge detection[21].

One approach to incorporating prior information is to adopt a Bayesian estimation approach such as maximum *a posteriori* (MAP) estimation. If we adopt $g(x)$ as a prior distribution for the unknown image, then the MAP estimate is given by

$$\begin{aligned} \hat{x} &= \arg\max_x p(x|y) \\ &= \arg\max_x \{\log p(y|x) + \log g(x)\} \\ &= \arg\max_x \log p(y,x) \end{aligned} \tag{1}$$

When the prior distribution of X is Gaussian, the log likelihood $\log g(x)$ will be a quadratic function of x. If $p(y|x)$ is also Gaussian, the MAP estimate corresponds to $E\{X|Y\}$, and is therefore the minimum mean squared error estimate[22]. When the prior distribution is not Gaussian, the MAP estimate is still optimal with respect to the zero/one loss function[22], but the appropriateness of this criterion in the general case is not clear[18]. However, MAP estimation is computationally direct and has experimentally been shown to work well in a variety of problems[1, 2, 3, 4, 23].

A critical issue is the choice of prior distribution for X. We will use Markov random fields (MRF) since they restrict computation to be local but still include a very wide class of possible models. Gibbs Distributions are used to explicitly write the distributions of MRF's. A Gibbs distribution is any distribution which can be expressed in the form

$$g(x) = \frac{1}{Z} \exp\left\{ -\sum_{c \in C} V_c(x) \right\}$$

where Z is a normalizing constant, $V_c(\cdot)$ is any function of a local group of points c and C is the set of all such local groups. The key to the definition of the Gibbs distribution is the specification of these local groups of points. A local set of points, c, is called a clique if $\forall s, r \in c$, s and r are neighbors. If Gibbs distributions are restricted to use functions of cliques induced by the neighborhood system ∂s, then the random field X will have the property that

$$\forall s \in S \;\; p(x_s | x_r \,, r \neq s) = p(x_s | x_r \,, r \in \partial s) \;.$$

This is the fundamental property of an MRF. In fact, the Hammersley-Clifford theorem states that if X is a strictly positive random field, then X is a MRF if and only if the distribution of X has the form of a Gibbs distribution[24, 25]. The technical condition that X be strictly positive insures that points in X are zero with probability zero. This is required since the Gibbs distribution is strictly positive.

3 Existing Image Models

3.1 Gaussian Markov Random Fields

A common choice for the prior model is a Gaussian Markov random field (GMRF)[15, 16, 17]. The distribution for a Gaussian random field has the form

$$g(x) = \frac{\lambda\sqrt{2}}{(2\pi)^{N/2}}|B|^{1/2}\exp\left\{-\lambda^2 x^t B x\right\} \ . \tag{2}$$

where B is a symmetric positive definite matrix, λ is a constant, and x^t is the transpose of x. In order for this to correspond to a Gibbs distribution with neighborhood system ∂s, we also impose the constraint that $B_{sr} = 0$ when $s \notin \partial r$ and $s \neq r$. This distribution may then be rewritten to form the log likelihood

$$\log g(x) = -\lambda^2 \left(\sum_{s\in S} a_s x_s^2 \ + \sum_{\{s,r\}\in C} b_{sr}|x_s - x_r|^2 \right) + \text{constant}$$

where $a_s = \sum_{r\in S} B_{sr}$ and $b_{sr} = -B_{sr}$. Notice that the second sum is now over all distinct pairs of neighboring pixels. MAP estimation of X then results from minimization of the following cost function:

$$\hat{x} = \arg\min_x \left\{ -\log p(y|x) + \lambda^2 \left(\sum_{s\in S} a_s x_s^2 \ + \sum_{\{s,r\}\in C} b_{sr}|x_s - x_r|^2 \right) \right\}.$$

While the GMRF prior has many analytical advantages, it generally results in estimates $\hat{x}$ which are either excessively noisy or generally blurred. This is because the squared difference of pixel values applies too high a penalty to edges that often occur in images.

3.2 Nonconvex Log Prior Distributions

Non-Gaussian MRF's are interesting because they can potentially model both the edges and smooth regions of images. Initial approaches often used an additional unobserved random field called a line process which determines the location of edges[2, 26]. More recently, many approaches have focused on MRF's with simpler Gibbs distributions of the general form

$$\log g(x) = -\sum_{\{s,r\}\in C} b_{sr}\rho(\lambda|x_s - x_r|) \quad + \quad \text{constant} \tag{3}$$

where λ is a scaling parameter, and ρ is a monotone increasing, but not convex function[7, 8, 5, 6, 9, 12, 13, 27]. A typical function used by Blake and Zisserman[5] is

$$\rho\left(\Delta\right) = \min\{|\Delta|, T\}^2$$

where T is a variable threshold parameter. This function is shown in Fig. 1(a) for $T = 0.5$. Notice that the function is quadratic near zero, but the flat region beyond the value T allows sharp edges to form in the reconstructed image. Intuitively, if two pixels differ by a value greater than T/λ, then it is likely that they lie on opposite sides of an edge, and therefore their values should not be required to be close.

For the purposes of modeling images, this distribution has some significant practical and theoretical disadvantages. Since the function is nonconvex it is generally impractical to globally minimize. The MAP estimate can only be approximated using a number of different techniques[2, 5, 6]. In fact, the solution achieved often depends substantially on the method used to perform the minimization.

In addition to this computational issue, there is a disadvantage in the quality of reconstruction that results from such a nonconvex prior. The prior term $\rho(\Delta)$ in the cost functional does not increase with larger local pixel differences after the difference exceeds T/λ. Therefore, any image edge of a given spatial configuration and of magnitude greater than this threshold incurs the same cost under this prior, and no preference is expressed

among these edge magnitudes. Consequently, the MAP estimate may abruptly change as the magnitude of an edge in the input data Y increases from below the value T/λ to above. This may lead to an unnatural quality in the reconstruction, in which reconstructed edges greater than T/λ are sharp, yet those of lower magnitudes are smooth.

Another undesirable quality in these reconstructions is due to the fact that the MAP estimate, $\hat{x}$, is not continuously dependent on the input, y[28]. To illustrate this point consider the nonconvex functions shown in Fig. 2. Fig. 2a shows a function with two local minima at positions x_1 and x_2. Fig. 2b shows a small perturbation on the first function with the same two local minima. Notice that while the difference between the two functions is small, the difference between locations of the two global minima is large.

The unstable behavior of the MAP estimate under the nonconvex prior of Blake and Zisserman is illustrated in Fig. 3. Fig. 3a shows two noisy pulses with amplitudes 4.2 and 4.3 and additive white Gaussian noise of unit variance. Fig. 3b shows the two reconstructions of those pulses using the prior model

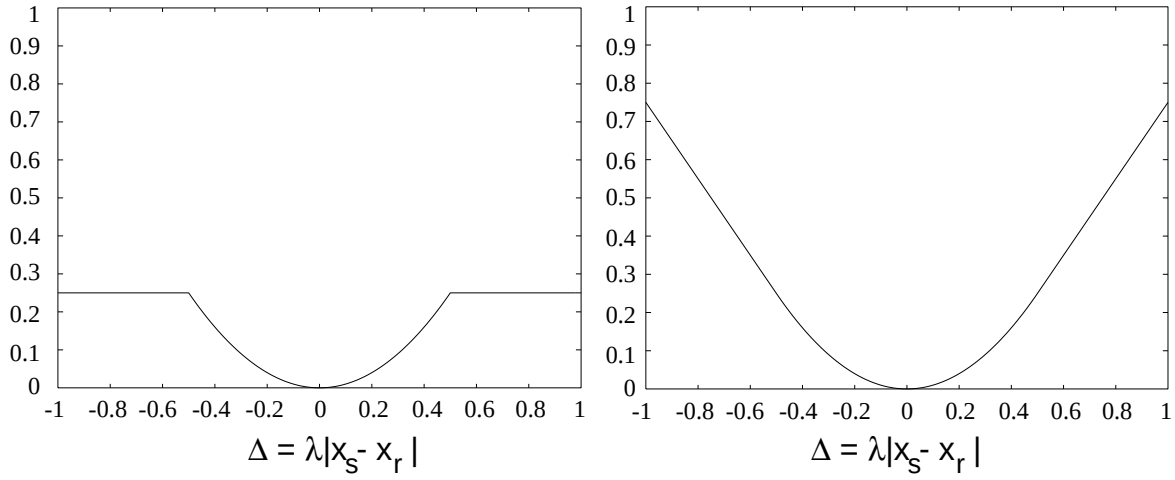
$$\log g(x) = -\sum_{s=1}^{49} \min\{5|x_s - x_{s+1}|, 1.75\}^2 \quad + \quad \text{constant} \ .$$

A small perturbation in the signal causes two very different reconstructions. Intuitively, one reconstruction has determined the presence of an edge while the other has not. However, this type of behavior is unnatural in the reconstruction of continuous tone images.

3.3 Convex Log Prior Distributions

More recently, convex functions have also been considered for $\rho(\cdot)$[10, 12, 13, 28]. Stevenson and Delp have studied the use of an alternative convex energy function for the problem of surface reconstruction[10]. They chose the Huber function first introduced in robust statistics[11].

$$\rho\left(\Delta\right) = \begin{cases} \Delta^2 & \text{if } |\Delta| \leq T \\ T^2 + 2T|\Delta - T| & \text{if } |\Delta| > T \end{cases}$$



(a) $\rho(\Delta) = (\min\{|\Delta|, 0.5\})^2$ (b) $\rho(\Delta) =$ Huber function

Figure 1: Examples of functions used for ρ. ρ is a function of Δ, the scaled difference between neighboring pixel values. (a) A nonconvex cost function. (b) A convex cost function.

