Adaptive Wavelet Graph Model for Bayesian Tomographic Reconstruction*

Thomas Frese*,[1], Charles A. Bouman[1] and Ken Sauer[2]

[1]School of Electrical and Computer Engineering, Purdue University, 1285 ECE Building, West Lafayette, IN 47907-1285, USA, Phone: (765) 494-0340, Fax: (765) 494-3358, Email: {frese, bouman}@ecn.purdue.edu

[2]University of Notre Dame, Department of Electrical Engineering, 270 Fitzpatrick Hall, Notre Dame, IN 46556-5637, USA, Phone: (219) 631-6999, Fax: (219) 631-4393, Email: sauer@nd.edu.

Abstract

We introduce an adaptive wavelet graph image model applicable to Bayesian tomographic reconstruction and other problems with non-local observations. The proposed model captures coarse-to-fine scale dependencies in the wavelet tree by modeling the conditional distribution of wavelet coefficients given overlapping windows of scaling coefficients containing coarse scale information. This results in a graph dependency structure which is more general than a quadtree, enabling the model to produce smooth estimates even for simple wavelet bases such as the Haar basis. The inter-scale dependencies of the wavelet graph model are specified using a spatially non-homogeneous Gaussian distribution with parameters at each scale and location. The parameters of this distribution are selected adaptively using nonlinear classification of coarse scale data. The nonlinear adaptation mechanism is based on a set of training images. In conjunction with the wavelet graph model, we present a computationally efficient multiresolution image reconstruction algorithm. This algorithm is based on iterative Bayesian space domain optimization using scale recursive updates of the wavelet graph prior model. In comparison to performing the optimization over the wavelet coefficients, the space domain formulation facilitates enforcement of pixel positivity constraints. Results indicate that the proposed framework can improve reconstruction quality over fixed resolution Bayesian methods.

EDICS numbers IP *2-WAVP, 3-TOMO*

Correspondence to: Thomas Frese
School of ECE, Purdue University
1285 Electrical and Computer Engineering Building
West Lafayette, IN 47907-1285
Office: (765) 494-0626, Fax: (765) 494-3358
Email: frese@ecn.purdue.edu

*This work was supported by the National Science Foundation under Grant MIP97-07763 and the State of Indiana 21st Century Research and Technology Fund "Indiana Center of Excellence in Biomedical Imaging".

1 Introduction

A major challenge for Bayesian image reconstruction methods is the design of image prior models that accurately account for edges as well as uniform and textured regions in images, yet result in tractable estimation algorithms. In comparison to Markov random field (MRF) priors, multiresolution models can improve accuracy and increase computational efficiency. However, little work has been done on applying multiresolution prior models to Bayesian tomographic reconstruction and other problems with non-local observations. In addition, a well known problem with many multiresolution approaches is that the resulting estimates exhibit blockiness which is usually the result of a quadtree dependency structure. In this work, we address these issues by introducing an adaptive wavelet graph prior model. The inter-scale dependencies of this model are not limited to a quadtree structure, resulting in smooth estimates even for simple wavelet bases such as the Haar basis. In conjunction with this model, we propose a fast, iterative multiresolution reconstruction algorithm that can incorporate space domain constraints such as positivity, and thus, is applicable to Bayesian tomographic reconstruction.

Markov random field priors[1, 2, 3] have enjoyed considerable success in Bayesian image reconstruction[4] and restoration[1]. However, MRF approaches are typically limited to modeling very local interactions in images. Several MRF potential functions have been proposed that provide good edge preservation without explicitly modeling edges[5, 6, 7, 8, 9, 10, 11]. In comparison to MRF priors, multiresolution methods can improve reconstruction quality and offer fast and robust estimation algorithms[12, 13, 14, 15, 16, 17]. Multiresolution models better account for long range interactions and can more easily be designed to separately account for edges, smooth and textured regions.

In recent years, multiresolution techniques have been developed which use linear system models on trees[18, 12, 13, 14, 19, 20, 21, 22, 23]. Nonlinear extensions of those methods have been applied to image restoration with both Gaussian and Poisson noise[24, 16, 17, 25, 26]. Other methods have been developed for image segmentation[27, 28, 29, 30]. Most of the existing work on multiresolution techniques has focused on applications where the observations are spatially localized. Typically, the observations are assumed to be conditionally independent given the local state of the model[18, 12, 20, 21, 16, 17]. For this class of problems, the application of multiresolution models defined on quadtrees is very appealing because it leads to non-iterative, scale-recursive estimation and realization algorithms.

Little work, however, has been done on applying multiresolution Bayesian estimation to problems with non-local observations as encountered in tomographic reconstruction. Perhaps, this is because positivity constraints, which are essential in tomographic reconstruction, are difficult to enforce in the wavelet do-

main. In [31], Wu, Herman, and Browne used a wavelet prior for MAP reconstruction of tomographic data. Their prior is based on a space-variant simultaneous autoregressive (SAR) model whose coefficients are extracted using anisotropic diffusion. Nowak and Kolaczyk have proposed a tomographic reconstruction technique using a wavelet prior[25, 26]. Their approach uses the expectation maximization (EM) algorithm to decouple the estimation problem from the linear projection transformation of the observations[6]. The quadtree structure of their model allows for closed-form EM updates; however, since the overall EM technique remains iterative, the computational benefit of using a quadtree structure is less clear. For the Poisson case, their approach is limited to the Haar wavelet basis. Saquib, Bouman, and Sauer proposed a prior model for tomography that used a multiresolution pyramid representation[32]. However, a disadvantage of this pyramid representation is that different scales contain redundant information. This makes formulation of a consistent Bayesian estimator difficult.

A general problem for multiresolution models formulated on quadtrees is blockiness of the resulting estimates[20, 33, 34, 28, 35, 36, 21]. Specifically, nodes that are spatially adjacent can be far apart in the quadtree so that their correlation is poorly modeled. A popular fix is to average multiple estimates obtained for different spatial alignments of the tree or wavelet basis[37, 33, 36]. More elegant approaches have used trees with nodes corresponding to overlapping portions of the image domain[35] or have performed state augmentation to account for the dependencies of general wavelet bases from within a quadtree structure[23]. These approaches have in common that their data representation is highly overcomplete which can make accurate modeling of sampled data difficult.

A more direct way to avoid blockiness is to use a dependency structure that is more general than the quadtree. For image segmentation, Bouman and Shapiro[28] have used a pyramidal graph where each node depends on a fixed size window at the next coarser scale. Kato, Berthod, and Zerubia proposed a fully 3-D MRF where each node's neighborhood consists of adjacent nodes at the same scale and its quadtree parent[27, 38]. A disadvantage of violating the tree constraint is that Bayesian reconstruction must be performed iteratively as compared to the recursive algorithms available for quadtree models. For applications such as tomographic reconstruction this is not a limitation since the forward model requires iterative optimization in any case.

In this work, we develop a stochastic multiresolution framework for Bayesian image reconstruction for problems with spatially non-local measurements. We propose a wavelet graph prior model in combination with a computationally efficient multiresolution reconstruction algorithm applicable to iterative tomographic reconstruction.

The basic concept of the proposed wavelet graph model is to exploit dependencies of wavelet coefficients

across scales. We capture these dependencies by modeling the wavelet coefficients at each scale and location as a function of a window of scaling coefficients at the same scale. This structure has several important implications: First, by conditioning the wavelet coefficients on overlapping windows of scaling coefficients, the dependencies are not limited to a quadtree structure. Secondly, the structure is such that the optimal wavelet graph model for a stationary process is homogeneous at each scale, resulting in a substantial reduction in the number of model parameters. Finally, the model is causal in scale, not overcomplete, and each wavelet coefficient is a function of only a few scaling coefficients. The resulting prior is not suitable for non-iterative scale-recursive optimization; however, it allows for very efficient iterative optimization using scale recursive updates.

The conditional distributions of the wavelet coefficients are modeled using a spatially non-homogeneous Gaussian distribution with image-adaptive parameters. As compared to a fully nonlinear approach, the Gaussian model is suitable for global optimization in a Bayesian framework while the adaptation can account for non-linear dependencies. The parameter adaptation is based on nonlinear classification of coarse scale data. The classifiers and class parameters used for the adaptation are obtained using training data. The training procedure allows the overall model to incorporate characteristics of typical reconstructions as prior information.

The proposed multiresolution reconstruction algorithm computes a coarse-to-fine scale sequence of Bayesian MAP estimates. Each estimate in the sequence is computed with fixed adaptation of the wavelet graph model followed by re-adaptation. The MAP estimates are computed in the space domain using scale-recursive updates of the multiresolution prior. The space domain formulation of the optimization is essential for application of our approach to tomographic reconstruction since it allows positivity constraints to be enforced independently of the wavelet basis.

The paper is organized as follows: Sections 2.1 and 2.2 introduce the image model and develop the overall structure of the multiresolution reconstruction algorithm. Section 2.3 provides a detailed discussion of the space domain optimization algorithm. Section 2.4 describes our implementation of the nonlinear classifiers used for the adaptation of the multiresolution prior. Section 3 discusses the application of the proposed prior to tomographic reconstruction and section 4 provides experimental results.

2 Image Model and Multiresolution Reconstruction Algorithm

2.1 Wavelet Graph Model

To define the notation for the wavelet decomposition, consider the one-dimensional case. Let x be the N-dimensional vector of the image pixel values in raster order and let x_s denote the pixel value at location

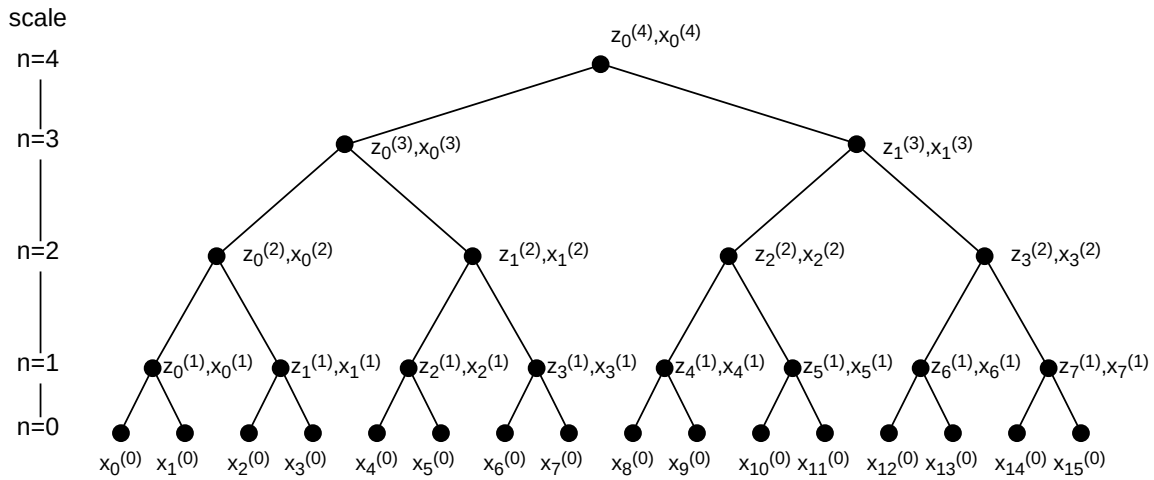


Figure 1: Wavelet decomposition in 1-D for $L = 4$ resolution levels. The wavelet transform of the original image $x^{(0)} = x$ is given by $z = (z^{(1)}, \cdots, z^{(L)}, x^{(L)})$.

s. We now consider the class of wavelet decompositions that can be computed using the recursions

$$x_s^{(n+1)} = \sum_k h_{k-2s} x_k^{(n)} \tag{1}$$

$$z_s^{(n+1)} = \sum_k g_{k-2s} x_k^{(n)} \tag{2}$$

where $x^{(0)} = x$, $0 \leq n < \log_2 N$, h is the lowpass analysis filter, and g is the highpass analysis filter. We assume both h and g to have finite support. In this notation, $x_s^{(n)}$ denotes the scaling coefficient and $z_s^{(n)}$ the wavelet or detail coefficient at scale n and location s. An illustration of this decomposition is shown in Fig. 1.

In the following, we will assume an L resolution wavelet decomposition. The non-overcomplete wavelet transform of x is then specified by the wavelet coefficients $z^{(1)} \dots z^{(L)}$ and the coarsest scale scaling coefficients $x^{(L)}$. Using matrix notation, the L resolution wavelet transform of x can be written as

$$z \stackrel{\triangle}{=} \begin{bmatrix} z^{(1)} \\ \vdots \\ z^{(L)} \\ x^{(L)} \end{bmatrix} = Wx \tag{3}$$

where for simplicity of notation, we suppress the dependence of the wavelet transform matrix W on L. Note, that the transform vector z includes the scaling coefficients $x^{(L)}$.

The basic concept of our image model is to exploit the dependencies of the wavelet coefficients across scales. To formulate the approach, we write the distribution $\log p_z(z)$ in terms of the conditional distri-

bution at each scale given the information at all coarser scales:

$$\log p_z(z) = \log p_{x^{(L)}}(x^{(L)}) + \sum_{n=1}^{L} \log p_{z^{(n)}|z^{(n+1)},\ldots,z^{(L)},x^{(L)}}(z^{(n)}|z^{(n+1)},\ldots,z^{(L)},x^{(L)}) \ . \tag{4}$$

Since the scaling coefficients $x^{(n)}$ contain exactly the same information as $z^{(n+1)},\ldots,z^{(L)},x^{(L)}$, we may rewrite (4) as

$$\log p_z(z) = \log p_{x^{(L)}}(x^{(L)}) + \sum_{n=1}^{L} \log p_{z^{(n)}|x^{(n)}}(z^{(n)}|x^{(n)}) \tag{5}$$

where $x^{(n)}$ is a function of $z^{(n+1)},\ldots,z^{(L)},x^{(L)}$. The distribution assumption for $x^{(L)}$ is typically not important due to the high signal-to-noise ratio at the coarsest resolution. Therefore, we assume that $x^{(L)}$ is uniformly distributed[1]. Thus, we can write

$$\log p_z(z) = \sum_{n=1}^{L} \log p_{z^{(n)}|x^{(n)}}(z^{(n)}|x^{(n)}) + \text{const} \ . \tag{6}$$

To obtain a practical model, we assume the wavelet coefficients at different locations to be conditionally independent given the scaling coefficients at the same scale. Furthermore, we assume the wavelet coefficients at each location only to depend on a small window of scaling coefficients. Let s denote a spatial location at a given scale n such that $z_s^{(n)}$ is the vector of the wavelet coefficients at location s. For the 2-D case, $z_s^{(n)}$ has three components corresponding to the high-low, low-high, and high-high coefficients of a separable wavelet decomposition. We then define ∂s as a window with finite support centered at position s with circular boundary conditions. Then $x_{\partial s}^{(n)}$ is the set of scaling coefficients within the window ∂s at resolution n. Further, let $S^{(n)}$ denote the set of all locations s of the wavelet decomposition at scale n. Using this notation, our assumptions are that the $z_s^{(n)}$ are conditionally independent and that $z_s^{(n)}$ depends only on $x_{\partial s}^{(n)}$. Applying these assumptions to (6) results in the model

$$\boxed{\log p_z(z) = \sum_{n=1}^{L} \sum_{s \in S^{(n)}} \log p_{z_s^{(n)}|x_{\partial s}^{(n)}}(z_s^{(n)}|x_{\partial s}^{(n)}) + \text{const} \ .} \tag{7}$$

We will call any model of the form (7) a wavelet graph model. Figure 2 illustrates the spatial dependencies of the model (7) for the case of a one dimensional signal and a three point window $\partial s = \{s-1, s, s+1\}$. In this case, the conditional distribution of $z_s^{(n)}$ depends only on the three scaling coefficients $x_{\partial s}^{(n)} = \{x_{s-1}^{(n)}, x_s^{(n)}, x_{s+1}^{(n)}\}$.

An important advantage of the structure (7) is that the optimal wavelet graph model for a stationary process is homogeneous. By homogeneous, we mean that $\log p_{z_s^{(n)}|x_{\partial s}^{(n)}}(z_s^{(n)}|x_{\partial s}^{(n)}) = f_n(z_s^{(n)}, x_{\partial s}^{(n)})$ for some

[1]For applications where $x^{(L)}$ is not guaranteed to be within a compact set, the uniform distribution of $x^{(L)}$ is improper, however, in our experience, this does not cause practical problems with the model. For tomography applications, $x^{(L)}$ is within a compact set since its components are non-negative and bounded from above by a constant proportional to the maximum emission rate or dosage used with the data acquisition system.

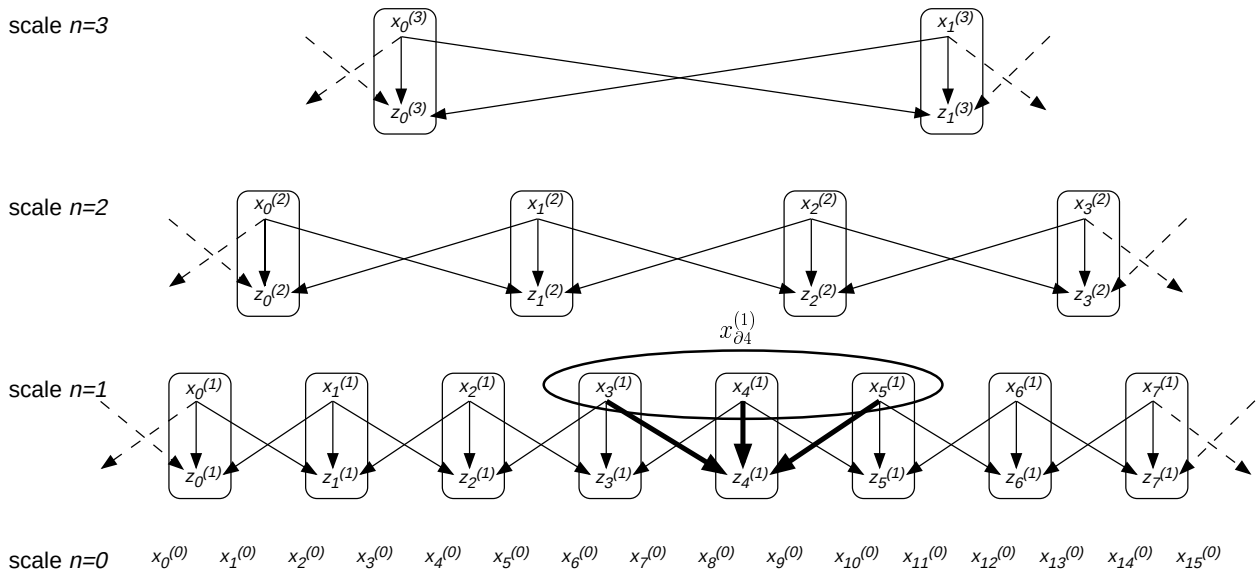


Figure 2: Spatial dependencies for the 1-D case. Using a prediction window $\partial s = \{s-1, s, s+1\}$, each wavelet coefficient $z_s^{(n)}$ depends on the three scaling coefficients $\{x_{s-1}^{(n)}, x_s^{(n)}, x_{s+1}^{(n)}\}$. Notice that the scaling coefficients $x^{(n)}$ contain all the information at coarser scales $l > n$.

functions f_n that do not depend on s. The following theorem, proven in Appendix A, makes the statement above precise:

Theorem 1 *Let $\{X_s^{(0)}\}_{s=0}^{2^N-1}$ be a 1-D discrete-time random process that is circularly stationary, i.e. $X_s^{(0)}$ and $V_s = X_{(s-k) \bmod 2^N}^{(0)}$ have the same distribution. Let $X_s^{(n)}$ and $Z_s^{(n)}$ for $1 \leq n \leq L$ be the wavelet decomposition of $X_s^{(0)}$ as specified by (1) and (2) using circular boundary conditions and let $\log p_z(z)$ be a wavelet graph model of the form (7) with*

$$\log p_{z_s^{(n)}|x_{\partial s}^{(n)}}(z_s^{(n)}|x_{\partial s}^{(n)}) = f_n(z_s^{(n)}, x_{\partial s}^{(n)}, \nu_s^{(n)})$$

where the $\nu_s^{(n)}$ are parameters of the model. Assume there exists a unique minimizer of the relative entropy (Kullback-Leibler distance)

$$\nu^* = \arg\min_{\nu} E[-\log p_z(Z)] \ . \tag{8}$$

Then, $(\nu_s^{(n)})^*$ is not a function of s, implying a homogeneous wavelet graph model $\log p_z(z)$.

The property stated in this theorem greatly simplifies parameter estimation since we only need to estimate a single set of parameters at each scale. In general, this would not be the case if we conditioned the inter-scale dependencies on wavelet coefficients instead of the window of scaling coefficients. This homogeneity property is very important since it dramatically reduces the number of free parameters in the model,

thereby allowing practical model estimation from sampled data. In section 2.4, we will use this property to justify the design of a single nonlinear classifier for each scale.