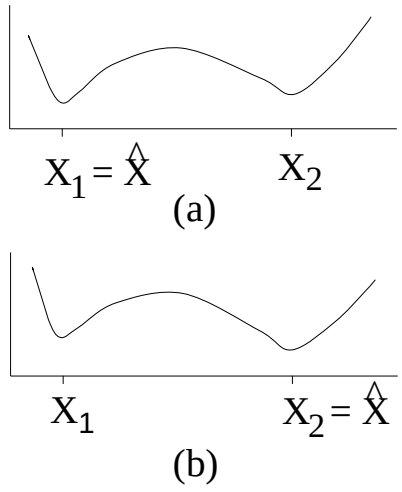


Figure 2: This figure illustrate how small changes in a nonconvex function can result in large changes in the location of the functions minimum value.

This function is shown in Fig. 1(b) for $T = 0.5$. For values greater than T the linear region of this function also allows sharp edges, yet convexity makes the MAP estimate efficient to compute.

In separate but related work, Green[12] employed the function

$$\rho(\Delta) = \beta \log \cosh(\Delta),$$

which produced useful Bayesian estimates of emission tomograms, while providing the aforementioned computational advantages. Proper choice of β and the scaling parameter λ give this potential function very similar behavior to that of the Huber function. Lange derived several other strictly convex potential functions in a study of convergence of the expectation maximization algorithm[13].

A disadvantage of these methods is that parameter choice for the potential functions requires knowledge of the edge magnitudes in the original images. Even if their values could be estimated accurately, it is not clear that a single value of, e.g. T can accurately describe edges in real images. In practice, all edges in an image do not have a single size.

4 A Stable Scale Invariant Approach

The conclusion of the previous sections is that it is desirable for the MAP estimate to be stable, and not depend on an absolute parameter of scale such as T. On first inspection, it may seem that these goals are incompatible with the requirement of preserving image edges. However, many nonlinear operations such as median filters[29] and stack filters[30] have been developed which preserve edges with out explicit prior knowledge of edge size. Since both of these nonlinear operations are homogeneous, any scaling of the input data results in a proportional scaling of the output image. Therefore, these edge preserving operations can not depend on any *a priori* knowledge of edge size.

In the following two subsections, we give conditions which guarantee both stability and

scale invariant behavior for the MAP estimate.

4.1 Stable MAP Estimation

Hadamard originally defined a problem to be well posed if its solution: (1) exists, (2) is unique, (3) depends continuously on the data. Condition (3), called stability, insures that small perturbations in the data do not cause dramatic change in the solution. In Section 3.2, we showed that a typical nonconvex function for $\rho(\cdot)$ did not result in a stable MAP estimator.

The problem of regularizing ill-posed problems has been the subject of much research[18, 19, 20, 21]. Tikhonov[31] has introduced methods for regularizing deterministic problems by introducing stabilizing functionals which play a role analogous to the log prior distribution of MAP estimation. In this work, Tikhonov also determined that these stabilizing functionals must meet certain conditions to guarantee that the resulting problem is well-posed. In particular, the stabilizing functionals are required to be "quasimonotone". A quasimonotone function is defined by Tikhonov to be one which contains no local minima other than the global minimum. Following the spirit of Tikhonov's work, we prove in Appendix A the following theorem[32]:

Theorem 1 *Let $f(\cdot,\cdot)$ be a continuous functional $f : U \times V \to I\!R$ such that for all $y \in V$ $f(\cdot, y)$ is strictly convex with a local minimum. Then*

$$\arg\min_{x \in U} f(x, y)$$

is a unique and continuous function of y.

This theorem may be directly applied to MAP estimation of finite dimensional images by choosing $f(x, y) = \log p(x, y)$. The MAP estimate is then guaranteed to be well-posed (and therefore stable) by choosing $\log p(x, y)$ to be a strictly convex function of x and a continuous function of (x, y).

Since in many physical problems $\log p(y|x)$ is accurately modeled as a convex function of x, a strictly convex log prior, $\log g(x)$, will insure strict convexity of $\log p(x,y)$ in x. Therefore, strict convexity of $\rho(\cdot)$ will generally insure stability of the MAP estimate when priors with the form of (3) are used. (In fact, simple convexity of $\rho(\cdot)$ may be sufficient to insure strict convexity of $\log p(x,y)$.)

4.2 Scale-Invariant MAP Estimation

The objective of this section is to develop a prior model which does not depend on an explicit edge size parameter such as T. To achieve this, we would like to find MAP estimators which are homogeneous operators. Specifically, scaling of the input y by any constant α should result in simple scaling of the resulting MAP estimator $\hat{x}$ by the same constant. The homogeneity property is insured if for all real constants α and for all inputs y

$$\arg\max_x \log p(\alpha y, \alpha x) = \arg\max_x \log p(y,x) \ .$$

Such an equality may be insured by requiring that

$$\log p(\alpha y, \alpha x) = \beta(\alpha, y) \log p(y,x) + \gamma(\alpha, y) \tag{4}$$

where β and γ are functions of α and y. A reasonable method for assuring that (4) holds is to require that the likelihood of Y given X and the prior distribution have the form

$$\log p(\alpha y|\alpha x) = \beta(\alpha) \log p(y|x) + \gamma_1(\alpha) \tag{5}$$

$$\log g(\alpha x) = \beta(\alpha) \log g(x) + \gamma_2(\alpha) \ . \tag{6}$$

These are the basic relations which we will use to enforce homogeneity in the MAP estimator. We will call functions such as p and g in (5) and (6) *scalable* due to the self-similar behavior of their respective logarithms.

The form of the function $\log p(y|x)$ is usually determined by the physics of a problem. However, the restriction that it be scalable is not unreasonable. To see this, consider the

random vector, Z, of independent and identically distributed random variables, Z_s, with the generalized Gaussian distribution[14, 33]

$$p_z(z) = \frac{q}{2\Gamma(1/q)} \exp(-|z|^q) \tag{7}$$

parameterized by q. When $q = 2$ the components of Z have a Gaussian distribution. When $q = 1$ they have a Laplacian distribution, and for $1 < q < 2$ the distribution has intermediate behavior. This noise model is commonly used in robust statistical estimation since it captures the heavy tailed behavior that is often exhibited by real noise distributions. If Y has the form

$$Y = \mathbf{A}X + \mathbf{D}^{-1}Z \tag{8}$$

where $\mathbf{A}$ and $\mathbf{D}$ are matrices, then $\log p(y|x)$ is a scalable function according to (5) and is given by

$$\log p(y|x) = -||\mathbf{D}(Y - \mathbf{A}X)||_q^q + \text{constant} \tag{9}$$

where $||\cdot||_q$ is the l_q norm.

If, in addition to $p(x|y)$, $g(x)$ is scalable with the same constants α and β, then the MAP estimator will be homogeneous. Also, we argued in Section 3.2 that the $\log g(x)$ should be a strictly convex function in order to insure stability of the solution to perturbations of the data. Enforcing these two requirements leads to the following theorem, proved in Appendix B.

Theorem 2 *The function $g : \mathbb{R}^N \to \mathbb{R}$ is a scalable density function with a convex log density function $\log g(x)$ if and only if*

$$-\log g(x) = ||x||^p + c$$

for some norm $||\cdot||$ and constants $p \geq 1$, and c.

4.3 Generalized Gaussian MRF

Theorem 2 leaves available a wide variety of possible choices for $g(x)$. However, we propose a simple generalization of Gaussian MRF's based on the concept of generalized Gaussian noise. This model has the functional form similar to (3), but uses $\rho(\Delta) = |\Delta|^p$,

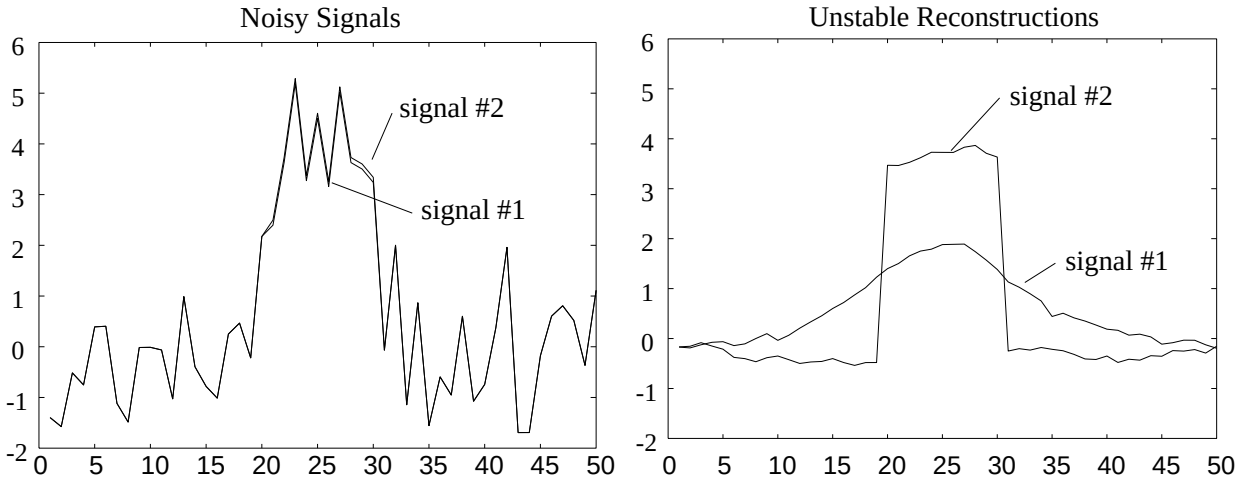
$$\log g(x) = -\lambda^p \left(\sum_s a_s |x_s|^p + \sum_{\{s,r\} \in C} b_{s,r} |x_s - x_r|^p \right) + \text{constant}, \tag{10}$$

where $1 \leq p \leq 2$, and λ is a parameter which is inversely proportional to the scale of x. We call the class of random fields with this distribution generalized Gaussian Markov random fields (GGMRF) since this model is contained within the more general class of MRF's and includes all Gaussian MRF's when $p = 2$. We also note that Besag[34] suggested a model similar to the GGMRF model with $p = 1$ which he referred to as the "median pixel prior" [1]. More recently, DeVore and Lucier have argued that the optimization problem resulting from the use of the GGMRF prior can be motivated based on approximation smoothness[45].