We first consider the case of a spatially non-homogeneous Gaussian model. In this case, the conditional distributions $\log p_{z_s^{(n)}|x_{\partial s}^{(n)}}(z_s^{(n)}|x_{\partial s}^{(n)})$ must be of the form

$$\log p_{z_s^{(n)}|x_{\partial s}^{(n)}}(z_s^{(n)}|x_{\partial s}^{(n)}) = -\frac{1}{2}\|z_s^{(n)} - A_s^{(n)}x_{\partial s}^{(n)} - \sigma_n b_s^{(n)}\|^2_{\sigma_n^{-2}B_s^{(n)}} + \text{const} \tag{9}$$

where $A_s^{(n)}$ is a matrix, $b_s^{(n)}$ is a column vector, $B_s^{(n)}$ is a positive definite matrix, σ_n is a scaling constant, and $\|\cdot\|_B$ denotes the norm such that $\|\varepsilon\|_B^2 = \varepsilon^t B\varepsilon$. We note that for this model, the conditional mean of $z_s^{(n)}$ is an affine function of $x_{\partial s}^{(n)}$ given by

$$\mu_s^{(n)} \triangleq A_s^{(n)}x_{\partial s}^{(n)} + \sigma_n b_s^{(n)} \ , \tag{10}$$

and when $\sigma_n^2 = 1$, the conditional covariance of $z_s^{(n)}$ is given by $R_s^{(n)} \triangleq (B_s^{(n)})^{-1}$. Consequently, the model is parameterized by $A_s^{(n)}, b_s^{(n)}, B_s^{(n)}$ and a global scaling constant σ_n for each scale n. The scaling constants σ_n will play an important role later by allowing us to use the same adaptation mechanism for different amplitude scalings of the image x. Assuming the wavelet transform W is orthonormal, then $\det W = 1$, and we may express $\log p_x(x)$ as

$$\begin{aligned} \log p_x(x) &= \log p_z(Wx) & (11)\\ &= -\frac{1}{2}\sum_{n=1}^{L}\sum_{s\in S^{(n)}} \|z_s^{(n)} - A_s^{(n)}x_{\partial s}^{(n)} - \sigma_n b_s^{(n)}\|^2_{\sigma_n^{-2}B_s^{(n)}} + \text{const} & (12)\end{aligned}$$

where $z_s^{(n)} = (Wx)_s^{(n)}$. The model (12) is used as the prior distribution for the Bayesian reconstruction of $x = x^{(0)}$.

To formulate a multiresolution reconstruction algorithm, we also want to directly calculate coarse scale reconstructions using the prior model only for coefficients at scales coarser than the reconstruction scale. Thus, we define a coarse scale prior model for the direct reconstruction of the scaling coefficients $x^{(l)}$ at scale l as

$$\begin{aligned} \log p_{x^{(l)}}(x^{(l)}) &= \log p_{z^{(l+1)},\ldots,z^{(L)},x^{(L)}}(z^{(l+1)},\ldots,z^{(L)},x^{(L)}) & (13)\\ &= -\frac{1}{2}\sum_{n=l+1}^{L}\sum_{s\in S^{(n)}} \|z_s^{(n)} - A_s^{(n)}x_{\partial s}^{(n)} - \sigma_n b_s^{(n)}\|^2_{\sigma_n^{-2}B_s^{(n)}} + \text{const} \ . & (14)\end{aligned}$$

Given noisy measurements y and a physical data model $\log p_{y|x}(y|x)$, we obtain the data model $\log p_{y|x^{(l)}}(y|x^{(l)})$ at scale l as

$$\log p_{y|x^{(l)}}(y|x^{(l)}) = \log p_{y|x}(y|\mathcal{I}_l^0 x^{(l)}) \tag{15}$$

where $\mathcal{I}_l^n$ denotes the interpolation from scale l to scale n. The interpolation $\mathcal{I}_l^0 x^{(l)}$ is obtained as the wavelet reconstruction of $x^{(l)}$ assuming that $z^{(n)} = 0$ for $1 \leq n \leq l$.

Based on (14) and (15), the Bayesian MAP estimate $\hat{x}^{(l)}$ of the scaling coefficients $x^{(l)}$ at scale l is the solution to the optimization problem

$$\hat{x}^{(l)} = \arg \max_{x^{(l)} \geq 0} \{\log p_{y|x^{(l)}}(y|x^{(l)}) + \log p_{x^{(l)}}(x^{(l)})\} \ . \tag{16}$$

We have included the positivity constraint $x^{(l)} \geq 0$ since it is important for tomographic reconstruction. Note that for the special case $l = 0$, expression (16) is the standard MAP equation for the image x.

2.2 Spatially Adaptive Multiscale Reconstruction

We use a multiresolution algorithm to perform the image reconstruction and to adaptively select the parameters of the linear model. As illustrated in Fig. 3, the basic concept of the multiresolution algorithm is to compute a sequence of Bayesian MAP estimates from coarse to fine scale. The algorithm starts with the reconstruction of the scaling coefficients $x^{(L)}$ at the coarsest scale L and then successively performs the reconstructions at the finer scales $l = L-1, \ldots, 0$. At each step in this sequence, the current reconstruction is used to initialize the model parameters at the next finer scale and to re-adapt the parameters at the coarser scales. Let $\theta_s^{(l)}$ denote the vector of model parameters at location s and scale l

$$\theta_s^{(l)} = [A_s^{(l)}, b_s^{(l)}, B_s^{(l)}] \ . \tag{17}$$

After computing the MAP reconstruction $\hat{x}^{(l)}$ at scale l, we update all the parameters $\theta_s^{(n)}$ for $n \geq l$. These new parameters are then used to reconstruct $x^{(l-1)}$ at the next finer scale.

We update the parameters $\theta_s^{(n)}$ using a nonlinear classification method derived from recent work in image interpolation[39]. More specifically, we update $\theta_s^{(n)}$ by applying the nonlinear operator $T^{(n)}[\cdot]$ to the window of scaling coefficients $\hat{x}_{\partial s}^{(n)}$:

$$\theta_s^{(n)} \leftarrow T^{(n)} \left[\frac{\hat{x}_{\partial s}^{(n)}}{\sigma_n} \right] \ . \tag{18}$$

Note that based on Theorem 1, we can use a single $T^{(n)}[\cdot]$ for all locations s at scale n. The normalization by σ_n in (18) is included to account for possible scaling of the image $x^{(n)}$. To obtain a robust estimate, we compute the 10% trimmed mean[40] of $\hat{x}^{(n)}$ over the approximate support of the active image region:

$$\sigma_n \leftarrow AVG_{trim10\%}[\hat{x}^{(n)}] \ . \tag{19}$$

The nonlinear operators $T^{(n)}[\cdot]$ are obtained during a training phase. The structure of the $T^{(n)}[\cdot]$ and the training procedure are explained in section 2.4.

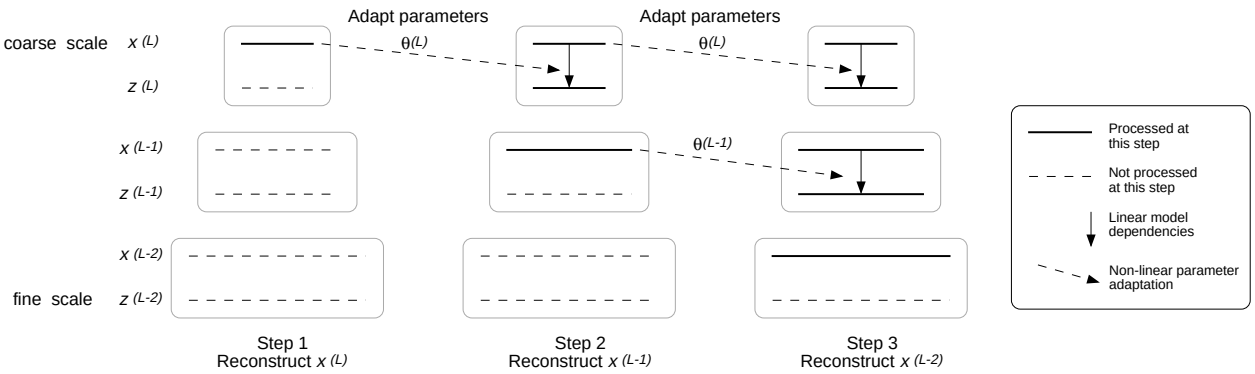


Figure 3: Illustration of the multiresolution reconstruction algorithm. Starting at the coarsest scale $l = L$, we calculate a coarse-to-fine scale sequence of MAP estimates $x^{(l)}$. Each estimate $x^{(l)}$ takes into account the contributions of the adapted prior model at *all* coarser scales $n > l$. After computing $x^{(l)}$, we initialize the parameters $\theta_s^{(l)}$ and re-adapt $\theta_s^{(n)}$ for $n > l$.

1. Initialize $\hat{x}^{(n)} = 0$ and $\hat{z}^{(n)} = 0$ for all scales n.
2. Compute maximum likelihood (ML) reconstruction $\hat{x}^{(L)}$.
3. For $l = L - 1$ downto 0 {
   (a) Adapt Gaussian model: For $l < n \leq L$, $\forall s$, assign $\theta_s^{(n)}$ and σ_n using (18), (19).
   (b) Initialize $x^{(l)} \leftarrow \mathcal{I}_{l+1}^{l} \hat{x}^{(l+1)}$.
   (c) Calculate MAP reconstruction $\hat{x}^{(l)}$ for Gaussian wavelet graph prior model, using (16).
   (d) Recompute $\hat{x}^{(n)}, \hat{z}^{(n)}$ for $l \leq n \leq L$ as wavelet decomposition of $\hat{x}^{(l)}$.
   }

Figure 4: Summary of multiresolution reconstruction algorithm.

The coarse-to-fine scale multiresolution reconstruction algorithm is summarized in Fig. 4. Note that the final fine scale reconstruction produced by this algorithm is not a conventional MAP estimate. Rather, it is a MAP estimate with respect to a data dependent prior. This prior is the spatially non-homogeneous linear model at all scales with the parameters obtained in the last adaptation step.

2.3 MAP Optimization for Gaussian Wavelet Graph Model

In this section, we describe a computationally efficient implementation of the MAP optimization in step 3.c of Fig. 4. The MAP optimization is performed for the Gaussian wavelet graph prior model with fixed model parameters $\theta_s^{(l)} = [A_s^{(l)}, b_s^{(l)}, B_s^{(l)}]$ and σ_n. The parameter selection is described in section 2.4.

The positivity constraint, $x^{(l)} \geq 0$, is an essential component of the MAP optimization equation (16). However, enforcement of positivity can be very difficult in the wavelet domain, particularly for general wavelet transforms. Figure 5 illustrates our approach for solving this problem. The optimization is performed in the space domain, while the prior model is formulated in the wavelet domain and the system

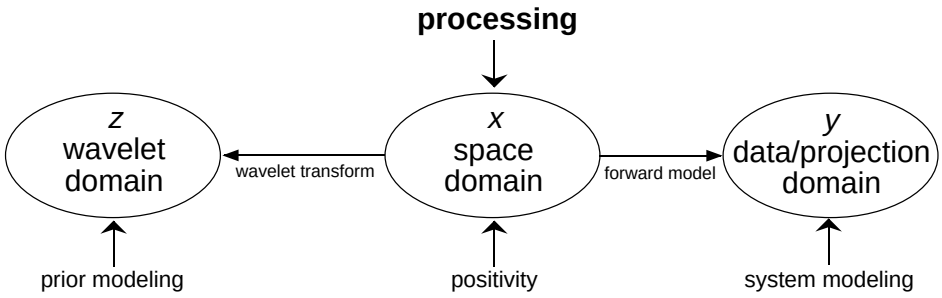


Figure 5: The image is modeled in the wavelet domain but the MAP optimization is performed in the space domain. This allows the positivity constraint to be easily enforced.

model is formulated in the projection domain. This makes enforcement of positivity simple. Another advantage of space domain optimization is that it simplifies the forward model in tomography. This is because the transformation from z to y is generally less sparse than the transformation from x to y.

Below, we derive the space domain MAP optimization algorithm. Since (16) has the same structure for any l, we develop the algorithm for reconstruction at the finest scale $l = 0$. This allows us to simplify the notation by omitting the superscript (l) and writing $x = x^{(0)}$. The solutions for $l > 0$ are obtained using the same algorithm as described at the end of this section.

To derive the optimization algorithm, define matrices $U^{(n)}$ as the subsets of the inverse wavelet transform W^{-1} such that

$$x^{(n)} = U^{(n)} z = U^{(n)} \begin{bmatrix} z^{(1)} \\ \vdots \\ z^{(L)} \\ x^{(L)} \end{bmatrix} . \tag{20}$$

Further, let $U_{\partial s}^{(n)}$ denote the rows of $U^{(n)}$ such that $x_{\partial s}^{(n)} = U_{\partial s}^{(n)} z$. Using this notation, we can rewrite (10) as

$$\mu_s^{(n)} = A_s^{(n)} U_{\partial s}^{(n)} z + \sigma_n b_s^{(n)} \ . \tag{21}$$

To simplify the notation, let us define $\tilde{A}_s^{(n)} = A_s^{(n)} U_{\partial s}^{(n)}$, $\tilde{b}_s^{(n)} = \sigma_n b_s^{(n)}$, and $\tilde{B}_s^{(n)} = \sigma_n^{-2} B_s^{(n)}$. Furthermore, let $\tilde{A}$, $\tilde{b}$, and $\tilde{B}$ denote the parameters for all locations s and all scales n. We can then re-write the model (12) as

$$\log p_x(x) = -\frac{1}{2} \|z - \tilde{A}z - \tilde{b}\|_{\tilde{B}}^2 + \text{const} \tag{22}$$

$$= -\frac{1}{2} \|(I - \tilde{A})Wx - \tilde{b}\|_{\tilde{B}}^2 + \text{const} \tag{23}$$

where $x = W^{-1} z$.

Given the space domain formulation (23), a variety of optimization strategies can be used to perform the constrained optimization of (16). Below, we derive a coordinate descent strategy which is the focus

of our work.

To optimize $\log p_x(x)$ with respect to a single pixel value x_i, we need the first and second derivatives with respect to x_i. Let us define ε as the prediction error $z - \mu$ of the linear model in the wavelet domain

$$\varepsilon = (I - \tilde{A})Wx - \tilde{b} \ . \tag{24}$$

Based on (23), we can then write the gradient $\nabla_x \log p_x(x)$ as

$$\begin{aligned} \nabla_x \log p_x(x) &= x^t W^t (I - \tilde{A})^t \tilde{B}(I - \tilde{A})W - \tilde{b}^t \tilde{B}(I - \tilde{A})W & (25) \\ &= \varepsilon^t \tilde{B}(I - \tilde{A})W \ . & (26) \end{aligned}$$

If we now let e_i denote the unitary vector in direction x_i, we can write the first derivative α_1 as

$$\begin{aligned} \alpha_1 &= \frac{\partial}{\partial x_i} \log p_x(x) & (27) \\ &= (\nabla_x \log p_x(x)) \ e_i & (28) \\ &= \varepsilon^t \tilde{B}(I - \tilde{A})W_{*i} & (29) \end{aligned}$$

where W_{*i} denotes the i^{th} column of W. Similarly, for the second derivative α_2 we obtain

$$\begin{aligned} \alpha_2 &= \frac{\partial^2}{\partial x_i^2} \log p_x(x) & (30) \\ &= W_{*i}^t (I - \tilde{A})^t \tilde{B}(I - \tilde{A})W_{*i} \ . & (31) \end{aligned}$$

Notice that $(I - \tilde{A})W_{*i}$ is the derivative of the prediction errors in the wavelet domain with respect to x_i. Let us define $d\varepsilon = (I - \tilde{A})W_{*i}$, then

$$\begin{aligned} \alpha_1 &= \varepsilon^t \tilde{B} d\varepsilon & (32) \\ \alpha_2 &= (d\varepsilon)^t \tilde{B} d\varepsilon \ . & (33) \end{aligned}$$

The prior $\log p_x$ can now be written as a function of the perturbation γ of pixel x_i

$$\log p_x(x + \gamma e_i) = \alpha_1 \gamma + \frac{1}{2}\alpha_2 \gamma^2 + \text{const} \ . \tag{34}$$

Using (34), we can apply a standard iterative coordinate descent (ICD)[41, 42] algorithm in x to iteratively optimize the MAP equation (16) with respect to a single pixel at a time. Figure 6 summarizes the basic steps of the ICD optimization algorithm for MAP reconstruction using the Gaussian wavelet graph prior model. The prediction errors ε are kept as a state vector. The set $S = S^{(0)}$ denotes the set of image pixels at scale $n = 0$. The positivity constraint is enforced by limiting Δ in step 2c to values $\Delta \geq -x_i$. If the update in step 2c is performed exactly and the log-posterior function is strictly convex and continuously

1. initialize $\varepsilon = (I - \tilde{A})Wx - \tilde{b}$
2. until convergence do
   for $i \in S$ {
   (a) $d\varepsilon = (I - \tilde{A})W_{*i}$
   (b) $\alpha_1 = \varepsilon^t \tilde{B} d\varepsilon$, $\alpha_2 = (d\varepsilon)^t \tilde{B} d\varepsilon$
   (c) $\Delta = \arg\max_{\gamma \geq -x_i} \{\log p_{y|x}(y|x + \gamma e_i) + \alpha_1 \gamma + 0.5\alpha_2 \gamma^2\}$
   (d) $x_i \leftarrow x_i + \Delta$
   (e) $\varepsilon \leftarrow \varepsilon + d\varepsilon\Delta$
   }

Figure 6: General formulation of the iterative coordinate descent (ICD) optimization algorithm for space domain MAP reconstruction using the Gaussian wavelet graph prior model. The prediction errors ε are kept as a state vector. The set $S = S^{(0)}$ denotes the set of image pixels at scale $n = 0$.

differentiable, the constrained ICD optimization will converge to a global minimum. However, we use a Newton-Raphson procedure for the update in step 2c which in practice has robust convergence. See [43] for the details of the convergence analysis with Newton-Raphson updates.

Below, we derive a computationally efficient implementation of the ICD algorithm by incorporating our model assumptions of conditional independence and limited spatial support. For illustration, we augment the tree structure of the linear model as shown in Fig. 7 for the 1-D case. In this representation, each tree node $\mathcal{N}_s^{(n)}$ contains the wavelet coefficients $z_s^{(n)}$, the scaling coefficient $x_s^{(n)}$ as well as the current prediction error $\varepsilon_s^{(n)} = z_s^{(n)} - A_s^{(n)} x_{\partial s}^{(n)} - \tilde{b}_s^{(n)}$. In addition, $\mathcal{N}_s^{(n)}$ contains the temporary variables $dz_s^{(n)}$ and $dx_s^{(n)}$. These variables are the derivatives of $z_s^{(n)}$ and $x_s^{(n)}$ with respect to the pixel value x_i that is currently being updated in the space domain. To compute $dz_s^{(n)}$ and $dx_s^{(n)}$, let $W^{(n)}$ denote the subset of the wavelet transform such that $z^{(n)} = W^{(n)} x$. Further, let $D^{(n)}$ denote the decimation operation used to obtain the scaling coefficients $x^{(n)}$ from the original image x, so that

$$x^{(n)} = D^{(n)} x^{(0)} = U^{(n)} W x^{(0)} \ . \tag{35}$$

For the update of pixel value x_i, we compute $dz_s^{(n)}$ and $dx_s^{(n)}$ as

$$dz_s^{(n)} = W_{si}^{(n)} \tag{36}$$

$$dx_s^{(n)} = D_{si}^{(n)} \ . \tag{37}$$

The notation in (36) is for the 1-D case, where both $dz_s^{(n)}$ and $dx_s^{(n)}$ are scalars. For the 2-D case, $dz_s^{(n)}$ is the three-component vector containing the derivatives corresponding to the high-low, low-high, and high-high components of $z_s^{(n)}$. Define $dx_{\partial s}^{(n)}$ as the vector with components $dx_k^{(n)}$ for $k \in \partial s$. Using this

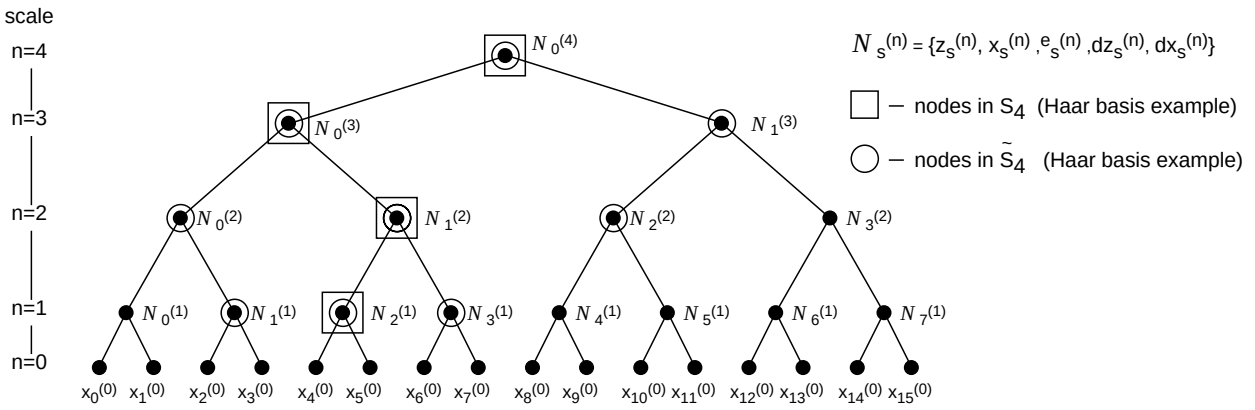


Figure 7: Illustration of ICD update computation for pixel $x_4^{(0)}$ using a Haar wavelet basis and a 3-point window ∂s. To update $x_4^{(0)}$, only the nodes in the set $\tilde{S}_4^{(n)}$, shown in circles, must be considered. Each node $\mathcal{N}_s^{(n)}$ for $n > 0$ contains the variables $z_s^{(n)}, x_s^{(n)}, \varepsilon_s^{(n)}$ as well as the temporary variables $dz_s^{(n)}$ and $dx_s^{(n)}$ required for the computation of the ICD update.

notation, we can write $d\varepsilon_s^{(n)}$ as

$$d\varepsilon_s^{(n)} = dz_s^{(n)} - A_s^{(n)} dx_{\partial s}^{(n)} \ . \tag{38}$$

To compute $d\varepsilon$ efficiently, we want to consider only the locations (s, n) for which $d\varepsilon_s^{(n)}$ is nonzero. Let us define the sets $S_i^{(n)}$ as

$$S_i^{(n)} = \{s : dz_s^{(n)} \neq 0 \text{ or } dx_s^{(n)} \neq 0\} \ . \tag{39}$$

Notice, that the sets $S_i^{(n)}, 1 \leq n \leq L$, are only a function of the wavelet transform W_{*i}. The change $dx_{\partial s}^{(n)}$ is nonzero only at locations whose prediction window includes nodes in $S_i^{(n)}$. We define the set of these locations as $\tilde{S}_i^{(n)}$

$$\begin{aligned} \tilde{S}_i^{(n)} &= \{k : s \in \partial k \text{ for some } s \in S_i^{(n)}\} & (40) \\ &= \bigcup_{s \in S_i^{(n)}} \partial s \ . & (41) \end{aligned}$$

The last equation is a result of the symmetry of the prediction window. Notice, that since the prediction window ∂s includes s, we have $S_i^{(n)} \subset \tilde{S}_i^{(n)}$ such that both $dz_s^{(n)} = 0$ and $dx_{\partial s}^{(n)} = 0$ for $s \notin \tilde{S}_i^{(n)}$ and consequently $d\varepsilon_s^{(n)} = 0$ for $s \notin \tilde{S}_i^{(n)}$. Figure 7 illustrates the definitions of $S_i^{(n)}$ and $\tilde{S}_i^{(n)}$ for the specific example of a 1-D Haar wavelet decomposition.