As in the case of the GMRF, not all values of the parameters a_s and $b_{s,r}$ will lead to a consistent model. In fact, $g(x)$ will be well defined only if $-\log g(x)$ is a positive definite function of x. A sufficient condition for positive definiteness is that $a_s > 0$ and $b_{s,r} > 0$. This condition also insures that $-\log g(x)$ is convex. In practice, we may choose $a_s = 0$, which results in an ill defined density. However, this is not a practical difficulty since the function $\log p(y|x)$ causes the MAP estimate to be unique.

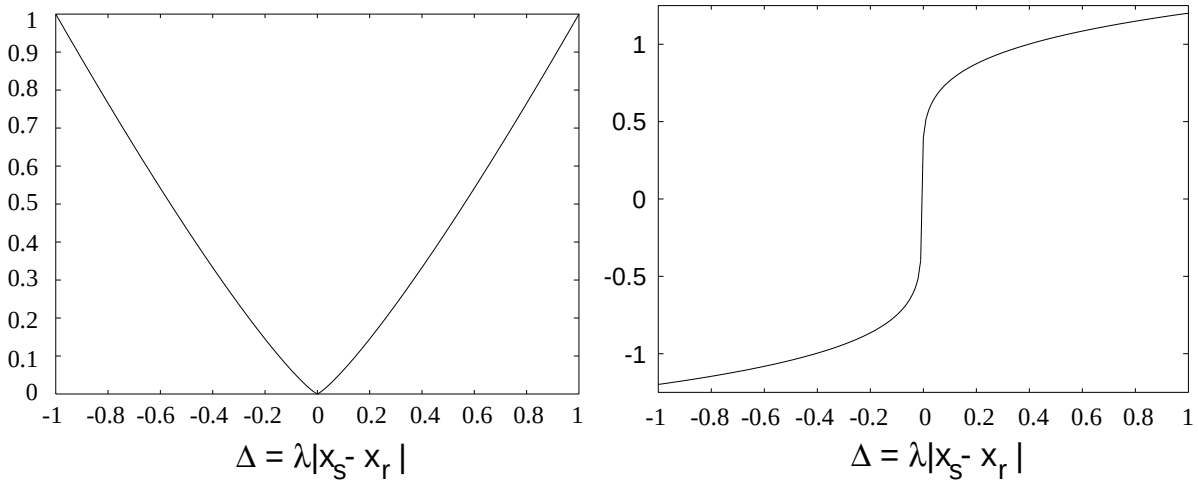
The choice of p is critical in determining the character of the model. Larger values of p discourage abrupt discontinuities while smaller values of p allow them. Fig. 4a shows the function $\rho(\Delta) = |\Delta|^{1.2}$. The derivative of $\rho(\cdot)$ is also shown in Fig. 4b. This function determines the tendency of neighboring pixels to be attracted and plays a role analogous to the influence function of robust statistics [11, 35].

[1] This approach modeled the unknown field X as the additive combination of the median pixel prior with an independent Gaussian MRF. The relative smoothness of the model was then controlled by the relative mixture of the Gaussian MRF and median pixel prior.



(a) (b)

Figure 3: Unstable reconstruction of two noisy pulses. a) Noisy square pulses with magnitudes 4.2(#1) and 4.3 (#2) in the interval [20,29], and additive white Gaussian noise of unit variance. b) Resulting MAP estimates using Blake and Zisserman function with $T = 1.75$, $\lambda = 5$, and $b_{sr} = 1$ for adjacent points. Optimization was performed using 10^5 iterations of simulated annealing.



(a) $\rho(\Delta) = |\Delta|^{1.2}$ (b) $\frac{d\rho(\Delta)}{d\Delta}$ = influence function

Figure 4: (a) An example of the proposed scale invariant convex cost function when $p = 1.2$. ρ is a function of Δ the difference between neighboring pixel values. (b) The derivative of ρ represents the attraction between two points separated by Δ.

The one-dimensional case with $p = 1$ provides insight into edge reconstruction. The prior distribution then has the form

$$\log g(x) = -\lambda \sum_{s=1}^{N-1} |x_s - x_{s+1}| + \text{constant} \ . \tag{11}$$

As long as x is a monotone (increasing or decreasing) function, then

$$\sum_{s=1}^{N-1} \lambda |x_s - x_{s+1}| = \lambda |x_1 - x_N| \ .$$

Therefore, the total cost is simply the difference between the starting and ending values. This means that abrupt edges in the reconstruction have no greater cost than smooth edges. Fig. 5 illustrates this fact and indicates that nonconvex functions are not required for the reconstruction of sharp edges.

We can experimentally illustrate the properties of the GGMRF prior with $p = 1$ by computing the MAP estimate of a noisy pulse. Fig. 6a shows 6 noisy pulses of widely varying amplitudes. Each signal is formed using the same noise and pulse shape as in the previous example of Fig. 3a. Fig. 6b shows the corresponding MAP estimates using the GGMRF prior of (11) with $\lambda = 5$. The solutions exhibit the continuous dependence on data guaranteed by Theorem 1. Note that although the prior term is not strictly convex when $p = 1$, the sum of the prior with $\log p(y|x)$ for the Gaussian noise is a strictly convex function of x.

Theorem 1 can also be used to show that the MAP solution under the GGMRF is a continuous function of p. To see this consider the joint density of (x, y) given the parameter p. Since the logarithm of this function is strictly convex in x, and a continuous function of (x, y, p) for $p \geq 1$, the MAP estimate must be a continuous function of the parameter p. This property is illustrated in Fig. 7 which shows the MAP estimate for various p and fixed $\lambda = 5$. The input signal is the same noisy pulse used in previous examples with a fixed amplitude of 4.0. The parameter p ranges from 1.01 to 2.0, with reconstructions varying in character from the smoothing of the Gaussian prior, to the relatively abrupt-edge estimate for $p = 1.01$.

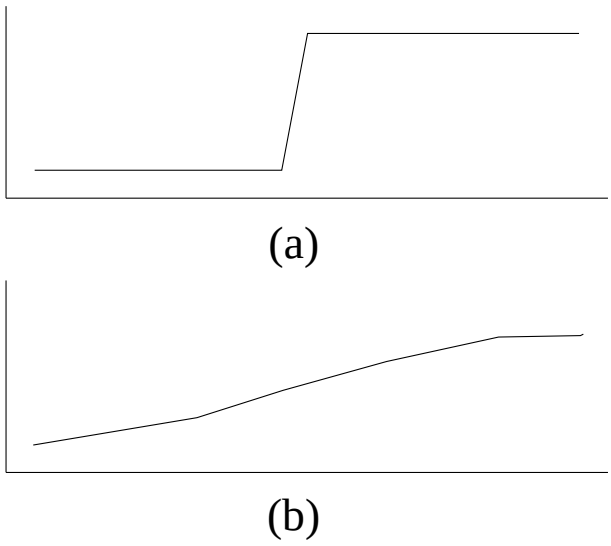


Figure 5: When $p = 1$, any monotone functions which are equal at the begining and end must have the same total cost. Therefore, both the sharp edge and the smooth edge have the same cost.

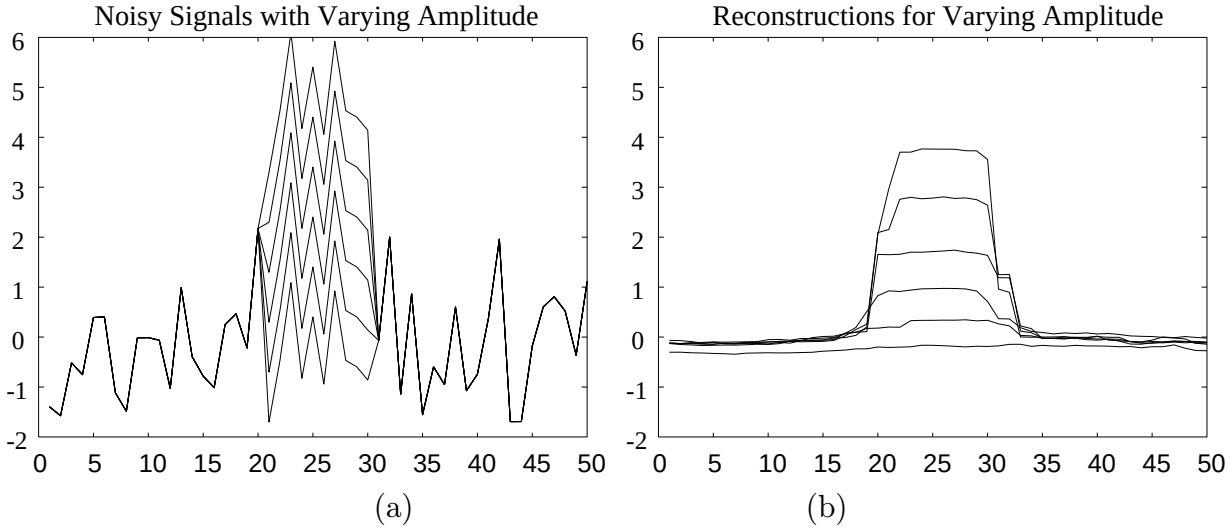


Figure 6: MAP estimates using GGMRF prior with $p = 1$, $\lambda = 5$. (a) Noisy square pulses with amplitudes of 0,1,2,3,4, and 5. Each contains additive white Gaussian noise of unit variance. (b) Resulting MAP estimates using GGMRF prior with $p = 1$, $\lambda = 5$, and $b_{sr} = 1$ for adjacent points. Optimization was performed using 10^4 full iterations of simulated annealing.