We can now compute the nonzero components of $d\varepsilon$ as

$$d\varepsilon_s^{(n)} = dz_s^{(n)} - A_s^{(n)} dx_{\partial s}^{(n)} \qquad s \in \tilde{S}_i^{(n)} \ . \tag{42}$$

1. initialize $\varepsilon = (I - \tilde{A})Wx - \tilde{b}$
2. until convergence do
   for $i \in S$ {
   (a) (*) compute $dz_s^{(n)} = W_{si}^{(n)},\ dx_s^{(n)} = D_{si}^{(n)},\ \forall s, n > 0$
   (b) (*) assign $S_i^{(n)} = \{s : dz_s^{(n)} \neq 0 \vee dx_s^{(n)} \neq 0\}$
   (c) (*) assign $\tilde{S}_i^{(n)} = \cup_{s \in S_i^{(n)}} \partial s$
   (d) (*) $d\varepsilon_s^{(n)} = dz_s^{(n)} - A_s^{(n)} dx_{\partial s}^{(n)}$ for $s \in \tilde{S}_i^{(n)}$
   (e) $\alpha_1 = \sum_{n=1}^{L} \sum_{s \in \tilde{S}_i^{(n)}} \varepsilon_s^{(n)} (\sigma_n^{-2} B_s^{(n)}) d\varepsilon_s^{(n)}$
   (f) (*) $\alpha_2 = \sum_{n=1}^{L} \sum_{s \in \tilde{S}_i^{(n)}} d\varepsilon_s^{(n)} (\sigma_n^{-2} B_s^{(n)}) d\varepsilon_s^{(n)}$
   (g) $\Delta = \arg\max_\gamma \left\{ \log p_{y|x}(y|x + \gamma e_i) + \alpha_1 \gamma + 0.5 \alpha_2 \gamma^2 \right\}$
   (h) $x_i \leftarrow x_i + \Delta$
   (i) $\varepsilon_s^{(n)} \leftarrow \varepsilon_s^{(n)} + d\varepsilon_s^{(n)} \Delta$ for $s \in \tilde{S}_i^{(n)}$
   }
3. recompute $\hat{x}^{(n)}, \hat{z}^{(n)}$ for $1 \leq n \leq L$ as wavelet decomposition of $\hat{x}$

Figure 8: Detailed algorithm for efficient ICD updates using wavelet graph model. The operations marked with (*) can be pre-computed. Excluding the forward model term of step 2.f, the computational complexity is order $N(\log N)^2$ for one full update of $\hat{x}^{(0)}$.

The first and second derivatives α_1 and α_2 are then given by

$$\alpha_1 = \sum_{n=1}^{L} \sum_{s \in \tilde{S}_i^{(n)}} \varepsilon_s^{(n)} \tilde{B}_s^{(n)} d\varepsilon_s^{(n)} \tag{43}$$

$$\alpha_2 = \sum_{n=1}^{L} \sum_{s \in \tilde{S}_i^{(n)}} d\varepsilon_s^{(n)} \tilde{B}_s^{(n)} d\varepsilon_s^{(n)} \ . \tag{44}$$

The derivatives α_1 and α_2 are used to perform the MAP optimization with respect to x_i in steps 2.c and 2.d of Fig. 6. After updating x_i, the state variables $\varepsilon_s^{(n)}$ are updated as

$$\varepsilon_s^{(n)} \leftarrow \varepsilon_s^{(n)} + d\varepsilon_s^{(n)} \Delta \qquad s \in \tilde{S}_i^{(n)} \tag{45}$$

for $1 \leq n < L$. This completes the efficient implementation of the iterative coordinate descent optimization. Note that $dx_s^{(n)}, dz_s^{(n)}, S_i^{(n)}, \tilde{S}_i^{(n)}, d\varepsilon_s^{(n)}$, and α_2 are not data dependent but are only a function of i. If desired, these variables can be pre-computed and stored for all i. Figure 8 summarizes the optimization algorithm.

While the development above assumes optimization at the finest scale $l = 0$, the same algorithm can be used for the coarse scale reconstructions $\hat{x}^{(l)}$ for $l > 0$. To optimize (16) for $l > 0$, we treat scale l as if it were the finest scale $n = 0$. Thus, instead of iterating over the pixel locations i at fine scale, we now index over $i \in S^{(l)}$. The state variables $z_s^{(n)}, x_s^{(n)}, \varepsilon_s^{(n)}$ as well as the directions $dz_s^{(n)}, dx_s^{(n)}, d\varepsilon_s^{(n)}$ are computed using the wavelet transform of $x^{(l)}$ and are computed for $n > l$ only.

The computational complexity associated with the optimization of the linear image model is on the order of $KN^{(l)}(\log N^{(l)})^2$ multiplications for one full update of $\hat{x}^{(l)}$, where $N^{(l)}$ is the number of image pixels at the reconstruction scale l and $K = |\partial s|$ is the number of coefficients in the window ∂s. For the special case of a Haar wavelet basis, this complexity reduces to $KN^{(l)} \log N^{(l)}$. The details can be found in Appendix B.

Note, that the concept of optimizing the Gaussian wavelet graph model in the space domain is general and not limited to the ICD algorithm. Specifically, given (23) and (26), other optimization methods such as preconditioned conjugate gradient (PCG)[44] can be used instead of ICD. In our experimentation, however, we have found that PCG with preconditioner as in [44] is not well suited for use with the multiresolution algorithm in Fig. 3 due to poor convergence of the PCG algorithm for non-constant initialization.

2.4 Nonlinear Classifiers for Parameter Selection

The nonlinear operators $T^{(n)}[\cdot]$ used for the parameter selection (18) are obtained using a method derived from recent work in image interpolation[39]. The selection of $\theta_s^{(n)}$ is performed by first classifying the input vector $x_{\partial s}^{(n)}/\sigma_n$ into a class $k_{n,s}$ and then selecting $\theta_s^{(n)}$ as a parameter vector associated with this class. To simplify the notation, we denote the parameter vector associated with class $k_{n,s}$ as $\theta_{k_{n,s}}^{(n)} = [A_{k_{n,s}}^{(n)}, b_{k_{n,s}}^{(n)}, B_{k_{n,s}}^{(n)}]$. Let $\mathcal{K}^{(n)}[\cdot]$ denote the classifier at scale n. The classification and parameter assignment can then be written as

$$k_{n,s} \leftarrow \mathcal{K}^{(n)}\left[\frac{\hat{x}_{\partial s}^{(n)}}{\sigma_n}\right] \tag{46}$$

$$\theta_s^{(n)} \leftarrow \theta_{k_{n,s}}^{(n)} \ . \tag{47}$$

The classifiers $\mathcal{K}^{(n)}[\cdot]$ and parameters $\theta_{k_{n,s}}^{(n)}$ are obtained during a training phase. For our implementation, we use a tree-based agglomerative clustering method which is described in detail in Appendix C. To summarize the approach, we initially partition the space of training samples $\{x_{\partial s}^{(n)}/\sigma_n, z_s^{(n)}/\sigma_n\}$ by performing a vector quantization (VQ) on $\{x_{\partial s}^{(n)}/\sigma_n\}$. For each cluster k, we then calculate the filters $[A_k^{(n)}, b_k^{(n)}]$ as the minimum mean square error linear predictors for $z_s^{(n)}$. Starting with this initial partitioning, we then form a cluster tree by merging pairs of clusters in a greedy fashion. At any given stage, we combine the two clusters whose merging results in the smallest increase in prediction error on the training set. Thus we form a binary tree where each node is associated with its optimal linear prediction filter for the conditional mean. To not overfit the classification model, we perform optimal tree pruning[45, 46] using a second data set for cross-validation.

The matrices $B_k^{(n)}$ are computed as

$$B_k^{(n)} = \frac{1}{w^2} R_k^{-1} \tag{48}$$

where R_k is the conditional sample covariance for class k (see Appendix), and w is a regularization parameter. The effect of w is similar to that of the scale parameter of a Gaussian Markov random field (GMRF) prior. Smaller values of w imply stronger regularization resulting in smoother images whereas larger values will result in less regularized, noisier images. In practical applications, w can be adjusted experimentally.

We have found it to be advantageous to constrain the classifiers $\mathcal{K}^{(n)}[\cdot]$ and the linear model predictors $A_s^{(n)} x_{\partial s}^{(n)}$ to only depend on $x_{\partial s}^{(n)} - x_s^{(n)}$, that is the difference between $x_{\partial s}^{(n)}$ and its center scaling coefficient. This constraint makes the model invariant to additive shifts in the gray value of the image and therefore improves robustness of training on smaller training sets. The details are listed in Appendix C.

3 Tomographic Reconstruction

In this section, we discuss the application of the proposed model to Bayesian tomographic reconstruction. First, we present the statistical data models $\log p_{y|x}(y|x)$ for both emission and transmission tomography using the exact Poisson counting statistics[42]. Let y denote the vector of photon counts for all M projections at different angles and displacements. Furthermore, let P be the tomographic projection matrix so that P_{i*} denotes the vector formed by its i^{th} row. For transmission tomography, the log-likelihood $\log p_{y|x}(y|x)$ may then be written as

$$\log p_{y|x}(y|x) = \sum_{i=1}^{M} \left(-y_T e^{-P_{i*}x} + y_i (\log y_T - P_{i*}x) - \log(y_i!) \right) \qquad \text{(transmission)} \tag{49}$$

where y_T denotes the dosage. For emission tomography, the log-likelihood is given by

$$\log p_{y|x}(y|x) = \sum_{i=1}^{M} (- P_{i*}x \ + y_i \log(P_{i*}x) - \log(y_i!)) \qquad \text{(emission)} \ . \tag{50}$$

Both (49) and (50) have the common form

$$\log p_{y|x}(y|x) = -\sum_{i=1}^{M} f_i(P_{i*}x) \tag{51}$$

where the $f_i(\cdot)$ are convex and differentiable.

Based on the model above, we can compute the coarse scale data models $\log p_{y|x^{(l)}}(y|x^{(l)})$ for $l > 0$. Combining (15) and (51), we obtain

$$\log p_{y|x^{(l)}}(y|x^{(l)}) \quad = \quad -\sum_{i=1}^{M} f_i(P_{i*}\mathcal{I}_l^0 x^{(l)}) \tag{52}$$

$$= \quad -\sum_{i=1}^{M} f_i(P_{i*}^{(l)} x^{(l)}) \tag{53}$$

where $P_{i*}^{(l)} = P_{i*}\mathcal{I}_l^0$. Thus, the coarse scale data models are equivalent to a standard model of the form (51) with a projection matrix $P^{(l)}$ whose columns are linear combinations of the columns of P.

To derive the MAP optimization for the tomographic data model, we write the equations for the emission case only, however, all methods analogously apply to the transmission case. Since the form of $\log p_{y|x^{(l)}}(y|x^{(l)})$ is the same for any l, we simplify the notation by omitting the superscripts (l). To implement the optimization of section 2.3 we need to solve

$$\Delta = \arg\max_{\gamma \geq -x_i} \left\{ \log p_{y|x}(y|x + \gamma e_i) + \alpha_1 \gamma + \frac{1}{2}\alpha_2 \gamma^2 \right\} \tag{54}$$

where the constraint $\gamma \geq -x_i$ enforces positivity in the space domain. The basic concept of the ICD algorithm[41, 42] is to solve (54) using a Newton-Raphson strategy. Importantly, the algorithm exploits the sparse nature of the projection matrix P by maintaining a state vector $\tilde{p} = Px$ of the current forward projection of x. Given $\tilde{p}$, we can write the first and second derivatives ψ_1 and ψ_2 of $\log p_{y|x}(y|x)$ with respect to the pixel value x_i as

$$\psi_1 \quad = \quad \frac{\partial}{\partial x_i} \log p_{y|x}(y|x) = -\sum_{\{k:P_{ki}>0\}} P_{ki} \left(1 - \frac{y_k}{\tilde{p}_k}\right) \tag{55}$$

$$\psi_2 \quad = \quad \frac{\partial^2}{\partial x_i^2} \log p_{y|x}(y|x) = -\sum_{\{k:P_{ki}>0\}} y_k \left(\frac{P_{ki}}{\tilde{p}_k}\right)^2 \ . \tag{56}$$

Based on the notation above, the second order approximation to $\log p_{y|x}(y|x + \gamma e_i)$ with respect to γ is

$$\log p_{y|x}(y|x + \gamma e_i) \approx \log p_{y|x}(y|x) + \psi_1 \gamma + \frac{1}{2}\psi_2 \gamma^2 \ . \tag{57}$$

Thus, the constrained Newton-Raphson update of (54) is given by

$$\Delta = \max\left\{ -x_i, -\frac{\psi_1 + \alpha_1}{\psi_2 + \alpha_2} \right\} \ . \tag{58}$$

The state vector $\tilde{p}$ can be updated efficiently using $\tilde{p_k}' = P_{ki}\Delta + \tilde{p_k}$ for $\{k : P_{ki} \neq 0\}$.

4 Experimental Results

In this section, we compare the proposed algorithm to two fixed-resolution Bayesian methods using Markov random field prior models and to convolution backprojection (CBP). The comparison is based on two sets of simulated emission tomography data; the first is a bar-phantom used for quantitative comparison of the algorithms in terms of reconstruction bias and noise variance, and the second is a more realistic case used for comparison of visual reconstruction quality, mean square error and computational efficiency.

The wavelet graph model (WGM) was implemented using an orthonormal 2-D Haar wavelet decomposition with $L = 5$ resolution levels. Two different training sets were used. For each set, the training samples were obtained by computing the wavelet decomposition of the training images. The number of reconstruction iterations was a fixed function of scale such that $K_0 * (\sqrt{2})^l$ iterations were performed for the reconstruction at scale l. Note, that the convergence speed could potentially be improved by using a stopping criterion based on the change in successive values of the log-posterior (16) at each scale. However, since for MAP reconstruction in general, very small changes of the log-posterior can result in large visual differences, we prefer to run a fixed number of iterations at each scale. For the evaluation of reconstruction quality, the number of iterations was set to be very large to insure complete convergence at each scale. The convergence behavior as a function of K_0 is shown in section 4.3.

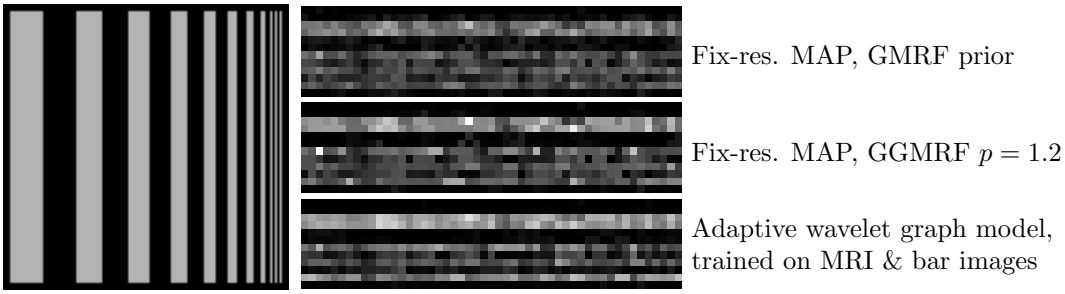
To demonstrate how the wavelet graph structure by itself can reduce the blockiness commonly encountered with Haar wavelet models, we also implemented a spatially homogeneous linear version of the wavelet graph model. The linear model uses no adaptation and only a single parameter vector at each scale such that $\theta_s^{(n)} = f(n)$.

The two fixed resolution MAP reconstruction algorithms were based on a Gaussian Markov random field (GMRF) and a generalized Gaussian Markov random field (GGMRF) prior model respectively. The GGMRF[10] is an edge-preserving, spatially homogeneous MRF that uses a non-quadratic penalty term. For the results shown here, the generalized Gaussian parameter was set to $p = 1.2$. The algorithms used ICD optimization with a large, fixed number of iterations to insure complete convergence. The CBP algorithm was implemented using a ramp filter and a generalized Hamming filter with frequency response $H(\omega) = H_{id}(\omega)\,(0.5 + 0.5\cos(\pi\omega/\omega_c))$ for $|\omega| \leq \omega_c$ where $H_{id}(\omega)$ denotes the ideal ramp filter.

4.1 Bar Phantom Results

Simulated emission data were generated using the bar phantom shown in Fig. 9(a) of size 115×115 pixels with values of 1.0 for the bars and 0.02 in the background. The image was embedded into a zero background of size 256×256 pixels which was forward projected at 128 angles and 256 displacements. The projection beam was assumed to be an ideal line. The data samples were formed by Poisson random variables with the appropriate means. The average number of counts per projection was 83.

Two different training sets were used for the proposed algorithm; the first set consisted of 40 MRI images of size 256×256 pixels and was intended to capture typical characteristics of medical images; the second set added 3000 amplitude and rotational variations of a bar phantom to demonstrate how the proposed method can be adapted when *a priori* knowledge about special characteristics of the phantoms



(a) Bar phantom (b) Magnified reconstruction of high frequency region (rotated)

Figure 9: Bar phantom and magnified high frequency region of sample reconstructions.

is available. The phantom shown in Fig. 9(a) was not included in the training. Because of the discrete nature of the second training set, we increased the value of δ in (70) of the Appendix to $\delta = 0.3$ to obtain reliable covariance parameters R_k. The window ∂s was set to 3×3 coefficients.

For quantitative comparison of the different algorithms, we calculated reconstruction bias and variance for the reconstructed image $\hat{x}_{ij}$ using the fact that the original phantom is constant along columns. We first calculated $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ as the mean and variance of the j^{th} column in the reconstruction $\hat{x}_{ij}$. Bias and variance were then computed as

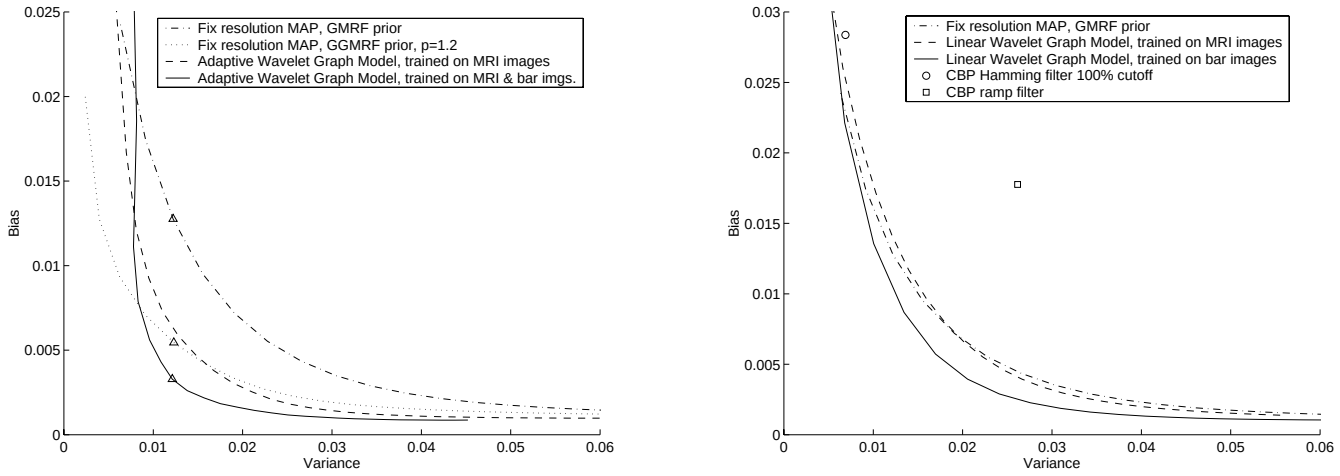
$$\text{bias} = \frac{1}{N}\sum_{j=1}^{N}(\hat{\mu}_j - \mu_j)^2 \tag{59}$$

$$\text{var} = \frac{1}{N}\sum_{j=1}^{N}\hat{\sigma}_j^2 \tag{60}$$

where μ_j is the value of the j^{th} column in the original phantom and $N = 115$ is the number of columns.

Bias and variance were computed for each reconstruction technique as a function of the regularization parameter. For the wavelet graph model, the parameter w in (48) was varied in the range $[0.1, 10]$, for the MRF priors the scale parameter σ was varied in the range $[0.1, 1.4]$, and for CBP reconstruction a ramp filter and a Hamming filter with cutoff $\omega_c = \pi$ were used. The proposed algorithm was initialized with a constant image and performed a fixed number of $25 * (\sqrt{2})^l$ reconstruction iterations at scale l. The fixed resolution ICD MAP algorithms were initialized with a CBP reconstruction and performed 40 iterations.