Let us assume that the observed distortion has the generalized Gaussian noise form of (9) and the prior distribution is from a GGMRF. Then both $g(x)$, and $p(y|x)$ will be scalable. If in addition, $p = q$ then the α and β parameters for both distributions will be the same, and the MAP estimator will be a homogeneous operation. More generally, if we write the MAP estimate, $\hat{x}$, explicitly as a function of the input data, y, and the prior scale parameter, λ, it is easily shown that

$$\hat{x}(\alpha y, \lambda) = \alpha \hat{x}(y, \alpha^{1-q/p}\lambda) \ . \tag{12}$$

When $p = q$, the relation $\hat{x}(\alpha y, \lambda) = \alpha \hat{x}(y, \lambda)$ holds for all α, and the MAP estimator is homogeneous. When $p \neq q$, the MAP estimator is not homogeneous, since the distributions for the prior and observation noise no longer coincide. However, (12) indicates the qualitative behavior of the MAP estimate would not be expected to change as the input is scaled since the result is proportional to a MAP estimate using a different regularization constant, $\alpha^{1-q/p}\lambda$.

5 Optimization Techniques

In this section, we discuss the minimization techniques which we will use to compute the MAP estimator. These methods are of fundamental importance for two reasons. First, they provide basic intuition for understanding MAP estimation using the GGMRF prior. Second, the minimization techniques connect the area of MAP estimation to the literature in weighted median filtering[29, 36, 37]. Since median filtering has been shown to be of broad practical importance in image filtering, we believe this suggests that methods based on the GGMRF prior can also be practically useful in a variety of image estimation applications.

We shall adopt a simplified problem for illustrating the issues of minimization. Assume that Y is formed by adding white noise to X,

$$Y = X + \sigma Z \tag{13}$$

where Z is defined in (7) and σ is a scale parameter (not equal to the standard deviation). We will also assume that the prior model is a homogeneous MRF (i.e. $b_{s-r} = b_{r-s}$ is used in

place of $b_{s,r}$), and the coefficients $a_s = 0$. The MAP estimate is then given by

$$\hat{x} = \arg\min_x \left\{ \sum_{s \in S} |y_s - x_s|^q + \sigma^q \lambda^p \sum_{\{s,r\} \in C} b_{s-r} |x_s - x_r|^p \right\} . \tag{14}$$

Since this cost function is convex for $p, q \geq 1$, finding a global minimum will be computationally feasible.

In general, (14) may be minimized using either global or local iterative techniques. Two examples of global iterative techniques are gradient descent and conjugate gradient[38]. Local minimization methods iteratively minimize the cost function at each pixel, x_s, of x. Since X is a MRF, minimization of the cost function with respect to x_s results in the following simple local computation.

$$\hat{x}_s = \arg\min_{x_s} \left\{ |y_s - x_s|^q + \sigma^q \lambda^p \sum_{r \in \partial s} b_{r-s} |x_s - x_r|^p \right\} \tag{15}$$

This is equivalent to the local operation used in the method ICM proposed by Besag[3]. In fact, it is also closely related to a numerical algorithm called Gauss-Seidel (GS) which has long been used for solving partial differential equations[47]. The GS interpretation of (15) will be important to us because it has been shown[42, 43] that GS iterations have fast numerical convergence for the transmission tomography problem described in the following section.

The discussion of minimization methods will be broken down into distinct cases depending on the values of p and q. When $p = q = 2$ the well-known Gaussian case occurs. Here the reconstruction may be thought of as the best linear estimate with the resulting edge blurring and nonrobustness to noise. The local minimization operation reduces to a linear average of the observed value y_s and the neighbors of x_s.

$$\hat{x}_s = \frac{y_s + (\sigma\lambda)^2 \sum_{r \in \partial s} b_{r-s} x_r}{1 + (\sigma\lambda)^2 \sum_{r \in \partial s} b_{r-s}}$$

5.1 Optimization for $p = q = 1$

When $p = q = 1$ the cost function is not strictly convex so Theorem 1 does not apply. However, this still represents an important limiting case as the distributions become heavy tailed. For $p = q = 1$ the cost function is a convex polytope in a high-dimensional space. Along the edges of the polytope, the function is not differentiable.

To illustrate the local updating equation of (15) we will first consider some special cases. If $\sigma^q \lambda^p b_r = 1$, then the local minimization operation reduces to the median of the observed pixel value, y_s, and the pixel's neighbors,

$$\hat{x}_s = \text{median}\left\{y_s, x_{r_1}, x_{r_2}, \cdots, x_{r_I}\right\}$$

where $x_{r_1}, \cdots, x_{r_I}$ are the I neighbors of the pixel x_s. This replacement operation is similar to the recursive median filter except it uses the original data value in place of the previous value for x_s in the median operation. This keeps the MAP estimate from drifting too far from the original data.

In the most general case of arbitrary coefficients, the solution of (15) with $p = q = 1$ is known as the weighted median. The weighted median is the value, $\hat{x}$, such that the total weight of pixels greater than $\hat{x}$ is as close as possible to the total weight of pixels less than $\hat{x}$. Since the weighted median has the flexibility to treat pixels differently as a function of position, it has attracted attention as a nonlinear filter for image processing[39, 40].

Median filters are known to be robust homogeneous filtering operations which preserve edges in practical image processing applications. So it is encouraging that they are similar to the local minimization operations of our MAP estimation problem. Surprisingly however, MAP estimation and median filtering are actually quite distinct because the local operations generally *do not* converge to the global MAP estimate. This happens because the local operations become "stuck" on the edges of the nondifferentiable polytope. In fact, it is well known that the recursive median filter converges to a root signal (which is generally not

constant)[36, 37].

Efficient numerical minimization for this case is a difficult problem which may depend on the specific application. However, some general observations may be made. The global MAP estimate may be approximately computed for $p = 1$ by alternating a complete pass of local minimization with a single iteration of a gradient-based method. Since the cost function is not differentiable, the gradient must be replaced by its generalization, the *subgradient*[48], which we choose according to

$$\frac{d|x|}{dx} = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases} .$$

The efficiency of minimization can be further improved by updating larger groups of pixels of constant value[41]. This approach, which we refer to as segmentation based optimization, corresponds to moving along the discontinuous edges of the log likelihood function. Finally, multigrid[44] and wavelet[45] techniques show promise for reliable and fast solution of this MAP optimization problem.

5.2 Optimization for $1 < p, q < 2$

When $1 < p = q < 2$, the cost function is differentiable, and it may be easily shown that the iterative local minimization of (15) converges to the global MAP estimate. In this case, the global minimum is the only point for which the gradient is zero. This local operation is, of course, nonlinear. When $p = q$, it is also homogeneous (as is the entire MAP estimator). The operation of (15) has the form of a least powers M-estimator used in robust statistics[35, 46]. In practice, a value of $q = 1.2$ has been found to yield a good compromise between asymptotic efficiency and robustness for M-estimation in real data[35].

For low values of p and/or q, convergence is very slow, due to the nondifferentiability of the GGMRF term in the limit as they approach 1. When $p \approx 1$ or $q \approx 1$, the numerical methods outlined in Section 5.1 may be used to accelerate convergence.

Due to the physical nature of a problem, we may often have $1 \leq p \neq q \leq 2$. In this case, the local operation for minimization is not homogeneous, but it does maintain a similar property, as described in Section 4.3.

6 Statistical Tomographic Reconstruction

In this section, we briefly describe the specific problem of statistical reconstruction of 2D cross-sections from integral projections. This inversion problem has been approached within the Bayesian framework for both emission[16, 17, 7, 12] and transmission[42, 43] tomography.

The 2-D Radon transform maps a function of two variables, which we denote by $x(s_1, s_2)$, into a function indexed by (θ, t) according to

$$p(\theta, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(s_1, s_2) \delta(t - s_1 \cos\theta - s_2 \sin\theta) ds_1 ds_2 \tag{16}$$

where $\delta()$ is an impulse function. Fig. 8 illustrates the collection of projection data for a single value of θ. The value of $p(\theta, t)$ represents the integral of $x(s_1, s_2)$ along the ray at orientation $\theta + \frac{\pi}{2}$, at a displacement t from the center of the field.

In practice, reconstruction requires finite-dimensional representation of both the projection data, p, and the modeled image, x. The projections may be discretized by computing them for only a finite set of M projection rays, $\{(\theta_i, t_i)\}_{i=0}^{M}$. The i^{th} projection is then written as $p_i = p(\theta_i, t_i)$. The Radon transform equations may now be written in the discrete form

$$p = \mathbf{A}x$$

where $\mathbf{A}$ is a sparse matrix whose $(i, j)th$ entry indicates the contribution of modeled pixel j to the i^{th} projection measurement.

In transmission tomography the projections, p, are not measured directly. Instead, raw data are in the form of the number of photons, y_i, detected after passing through an absorptive material. In [43], the following quadratic approximation is derived for the log likelihood

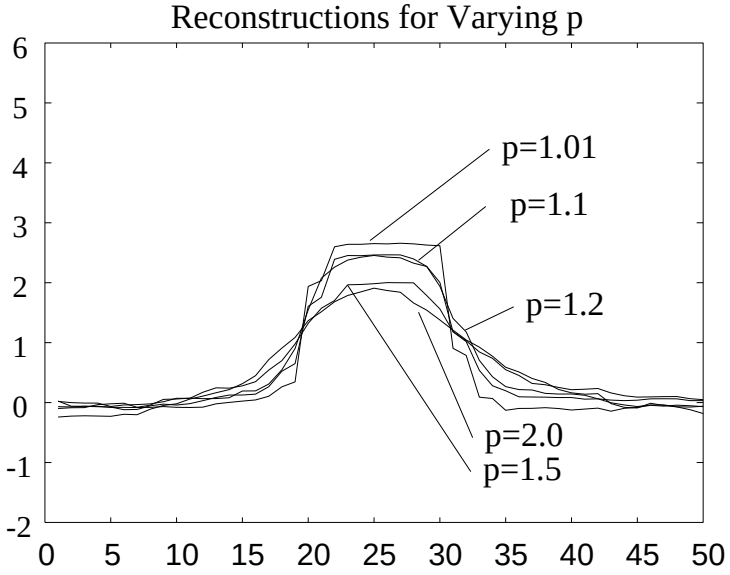


Figure 7: MAP estimates of square pulse of amplitude 4.0, using GGMRF prior with $\lambda = 5$, and varying p. Optimization was performed using 10^4 iterations of simulated annealing.

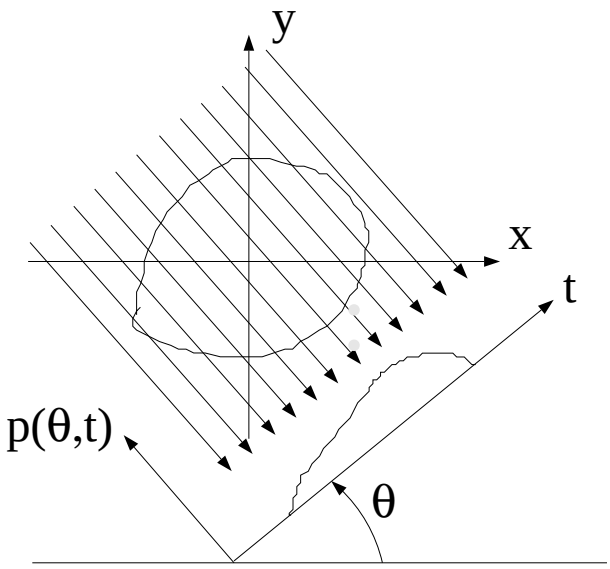


Figure 8: Projection data for angle θ, resulting in the one-dimensional function $p(\theta, t)$.

of the photon counts y given the image x:

$$\log p(y|x) \approx -\frac{1}{2}(\hat{p} - \mathbf{A}x)^t \mathbf{D}^2 (\hat{p} - \mathbf{A}x) + c(y), \tag{17}$$

where $\hat{p}_i$ and $\mathbf{D}$ are defined by

$$\begin{aligned} \hat{p}_i &= \log(y_T/y_i) \\ \mathbf{D} &= \text{diag}\{\sqrt{y_1}, \sqrt{y_2}, \cdots, \sqrt{y_M}\} \end{aligned}$$

for input photon count y_T.

While repeating the derivation of (17) is beyond the scope of this paper, general attributes of the approximation may be inferred from its structure. The matrix $\mathbf{D}$ more heavily weights errors corresponding to projections with large values of y_i. These projections pass through less dense objects, and consequently have higher signal-to-noise ratio. In the limit of opaque projections where no photons pass through the material, the approximation simply applies no weight to the measurement. The expression of (17) is quite accurate for reasonable ranges of photon counts, and offers analytical advantages over the more precise Poisson distribution.

In order to apply the MAP estimation techniques described above, we will require computationally efficient methods for implementing the minimization methods described in Section 5. In fact, these methods have already been developed for the tomography problem in [42, 43, 41, 44]. A previous paper[43] describes how the GS updates of (15) can be applied using the log likelihood function of (17). This work shows that the GS and gradient descent minimization methods require approximately equal amounts of computation per iteration through the data. However, when a Gaussian prior is used, the GS method is analytically shown to suppress high frequency error components more rapidly, while the gradient descent method suppresses low frequencies more rapidly.

7 Experimental Results

Under the approximation of the conditional log likelihood of the photon counts given in (17), we are restricted to $q = 2$ for the present experimental work, and will show the character of the results' dependence on the choice of p in the GGMRF. The results presented here were achieved primarily using GS iterations[43], with pixel-by-pixel updates. For $p \approx 1$, the multigrid[44] and segmentation techniques[41] mentioned above substantially improved convergence. Convergence rates are discussed in detail in these previous papers. As mentioned in Section 5, the GS iterations will in general not find the global minimum for $p = 1$, and will be slow in converging for other small values of p. We cannot in general verify the precision of our results as MAP estimates, especially for small p. However, we have made every effort, through large numbers of iterations where necessary, to assure that our results are very near the true MAP reconstructions.

The synthetic test phantom, shown in Fig. 9(a), consists of two distinct densities, 0.22cm^{-1} and 0.48cm^{-1}, both of which are within the range of human tissue in X-ray absorptivity. Increasing intensity in Fig. 9 represents higher absorptivity. The physical diameter is approximately 20cm. Projections are collected using only $y_T = 2000$ photons per ray, far below typical clinical dosages, making the lighter regions nearly opaque to the X-rays. With these values for y_T and object composition, photon counting noise may dominate the corruption of the reconstruction if conventional techniques such as convolution backprojection (CBP) are used. The best (by visual inspection) CBP reconstruction resulted from relatively severe lowpass filtering of projection data before inversion, and can be seen in Fig. 9(b).

This case is similar to the hollow projections problem, but note that our MAP reconstruction methods require no estimation of the dense regions' locations, or interpolation of projections. The algorithm can be applied directly to other limited data problems such as the limited-angle reconstruction.

The GGMRFs used featured 8-pixel neighborhoods, with unity weighting of nearest horizontal and vertical neighbors, and 0.7 for diagonal neighbors. We present MAP estimates for $\lambda = 5, 10$, and $p = 2.0, 1.2, 1.05, 1$ in Figs. 10 and 11. For the Gaussian prior, $\lambda = 10$ is equivalent to a standard deviation for each pixel, given its neighbors, of 0.027cm^{-1}. The reconstruction using the Gaussian prior ($p = 2$) suffers from the smoothing of edges as a cost of noise suppression. Smaller values of λ can sharpen object boundaries but at the expense of larger noise artifacts.

Figs. 10(d) and 11(d) show the results when a GGMRF prior is used with $p = 1$. Due to the property illustrated previously in Fig. 5, edges are allowed to form freely in this image, and boundary definition in these is appreciably higher than the other examples. The influence function of the absolute value prior does not fall to zero in the neighborhood of the origin, which helps suppress the noise in these estimates much more effectively than the Gaussian model. The other values of p yield intermediate behavior, as predicted by continuity property in p mentioned earlier. The value of $p = 1.2$, as suggested by Rey[35], produces an improved reconstruction, with limited edge smoothing, but very good noise artifact suppression. Reconstructions with $p = 1.05$ have nearly the same effect on noise as $p = 1.0$, but produce slightly softer edges.

The GGMRF MAP estimate with small values of p has substantially lower mean-squared error than the CBP image, or the MAP estimate with the Gaussian prior. But because the mean-squared error tends to be dominated by pixels at the edges of the high intensity regions, we have found it to be a misleading measure of performance. Alternatively, Fig. 12 shows a histogram of the absolute error in the reconstructed images for $p = 1$ and $p = 2$, and $\lambda = 10$. The pulses at the far right edge of each plot are the sum of all absolute errors above 0.05. Note the much greater concentration of errors at the lower magnitudes for the case $p = 1$. For phantoms with substantial constant-valued regions such as our example, error values may tend to cluster in magnitude with $p = 1$. The two large values in the $p = 1$

histogram on either side of the value 0.02 represent the errors from the two larger dense regions of the phantom, which are reproduced as nearly constant. As illustrated in Fig. 5, it is only the total local rise in a function which is penalized by this prior. The dense disks are therefore attenuated in amplitude inversely to their border lengths.

Many of the realistic settings for the tomographic problem at hand arise in nondestructive evaluation. One such example is shown in Figs. 13-14. The subject is a block of concrete, 20 cm on each side, with four steel reinforcing rods. The data for this case was collected by Lawrence Livermore National Laboratories, using a first-generation gamma-ray scanner, with an ^{192}Iridium source. Such a scanner effectively removes many scattering artifacts, but requires relatively long exposure to accumulate sufficient photon counts for reasonable signal-to-noise ratio. The CBP reconstruction of Fig. 13 is therefore degraded by noise artifacts somewhat similar to those in the previous example.