Figure 10(a) shows a comparison of the results for the adaptive wavelet graph model and the two fixed resolution Bayesian methods. Each plot corresponds to the bias/variance curve of a single reconstruction method as a function of the regularization parameter. The upper left hand corner of the graphs corresponds to strong regularization (high bias, low variance) whereas the lower right hand corner corresponds to weak regularization (low bias, high variance). The results indicate that the adaptive wavelet graph model trained on the set of MRI images performed significantly better than the GMRF method and performed



(a) Adaptive wavelet graph model (b) Linear models

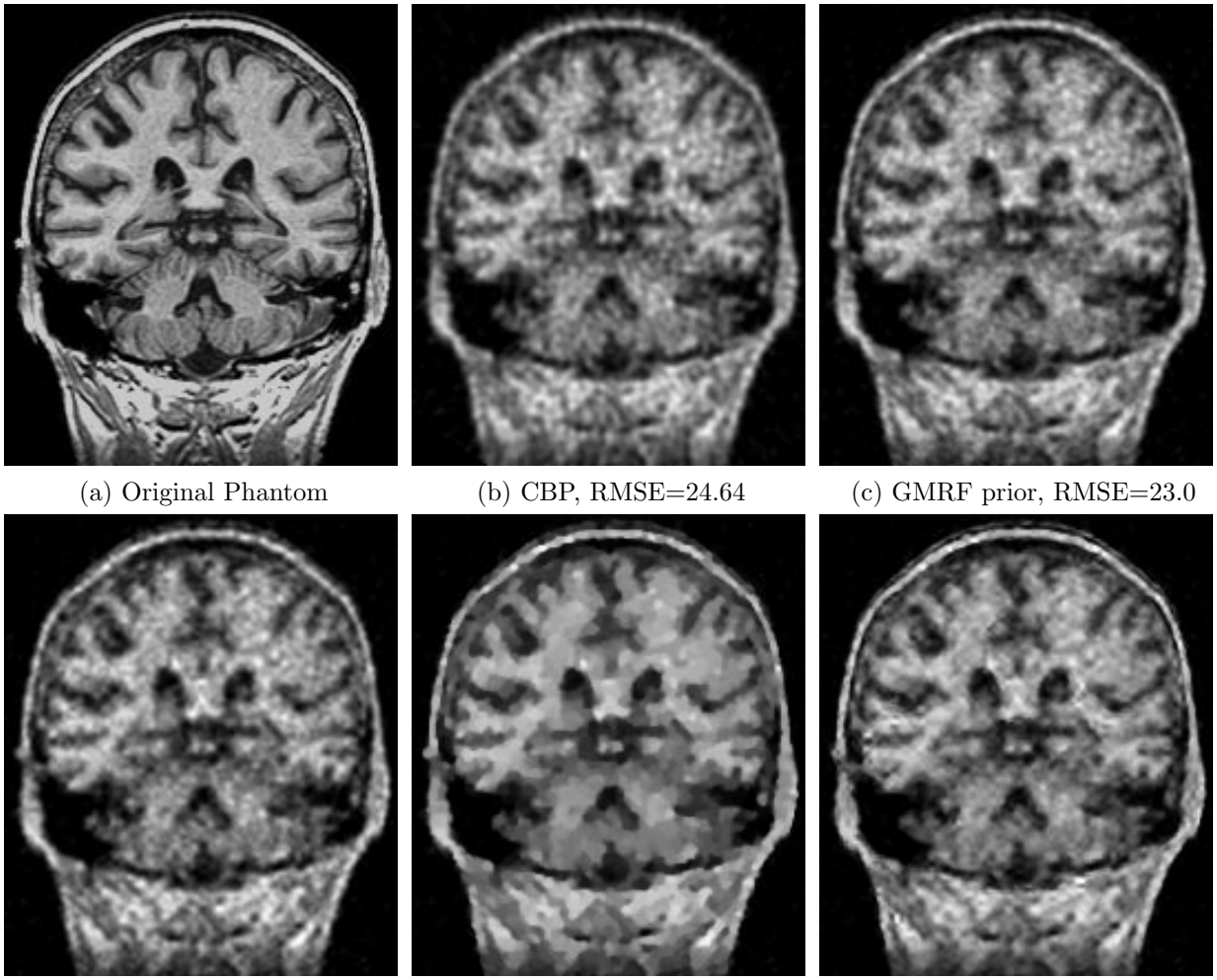
Figure 10: Bias-variance reconstruction performance of each algorithm as a function of regularization parameter. The triangles in (a) correspond to the bias-variance values for the sample reconstructions in Fig. 9(b).

comparably to the GGMRF based method. While the GGMRF performed best in the low variance region, its residual bias in the high variance region was slightly higher, introducing bias even when very little regularization is applied. When trained on the combination of bar phantoms and MRI images, the adaptive wavelet graph model outperformed both fixed resolution Bayesian methods. The smooth shape of the curve for the adaptive wavelet graph model indicates that the reconstructions are not overly sensitive to small variations of w, but rather, smoothly depend on the regularization parameter. Figure 9(b) shows the magnified high frequency region of sample reconstructions corresponding to the bias/variance points marked by triangles in Fig. 10(a). The adaptive wavelet graph model reconstruction better resolves the high frequency bars than the two fixed resolution Bayesian reconstructions.

Figure 10(b) shows the bias/variance curves for the linear, non-adaptive, wavelet graph model in comparison to fixed resolution ICD with a GMRF prior and to CBP. Trained on the set of MRI images, the linear wavelet graph model performed comparably to the GMRF method. When trained on the combination of bar phantoms and MRI images, the linear wavelet graph model achieved lower bias at equal variance as compared to the GMRF result.

4.2 Medical Image Phantom Results

Simulated emission data were generated from the magnetic resonance imaging (MRI) reconstruction image in Fig. 11(a). The 256×256 pixel image with mean 0.32 was forward projected at 128 angles and 256 displacements, assuming a field of view of 20cm and using Poisson noise. To better illustrate the deblurring



(a) Original Phantom (b) CBP, RMSE=24.64 (c) GMRF prior, RMSE=23.0

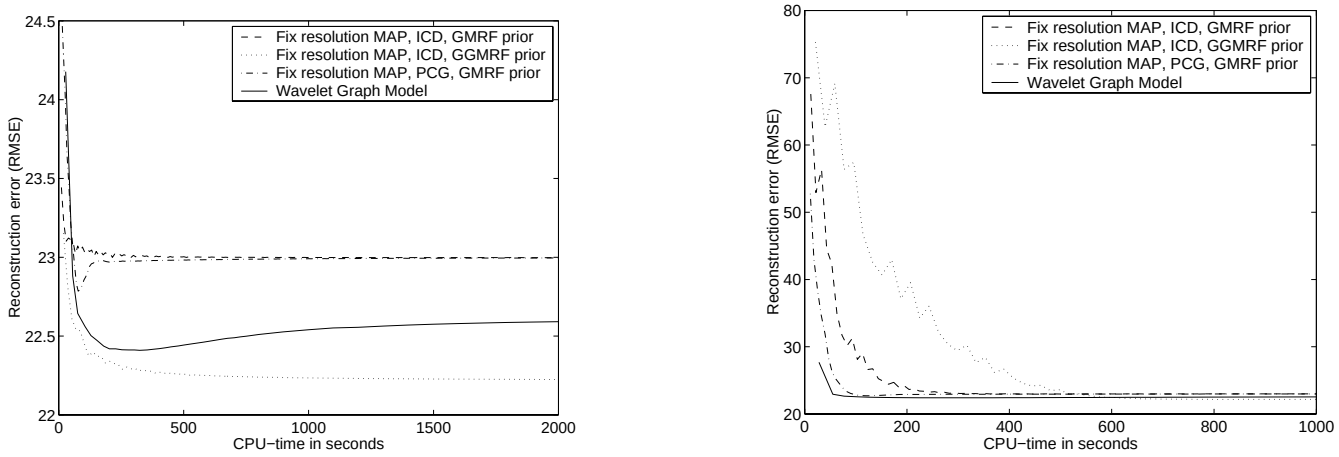
(d) Linear WGM, RMSE=23.59 (e) GGMRF prior, RMSE=22.21 (f) Adaptive WGM, RMSE=22.60

Figure 11: Original phantom (a) and reconstructions: (b) convolution backprojection, (c) fixed resolution MAP with GMRF prior, (d) proposed algorithm using linear wavelet graph model, (e) fixed resolution MAP with GGMRF prior, (f) proposed method using adaptive wavelet graph model.

potential of the algorithm, we assumed a projection beam with triangular profile of width 2.34mm which is three times the projection spacing. The average number of counts per projection was 235.

The adaptive and linear wavelet graph models were trained on the same set of 40 MRI images used for the bar phantom results. The phantom in Fig. 11(a) was not included in the training. The size of the window ∂s was set to 5 by 5 coefficients.

The regularization parameter for each reconstruction algorithm was adjusted manually to minimize reconstruction mean square error; this resulted in $\omega_c = 0.63\pi$ for the CBP Hamming filter, $\sigma = 0.2$ for the GMRF prior, $\sigma = 0.24$ for the GGMRF prior, $w = 1.075$ for the linear wavelet graph model, and $w = 1.0$ for the adaptive wavelet graph model. For each algorithm, a large number of iterations was used to insure complete convergence; specifically, $K_0 = 148$ for the proposed method with constant initialization and 500



(a) CBP initialization

(b) Constant initialization

Figure 12: Error convergence of the proposed algorithm compared to fixed resolution MAP reconstructions with GMRF and GGMRF prior models.

iterations for the fixed resolution ICD algorithms with CBP initialization.

The CBP reconstruction in Fig. 11(b) is noisy in the uniform image regions and contains blurry edges. The GMRF MAP reconstruction in Fig. 11(c) is less noisy than the CBP; however, the quadratic regularization function of the Gaussian MRF results in blurred edges. The linear wavelet graph model reconstruction (d) is slightly sharper than GMRF result (c) but contains some blocking artifacts and has higher RMSE. However, considering that this is a spatially homogeneous linear model using a Haar wavelet prior without shift-averaging, the reconstruction is surprisingly smooth, confirming the advantage of the wavelet graph structure. Fig. 11(e) shows the fixed resolution GGMRF MAP reconstruction. The result has sharper edges as compared to the GMRF case (c) and achieves the lowest mean square error of all four methods. Visually, however, the reconstruction (e) is of poor quality since it suffers from considerable loss of detail. The reconstruction using the adaptive wavelet graph model in Fig. 11(f) is superior to that of the other four methods. In comparison to the GMRF case (c), the reconstruction (f) contains sharper edges while the noise in the uniform regions is better suppressed. The mean square error is lower than for the GMRF case (c) but higher than for the GGMRF result (e). In comparison to (e), however, the proposed method preserves more detail.

4.3 Computational Efficiency Comparison

Reconstruction error convergence was compared for the proposed algorithm, fixed resolution ICD, and for a preconditioned conjugate gradient (PCG) MAP reconstruction algorithm for tomographic data which was developed in [44]. The PCG algorithm uses the GMRF prior model and was implemented with preconditioner and bent line search exactly as in [44] but did not use a factorization of the tomographic

projection matrix P which can potentially speed-up computation in practical applications[44].