The photon dosage, y_T, is not available for this data. Since the weighting matrix $\mathbf{D}$ of (17) is proportional to y_T, this represents an unknown scaling factor for the data component of the log posterior density function. Therefore, we cannot make a useful direct comparison across p of equal values of λ, as with the synthetic phantom. But the form of the optimization for a given p is defined entirely by the ratio y_T/λ^p, and we present these results using this ratio and p as the parameters. Reconstructions for $p = 2.0$ and $p = 1.0$ appear in Fig. 14, with the chosen $p = 2.0$ estimate being qualitatively the best across a wide range of y_T/λ^2. As in the previous example, Fig. 14(a) suffers from the trade-off between noise suppression and smoothing. Figs. 14(b)-(d) illustrate the GGMRF estimates under the absolute value prior with y_T/λ ranging from 4×10^4 to 4×10^2. Because we have no original image of this cross-section, we show MAP estimates over two orders of magnitude of y_T/λ. These estimates range from the apparently under-weighted prior of figure 14(b), to the excessively smoothed results in Fig. 14(d). We propose that Fig. 14c represents a useful approximation, with good retention of structural detail. Other examples of similar data are currently under

study.

8 Conclusion

The GGMRF has demonstrated analytical properties and experimental results which offer promise for applications in many problems of image estimation. In particular, the GGMRF prior leads to a MAP estimator which may be uniquely computed. Moreover, when $1 < p \leq 2$, this MAP estimator is guaranteed to be a continuous function of the input data. For any problem in which the noise parameter q equals the prior parameter p, the MAP estimator will be invariant to scaling of the data. This means that edge magnitudes need not be predetermined. When $p \neq q$, variations in the data scale are equivalent to variations in the signal-to-noise ratio, $\sigma^q \lambda^p$, used in the MAP estimate.

The computed tomography simulations presented here have dealt with materials expected to have sharp transitions between densities. The suitability of the GGMRF with small values of p to more smoothly varying images is, as yet, unclear. However, it is promising that median filters, which have been successfully applied in image processing, are closely related to MAP estimation with the GGMRF prior. As noted earlier, the Bayesian approach has the advantage of retaining the original data in its recursions.

The very slow convergence of the MAP estimate with small p is an impediment to the efficient application of these techniques. A major effort of our coming research will be directed toward speeding the MAP estimation process.

Acknowledgments

The authors thank Dr. Steve Azevedo and Lawrence Livermore National Laboratories for supplying data for this research, and Professors R. Stevenson and E. Delp for their helpful comments.

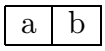


Figure 9: (a) Original phantom (left); (b) convolution backprojection reconstruction in low photon dosage with 128 projections at each of 128 angles (right); all synthetic phantom images are presented at a resolution of 128×128 pixels.

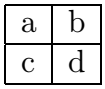


Figure 10: Estimates with $\lambda = 5.0$. (a) MAP estimate using Gaussian prior, $p = q = 2$ (upper left); (b)-(d) MAP estimates using Generalized Gaussian MRF, $p = 1.2$, 1.05, and 1.0, respectively.

a	b
c	d

Figure 11: Estimates with $\lambda = 10.0$. (a) MAP estimate using Gaussian prior, $p = q = 2$ (upper left); (b)-(d) MAP estimates using Generalized Gaussian MRF, $p = 1.2$, 1.05, and 1.0, respectively.

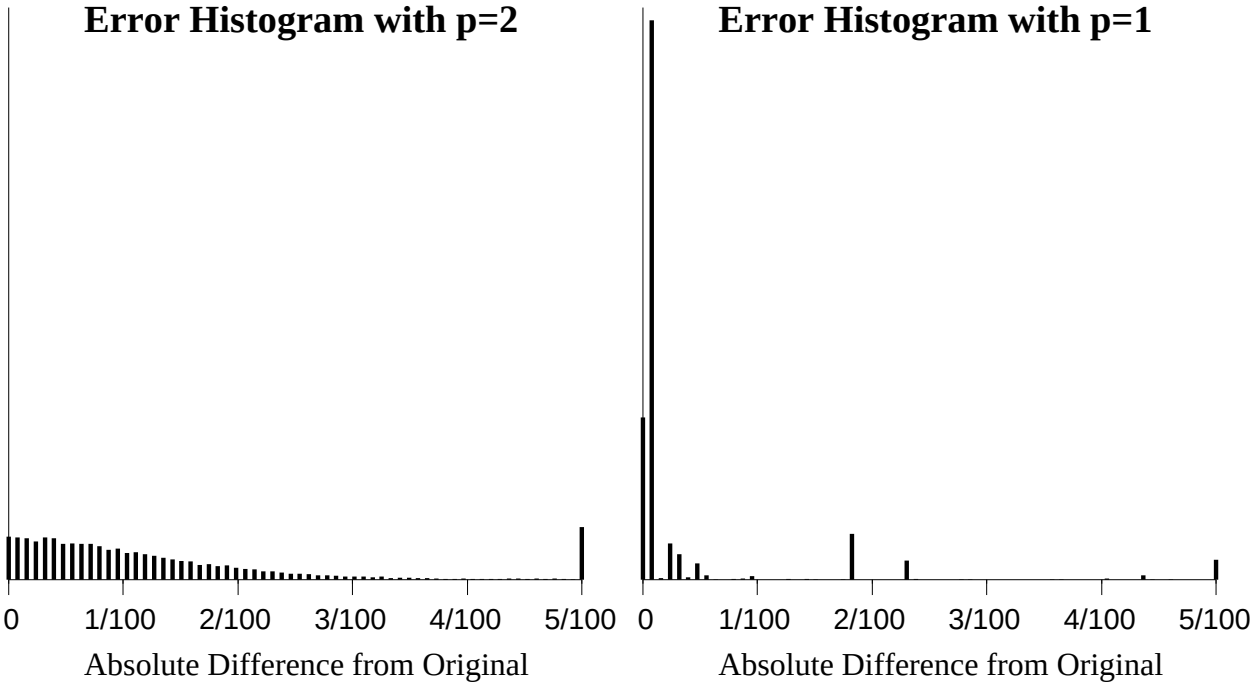


Figure 12: Histograms of absolute error in the reconstructed images with $\lambda = 10$ for the cases $p = 2.0$ and $p = 1.0$. Value at right of each is the total of occurrences for all differences greater than or equal to 0.05

Appendix A

Our objective in this appendix is to prove Theorem 1 by proving two lemmas which directly imply it. We first modify Tikhonov's original definitions slightly and define the following.

Definition 1 *A local minimum of a function is any point which is the minimum on some local open neighborhood of the point.*

Definition 2 *The functional $h : U \to I\!R$ is called quasimonotonic if $h(\cdot)$ has a unique global minimum, $h(x_o)$, at the point $x_o \in U$, $h(\cdot)$ contains no other local minima, and there exists a number $b > h(x_o)$ such that $\{x \in U : h(x) \leq b\}$ is compact.*

Lemma 1 *Let $f(\cdot, \cdot)$ be a continuous functional $f : U \times V \to I\!R$ such that for all $y \in V$ $f(\cdot, y)$ is quasimonotonic then*

$$\arg\min_{x \in U} f(x, y)$$

is a continuous function of y.

Proof of lemma 1:

We will prove this lemma for $f(\cdot, \cdot)$ defined on any general metric space, $U \times V$. The appropriate induced metrics on U and V will be denoted by $d_u(x, \tilde{x})$ and $d_v(y, \tilde{y})$ respectively. This is equivalent to $||x - \tilde{x}||$ and $||y - \tilde{y}||$ when U and V are vector spaces.

Choose any $y \in V$. By assumption, there is a unique global minimum.

$$\begin{aligned} \hat{x} &= \arg\min_{x \in U} f(x, y) \\ c &= f(\hat{x}, y) \end{aligned}$$

Our objective is then to show that for any $\epsilon > 0$, there is a $\delta > 0$, so that for all $\tilde{y}$ with $d_v(y, \tilde{y}) < \delta$

$$\arg\min_{x \in U} f(x, \tilde{y}) \in \mathcal{E}$$

where

$$\mathcal{E} = \{x : d_u(x, \hat{x}) < \epsilon\} \ .$$

By assumption, there exists a $b > 0$ such that

$$A_1 = \{x \in U : f(x, y) \leq b + c\}$$

is a compact set. If we define the sequence of sets

$$A_n = \{x \in U : f(x, y) \leq b/n + c\}$$

then each A_n contains $\hat{x}$ and must be compact since it is a closed subset of A_1. If $\bar{\mathcal{E}}$ denotes the closed set given by the complement of $\mathcal{E}$, then

$$O_n = A_n \cap \bar{\mathcal{E}}$$

is a sequence of compact sets none of which contain $\hat{x}$.

We will next show that for some N, $A_N \subset \mathcal{E}$ or equivalently O_N is empty. To show this, assume that O_n is not empty for any n. Since O_1 is compact, there is a sequence of points, $x_n \in O_n \subset O_1$, with a subsequence, x_{n_k} that converges in O_1. Since $f(\cdot, y)$ is continuous and $x_n \in A_n$, the limit of this subsequence must also be a global minimum of $f(\cdot, y)$. This contradicts the assumption that there is a unique global minimum to the function $f(\cdot, y)$.

Define the following three subsets of A_N.

$$\begin{aligned} \tilde{A}_N &= \{x \in U : f(x, y) < b/N + c\} \\ I &= \text{the interior points of } A_N \\ B &= \text{the boundary points of } A_N \end{aligned}$$

Then by the continuity of $f(\cdot, y)$, it may be shown that $\tilde{A}_N \subset I$. Therefore,

$$A_N - \tilde{A}_N \supset A_N - I = B \ .$$

This implies that for all $x \in B$, $f(x, y) = b/N + c$.

Since $f(\cdot,\cdot)$ is a continuous function, it is uniformly continuous on any compact set. Therefore, there exists a $\delta > 0$ such that for all $\tilde{y}$, with $d_v(y,\tilde{y}) < \delta$

$$\sup_{x \in A_N} |f(x,y) - f(x,\tilde{y})| < \frac{b}{4N} \ . \tag{18}$$

We will use this fact to show that for any choice of $\tilde{y}$, with $d_v(y,\tilde{y}) < \delta$ the global minimum of $f(\cdot,\tilde{y})$ is still a member of $A_N \subset \mathcal{E}$. Since A_N is compact and $f(\cdot,\tilde{y})$ is continuous, $f(\cdot,\tilde{y})$ must take on its minimum value at some point, $\tilde{x} \in A_N$. If we can show that the point $\tilde{x}$ is in the interior of A_N, then it must be the global minimum since it is a local minimum and $f(\cdot,\tilde{y})$ is assumed to have no local minima other than the global minimum.

Using the uniform continuity property of (18), we know that for all boundary points $x \in B$

$$f(x,\tilde{y}) \geq \frac{3b}{4N} + c \ , \tag{19}$$

and at the point $\hat{x}$

$$f(\hat{x},\tilde{y}) \leq \frac{b}{4N} + c \tag{20}$$

Therefore, $f(\cdot,\tilde{y})$ is less at the point $\hat{x}$ than at any point on the boundary of A_N. Since there is at least one point in the interior of A_N which is less than any point on the boundary, the minimum of $f(\cdot,\tilde{y})$ must fall on the interior of A_N.

Lemma 2 *Any strictly convex function f:* $\mathbb{R}^N \rightarrow \mathbb{R}$ *with a local minimum is quasimonotonic.*

Proof of lemma 2:

Strict convexity implies that at most one local minimum of f exists. Without loss of generality, assume the minimum occurs at $x = 0$, and the minimum value is $f(0) = 0$.

In order to show that $C = \{x \in \mathbb{R}^N : f(x) \leq b\}$ is compact for any $b > 0$, we invoke the Heine-Borel Theorem, which states that in $\mathbb{R}^N$, every closed and bounded set is compact.

Convexity implies continuity of f. Since the mapping $f : I\!R^N \rightarrow I\!R$ is continuous, $f^{-1}(S) \in I\!R^N$ is closed for every closed set S in $I\!R$. Therefore, $C = f^{-1}((-\infty, b])$ is a closed set.

By definition of the unique local minimum, there is an N-ball, $B = \{x : ||x|| \leq 1\}$, about 0 such that for $x \neq 0$ with $x \in B$, $f(x) > 0$. Given that f is continuous, the latter inequality holds on the surface of the ball, $D = \{x : ||x|| = 1\}$, a compact set. The continuous function f must attain a minimum in D, and we denote any point at which the minimum occurs as x_m. If we choose $b = f(x_m) > 0$, then $C \subset B$ and C is compact. To see this assume that $x_o \in C$, but $x_o \notin B$. Then defining $\lambda = \frac{1}{||x_o||}$, we have

$$\begin{aligned} \lambda f(x_o) + (1-\lambda) f(0) &> f(\lambda x_o) \\ &= b \end{aligned}$$

and this implies the contradiction $f(x_o) > b$.

Appendix B

Definition 3 *A strictly positive function $g(x)$ is called scalable if for each constant α, there exist two constants β and γ so that for almost every x*

$$\log g(\alpha x) = \beta \log g(x) + \gamma \ .$$

Proof of theorem 2:
($\Leftarrow$) We must prove that **(a)** $\exp\{-||x||^p\}$ defines a proper density function, **(b)** $||x||^p$ is convex, and **(c)** scalable.

a) Any norm has the property that in a finite dimensional space

$$\int_{I\!R^N} \exp\{-||x||^p\} < \infty \ .$$

Therefore, this forms a proper probability distribution.

b) We may use the triangle inequality and the convexity of $|\cdot|^p$ to show that $||x||^p$ is convex. For all $0 < \lambda < 1$,

$$\begin{aligned} ||\lambda x + (1-\lambda)y||^p &\leq (\lambda||x|| + (1-\lambda)||y||)^p \quad (21) \\ &\leq \lambda||x||^p + (1-\lambda)||y||^p \quad (22) \end{aligned}$$

c) X is scalable since

$$\begin{aligned} \log g(\alpha x) &= -||\alpha x||^p + c \\ &= -|\alpha|^p\{\log g(x)\} - (1-|\alpha|^p)c \end{aligned}$$

($\Rightarrow$) We must determine that $-\log g(x) = (f(x))^p + \text{constant}$ where f has the properties **(a)** for all $c \geq 0$, $f(cx) = cf(x)$, **(b)** for all x, $f(x) = 0$ implies $x = \theta$ where θ is the zero vector, and **(c)** f(x) obeys the triangle inequality.

a) Define the function $\tilde{u}(x) = -\log g(x)$. Since $\tilde{u}(x)$ is convex, it is a continuous function of x, and $\tilde{u}(\theta)$ exits where θ is the vector of zeros. By assumption we have that for any α there are β and γ so that

$$\tilde{u}(\alpha x) = \beta\tilde{u}(x) - \gamma \ .$$

If we define the new function, $u(x) = \tilde{u}(x) - \tilde{u}(\theta)$ then for all x

$$u(\alpha x) = \beta u(x)$$

Choose any $\alpha > 1$. Since $g(x)$ must integrate to 1, there must be an x such that $u(x) = u_o > 0$. Consider the three points $u(0x) = 0$, $u(x) = u_o$ and $u(\alpha x) = \beta u_o$. Convexity implies that $\beta \geq \alpha$. Therefore, we can find a $p \geq 1$ so that $\beta = \alpha^p$.

Define $f(x) = (u(x))^{1/p}$. Then for all integers $n \geq 1$, $f(\alpha^n x) = \alpha^n f(x)$. Choose $\delta = a^{1/m}$, then similarly

$$u(\alpha x) = u(\delta^m x) = \beta_\delta^m u(x)$$

where β_δ is chosen so that $u(\delta x) = \beta_\delta u(x)$. ¿From these relationships, we may infer that

$$\begin{aligned}
\beta_\delta^m &= \beta = \alpha^p \\
\beta_\delta &= \alpha^{p/m} \\
f(\delta^n x) &= \delta^n f(x) \\
f(\alpha^{1/m} x) &= \alpha^{1/m} f(x) \\
f(\alpha^{n/m} x) &= \alpha^{n/m} f(x) \ .
\end{aligned}$$

Since m and n are arbitrary integers and f is continuous, we have that for all $c \geq 1$, $f(cx) = cf(x)$. Let $0 < c < 1$, then $(1/c)f(cx) = f(x)$, and therefore $f(cx) = cf(x)$. Therefore, we have that for all $c \geq 0$, $f(cx) = cf(x)$.

b) Assume that there exists $f(x_o) = 0$ but $x_o \neq \theta$. For all y,

$$\begin{aligned}
u(x_o + y) &\leq (1/2)u(2x_o) + (1/2)u(2y) \\
&= (1/2)u(2y) \\
&= 2^{p-1}u(y)
\end{aligned}$$

where the first inequality is by convexity, and the second by the fact that $f(2x_o) = 0$. By continuity of f, we may define the set

$$A_\epsilon = \{y : y^t x_o = 0, \ ||y|| < \epsilon\}$$

such that for all $y \in A_\epsilon$, $u(y) < 1$. Using the above inequality, we have that

$$\begin{aligned}
\int_{I\!R} g(x)dx &\geq \int_{-\infty}^{\infty} \int_{A_\epsilon} \exp\left\{-u(\alpha x_o/||x_o||_2 + y) - u(\theta)\right\} dyd\alpha \\
&\geq \int_{-\infty}^{\infty} \int_{A_\epsilon} \exp\left\{-2^{p-1} - u(\theta)\right\} dyd\alpha \\
&= \infty \ .
\end{aligned}$$

c) First choose any x and y so that $f(x) = f(y) = c \neq 0$. Then for any $0 < \lambda < 1$,

$$f(\lambda x + (1-\lambda)y) = cf\left(\frac{\lambda}{c}x + \frac{(1-\lambda)}{c}y\right)$$

$$
\begin{aligned}
&= c\left(u\left(\frac{\lambda}{c}x + \frac{(1-\lambda)}{c}y\right)\right)^p \\
&= c\left(\lambda u(x/c) + (1-\lambda)u(y/c)\right)^p \\
&= c\left(\frac{\lambda}{c}u(x) + \frac{(1-\lambda)}{c}u(y)\right)^p \\
&= c \\
&= \lambda f(x) + (1-\lambda) f(y)\ .
\end{aligned}
$$

Now choose any $x, y \neq \theta$, then $f(x), f(y) \neq 0$. Define,

$$\lambda = \frac{f(x)}{f(x) + f(y)}$$

and also define $x\prime = x/\lambda$ and $y\prime = y/(1-\lambda)$. Then since $f(x\prime) = f(y\prime)$, we may apply the above result to yield the triangle inequality for $f(\cdot)$.

$$
\begin{aligned}
f(x+y) &= f(\lambda x\prime + (1-\lambda)y\prime) \\
&\geq \lambda f(x\prime) + (1-\lambda) f(y\prime) \\
&= f(x) + f(y)
\end{aligned}
$$

References

[1] H. Derin, H. Elliot, R. Cristi, and D. Geman, "Bayes Smoothing Algorithms for Segmentation of Binary Images Modeled by Markov Random Fields," *IEEE Trans. Trans. Pattern Anal. and Mach. Intell.*, vol. PAMI-6, no.6 , pp. 707-720, Nov. 1984.

[2] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. PAMI-6, no.6, pp. 721-741, Nov. 1984.

[3] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. Roy. Statist. Soc. B,* vol. 48, no. 3, pp. 259-302, 1986.

[4] H. Derin and H. Elliott, "Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields," *IEEE Trans. Pat. An. Mach. Intell.,* vol. PAMI-9, pp. 39-55, Jan. 1987.

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

[6] C. Bouman and B. Liu, "A Multiple Resolution Approach to Regularization," *Proc. SPIE Conf. on Visual Comm. and Image Proc.*, pp. 512-520, Cambridge, MA, Nov. 9-11, 1988.

[7] S. Geman and D. McClure, "Bayesian Image Analysis: An Application to Single Photon Emission Tomography," in *Proc. Statist. Comput. Sect. Amer. Stat. Assoc.*, Washington, DC, pp. 12-18, 1985.

[8] S. Geman and D. McClure, "Statistical Methods for Tomographic Image Reconstruction," *Bull. Int. Stat. Inst.*, vol. LII-4, pp. 5-21, 1987.

[9] T. Hebert and R. Leahy, "A Generalized EM Algorithm for 3-D Bayesian Reconstruction from Poisson data Using Gibbs Priors," *IEEE Trans. Med. Im.*, vol. 8, no. 2, pp. 194-202, June 1989.

[10] R. Stevenson and E. Delp, "Fitting Curves with Discontinuities," *Proc. of the First International Workshop on Robust Computer Vision,* pp. 127-136, Seattle, WA, Oct. 1-3, 1990.

[11] P. Huber, *Robust Statistics,* John Wiley & Sons, New York, NY, 1981.

[12] P. J. Green "Bayesian Reconstructions from Emission Tomography Data Using a Modified EM Algorithm,: *IEEE Trans. Med. Im.*, vol. 9, no. 1, pp. 84-93, March 1990.

[13] K. Lange, "Convergence of EM Image Reconstruction Algorithms with Gibbs Priors," *IEEE Trans. Med. Im.*, vol. 9, no. 4, pp. 439-446, Dec. 1990.

[14] S. A. Kassam, *Signal Detection in Non-Gaussian Noise,* Springer-Verlag, New York, 1988.

[15] G. T. Herman, H. Hurwitz, A. Lent, and H-P. Lung, "On the Bayesian Approach to Image Reconstruction," *Info. and Cont.,* vol. 42, pp. 60-71, 1979.

[16] K. M. Hanson and G. W. Wecksung, "Bayesian Approach to Limited-Angle Reconstruction in Computed Tomography," *J. Opt. Soc. Am.,* vol. 73, no. 11, pp. 1501-1509, Nov. 1983.

[17] E. Levitan and G. T. Herman, "A Maximum *A Posteriori* Probability Expectation Maximization Algorithm for Image Reconstruction in Emission Tomography," *IEEE Trans. Med. Imag.,* vol. MI-6, No. 3, pp. 185-192, 1987.

[18] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic Solution of Ill-Posed Problems in Computational Vision," *J. of the Am. Stat. Assoc.* vol. 82, pp 76-89, March 1987.

[19] B. Horn and B. Schunck, "Determining Optical Flow," *Artificial Intelligence,* vol. 17, pp. 185-203, 1981.

[20] K. Ikeuchi and B. Horn, "Numerical Shape from Shading and Occluding Boundaries," *Artificial Intelligence,* vol. 17, pp. 141-183, 1981.

[21] V. Torre and T. Poggio, "On Edge Detection," *IEEE Trans. Pat. An. Mach. Intell.,* vol. PAMI-8, no. 2, pp. 147-163, March 1986.

[22] H. Van Trees, *Detection, Estimation, and Modulation Theory,* John Wiley & Sons, New York, 1968.

[23] C. Bouman and B. Liu, "Multiple Resolution Segmentation of Textured Images," *IEEE Trans. on Pat. An. Mach. Intell.*, vol. 13, no. 2, pp. 99-113, Feb. 1990.

[24] J. Besag, "Spatial Interaction and the Statistical Analysis of Lattice Systems," *J. Royal Stat. Soc. B*, vol. 36, pp. 192-326, 1974.

[25] R. Kindermann and J. L. Snell, *Markov Random Fields and their Applications.* Providence: American Mathematical Society, 1980.

[26] J. Hutchinson, C. Koch, J. Luo and C. Mead, "Computing Motion Using Analog and Binary Resistive Networks," *Computer,* vol. 21, pp. 53-63, March 1988.

[27] D. Geman and G. Reynolds, "Constrained Restoration and the Recovery of Discontinuities," *IEEE Trans. Pat. An. Mach. Intell.,* vol. 14, no. 3, pp. 367-383, March 1992.

[28] C. Bouman and K. Sauer, "An Edge Preserving Method for Image Reconstruction from Integral Projections," *Proc. Conf. on Inf. Sci. and Sys.* Johns Hopkins University, Baltimore, MD, Mar. 20-22, 1991, pp. 382-387.

[29] N. Gallagher and G. Wise, "A Theoretical Analysis of the Properties of Median Filters," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-29, no. 6, pp. 1136-1141, Dec. 1981.

[30] E. J. Coyle and J. H. Lin, "Stack filters and the Mean Absolute Error Criterion," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 36, pp. 1244-1254, Aug. 1988.

[31] A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems,* Winston and Sons, New York, 1977.

[32] C. Bouman and K. Sauer, "A Generalized Gaussian Image Model for Edge-Preserving MAP Estimation," Technical Report TR-EE-92-1, School of Electrical Engineering, Purdue University, Jan. 1992.

[33] J. Zhang and J. W. Modestino, "A Model-Fitting Approach to Cluster Validation with Application to Stochastic Model-Based Image Segmentation," *Proc. IEEE Int'l Conf. on Acoust., Speech and Sig. Proc.*, pp. 1148-1151, New York, NY, April 11-14, 1988.

[34] J. Besag, "Towards Bayesian Image Analysis," *Journal of Applied Statistics,* vol. 16, no. 3, pp. 395-407, 1989.

[35] W. Rey, *Introduction to Robust and Quasi-Robust Statistical Methods,* Springer-Verlag, Berlin, 1980.

[36] T. Nodes and N. Gallagher, "Median Filters: Some Modifications and Their Properties," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-30, no. 5, pp. 739-746, Oct. 1982.

[37] J. P. Fitch, E. Coyle and N. Gallagher, "Root Properties and Convergence Rates of Median Filters," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-33, no. 1, pp. 230-239, Feb. 1985.

[38] F. Beckman, "The Solution of Linear Equations by the Conjugate Gradient Method," in eds. A. Ralston, H. Wilf and K. Enslein, *Mathematical Methods for Digital Computers,* Wiley, 1960.

[39] O. Yli-Harja, J. Astola and Y. Neuvo, "Analysis of the Properties of Median and Weighted Median Filters Using Threshold Logic and Stack Filter Representation," *IEEE Trans. Signal Processing,* vol. 39, no. 2, Feb. 1991.

[40] J. Astola and Y. Neuvo, "Matched Median Filtering," to appear in *IEEE Trans. on Communications.*

[41] K. Sauer and C. Bouman, "Bayesian Estimation of Transmission Tomograms Using Segmentation Based Optimization," *IEEE Trans. on Nuclear Science*, vol. 39, no. 4, pp. 1144-1152, Aug. 1992.

[42] K. Sauer and C. Bouman, "Bayesian Estimation from Projections with Low Photon Dosages," *Proc. of IEEE Int'l Conf. on Acoust., Speech and Sig. Proc.*, Toronto, Ontario, May 14-17, 1991, pp. 2593-2596.

[43] K. Sauer and C. Bouman, "A Local Update Strategy for Iterative Reconstruction from Projections," to appear Feb. 1993, *IEEE Trans. on Sig. Proc.*.

[44] C. Bouman and K. Sauer, "Nonlinear Multigrid Methods of Optimization in Bayesian Tomographic Image Reconstruction," to appear in the *Proc. of SPIE Conf. on Neural and Stochastic Methods in Image and Signal Processing*, San Diego, California, July 19-24, 1992.

[45] R. DeVore and B. Lucier, "Fast Wavelet Techniques for Near-Optimal Image Processing," *Proc. of IEEE Military Communications Conference*, San Diego, CA, October 11-14, 1992, pp. 1129-1135.

[46] P. Meer, D. Mintz, and A. Rosenfeld, "Robust Regression Methods for Computer Vision: A Review," *Inter. J. of Comp. Vision*, vol. 6, no 1, pp. 59-70, April 1983.

[47] D. M. Young, *Iterative Solution of Large Linear Systems,* Academic Press, New York, 1971.

[48] S.P. Boyd and C.H. Barrett, *Linear Controller Design-Limits of Performance,* Prentice-Hall, Englewood Cliffs, New Jersey, 1991.

Figure 13: CBP reconstruction of concrete block from 45×313 projections. Resolution is 314×314 pixels.

a	b
c	d

Figure 14: MAP estimates of concrete block with (a) $p = 2$ and $y_T/\lambda^2 = 8 \times 10^3$; (b)-(d) $p = 1$ with $y_T/\lambda = 4 \times 10^4, 4 \times 10^3$, and 4×10^2, respectively.