Figure 12 shows the reconstruction error convergence for the different algorithms as a function of CPU-time on a 700 MHz Pentium III. The plots correspond to reconstructions of the data set in Fig. 11. The reconstruction error was calculated as the RMSE to the ground truth image. For ICD and PCG, the CPU-time was measured after each iteration; for the proposed method each data point corresponds to the total execution time of the algorithm for a certain K_0. Figure 12(a) shows the convergence results when all algorithms are initialized to the CBP reconstruction of Fig. 11(b). ICD and PCG converge very quickly, particularly for the GMRF prior. The proposed algorithm has slightly slower convergence. The RMSE curves for the proposed algorithm and for PCG reach a minimum followed by a slight increase for larger CPU-time. This is not surprising since the RMSE measure favors the slightly oversmoothed intermediate results over the visually superior converged reconstructions.

Figure 12(b) shows the convergence results when all reconstruction algorithms were initialized to a constant image. The constant was calculated from the projection data as $(\sum_i y_i)/(\sum_i \sum_j P_{i,j})$ to match the number of total measured counts. This initialization is typically used for PCG in practice. For this case, the fixed resolution ICD algorithm has slow convergence due to the slow low-frequency convergence of the ICD algorithm[41]. The proposed algorithm converges fastest, indicating that the multiresolution technique can provide a computational advantage in cases where an initialization with the correct low-frequency behavior is not available. Specifically, the multiresolution approach allows us to use an ICD optimization technique without requiring a CBP initialization. This is an advantage for systems with non-circular geometry[47] or limited angle problems where CBP initializations are not easily obtained.

Since the computation of the proposed algorithm is largely dominated by the forward model, the size of the window ∂s for the wavelet graph model has a limited effect on efficiency. For the data set in Fig. 11, a window of size 5×5 as compared to 3×3 results in a 35% increase in CPU-time for the same K_0.

5 Conclusions

We propose a wavelet graph prior model in conjunction with a multiresolution Bayesian reconstruction algorithm applicable to tomographic reconstruction. The wavelet graph prior model has a dependency structure that is more general than a quadtree. This enables the model to produce smooth estimates even for a Haar wavelet basis. Furthermore, the wavelet graph structure is such that the optimal model for a stationary process is homogeneous, resulting in a substantial reduction in the number of model parameters. The multiresolution reconstruction algorithm uses the wavelet graph prior model but performs a sequence of MAP optimizations in the space domain. The space domain formulation allows us to efficiently enforce

the pixel positivity constraint and to preserve the sparseness of the tomographic projection operator. Our experimental results indicate that the proposed framework can improve reconstruction quality over commonly used fixed resolution Bayesian methods.

6 Acknowledgments

We would like to thank Dr. Ned C. Rouze for his valuable comments.

Appendix A: Proof of Theorem

First note that $X_s^{(n)}$ is a circularly stationary random process for each resolution $0 \leq n \leq L$. This results from the assumption that $X_s^{(0)}$ is circularly stationary together with the form of the recursions (1) and (2). Next, notice that

$$\begin{bmatrix} X_{\partial_s}^{(n+1)} \\ Z_s^{(n+1)} \end{bmatrix} = Q \begin{bmatrix} X^{(n)}_{(-p+2(s-w))\bmod 2^N} \\ X^{(n)}_{(-p+2(s-w)+1)\bmod 2^N} \\ \vdots \\ X^{(n)}_{(p+2(s+w))\bmod 2^N} \end{bmatrix}$$

where Q is a fixed matrix, where p is such that the support of the kernels h and g in (1) and (2) is within $[-p, p]$ and the centered window ∂s is of length $2w+1$. Based on the equation above and the fact that $X_s^{(n)}$ is a circularly stationary process, the distribution of $[X_{\partial_s}^{(n+1)}, Z_s^{(n+1)}]$ does not depend on s. Using the assumption that $\nu^* = \arg\max_\nu E[\log p_z(Z)]$ exists and is unique, we write

$$\begin{aligned}
\max_\nu E[\log p_z(Z)] &= \max_\nu E\left[\sum_{(s,n)\in S} \log p_{z_s^{(n)}|x_{\partial s}^{(n)}}(Z_s^{(n)}|X_{\partial s}^{(n)}) + \text{const}\right] \\
&= \max_\nu \sum_{(s,n)\in S} E\left[\log p_{z_s^{(n)}|x_{\partial s}^{(n)}}(Z_s^{(n)}|X_{\partial s}^{(n)})\right] + \text{const} \\
&= \sum_{(s,n)\in S} \max_{\nu_s^{(n)}} E\left[\log p_{z_s^{(n)}|x_{\partial s}^{(n)}}(Z_s^{(n)}|X_{\partial s}^{(n)})\right] + \text{const} \\
&= \sum_{(s,n)\in S} \max_{\nu_s^{(n)}} E[f_n(Z_s^{(n)}, X_{\partial s}^{(n)}, \nu_s^{(n)})] + \text{const}\ .
\end{aligned}$$

Since neither the f_n nor the distribution of $[X_{\partial_s}^{(n+1)}, Z_s^{(n+1)}]$ depend on s,

$$(\nu_s^{(n)})^* = \arg\max_{\nu_s^{(n)}} E[f_n(Z_s^{(n)}, X_{\partial s}^{(n)}, \nu_s^{(n)})]$$

is not a function of s. This proves the theorem.

Appendix B: Computational Complexity

The number of multiplications for optimizing the wavelet graph model with respect to a single scaling coefficient $x_i^{(l)}$ is proportional to the size of the set $\tilde{S}_i \stackrel{\triangle}{=} \bigcup_{n>l} \tilde{S}_i^{(n)}$, where the $\tilde{S}_i^{(n)}$ are as defined in (40). For a general wavelet transform, the size $|S_i^{(n)}|$ of the sets $S_i^{(n)}$ as defined in (39) is proportional to $n-l$. Define $S_i \stackrel{\triangle}{=} \bigcup_{n>l} S_i^{(n)}$, then

$$|S_i| = \sum_{n=l+1}^{L} |S_i^{(n)}| \quad \propto \quad \sum_{n=l+1}^{L} (n-l) \tag{61}$$

$$\propto \quad (L-l)^2 \leq (\log N^{(l)})^2 \ . \tag{62}$$

Based on (41), the size of the set $\tilde{S}_i$ is upper bounded by $K|S_i|$ where $K = |\partial s|$ denotes the number of coefficients in the fixed size window ∂s. Therefore, $|\tilde{S}_i| \propto K(\log N^{(l)})^2$ and consequently steps (42)-(45) can be executed in order $K(\log N^{(l)})^2$ multiplications. Thus, the complexity for a full update of $\hat{x}^{(l)}$ is order $KN^{(l)}(\log N^{(l)})^2$. For a Haar wavelet decomposition, this reduces to $KN^{(l)} \log N^{(l)}$ since $S_i^{(n)}$ contains only a single node at each scale n. The complexities for the 1-D and the 2-D case are the same.

Appendix C: Tree-structured Nonlinear Classifier

In this section, we describe the agglomerative clustering method used to obtain the classifiers $\mathcal{K}^{(n)}[\cdot]$ and the parameter vectors $\theta_k^{(n)}$ for class k. For the classifier at scale n, we assume a training set $\{x_{\partial s}^{(n)}/\sigma_n,\ z_s^{(n)}/\sigma_n\}$. We then define a normalized training set $\{v_s, w_s\}$ with samples $v_s = V(x_{\partial s}^{(n)} - x_s^{(n)})/\sigma_n$ and $w_s = z_s^{(n)}/\sigma_n$ where V is the matrix that eliminates the zero center component of $x_{\partial s}^{(n)} - x_s^{(n)}$ and is the identity otherwise. Thus, v_s has one fewer component than $x_{\partial s}^{(n)}$. Since the training is performed separately for each scale n, we simplify the notation by omitting the dependence on n. Our objective is to form a classification tree for v_s such that each tree node is associated with a MMSE linear predictor to predict w_s from v_s. We first perform a vector quantization (VQ) of $\{v_s\}$ with a pre-specified, fixed number of clusters K, currently $K = 150, 100, 100, 50, 50$ for scales $n = 0$ through $n = 4$. The distance metric for the VQ is the Euclidean distance. The number of iterations M for the VQ is set to a constant value, currently $M = 5$. The next step is to compute the MMSE linear predictors for the VQ clusters. Let $c_1, \ldots, c_K$ denote the K clusters. Define $\mu_{c_k,v}$ and $\mu_{c_k,w}$ as the means and $C_{c_k,vv}$ and $C_{c_k,ww}$ as the covariance matrices of the samples in cluster c_k with respect to v and w. Let $C_{c_k,vw}$ denote the cross-covariance of the samples in cluster c_k. We then compute a MMSE linear predictor $[A_{c_k}, b_{c_k}]$ for cluster c_k as

$$\tilde{A}_{c_k} \quad = \quad C_{c_k,vw}^t C_{c_k,vv}^{-1} \tag{63}$$

$$b_{c_k} \quad = \quad \mu_{c_k,w} - \tilde{A}_{c_k} \mu_{c_k,v} \tag{64}$$

$$A_{c_k} \;=\; \tilde{A}_{c_k} V \;. \tag{65}$$

Further, the total prediction error ε_{c_k} of cluster c_k over the training set is obtained as

$$\varepsilon_{c_k} = \mathrm{trace}(C_{c_k,ww} - 2\tilde{A}_{c_k} C_{c_k,vw} + \tilde{A}_{c_k} C_{c_k,vv} \tilde{A}_{c_k}^t) \;. \tag{66}$$

Now consider merging two clusters c_k and c_l into a new cluster $c_m = c_k \cup c_l$ with MMSE predictor $[A_{c_m}, b_{c_m}]$ and prediction error ε_{c_m}. The total increase in prediction error due to the merging is $\Delta\varepsilon_{c_k,c_l} = \varepsilon_{c_m} - (\varepsilon_{c_k} + \varepsilon_{c_l})$. Merging clusters based on minimum increase in prediction error is not sensible for small clusters whose linear predictor may be over-parameterized such that $\varepsilon_{c_k} = 0$. To merge small clusters in a meaningful way, we introduce a regularization term ρ_{c_k,c_l} based on cluster distance

$$\rho_{c_k,c_l} = \alpha(|c_k||\mu_{c_k,v} - \mu_{c_m,v}|^2 + |c_l||\mu_{c_l,v} - \mu_{c_m,v}|^2) \tag{67}$$

where α is a small regularization constant and $|c_k|$ denotes the number of samples in cluster k. We then define the cost function M_{c_k,c_l} for the merging of c_k and c_l as

$$M_{c_k,c_l} = \Delta\varepsilon_{c_k,c_l} + \rho_{c_k,c_l} \;. \tag{68}$$

Starting with the initial VQ partitioning, we successively combine the two clusters c_k, c_l whose merging results in the smallest M_{c_k,c_l}. This results in a binary tree where each node is associated with its optimal linear prediction filter for w_s. The leaves of the tree are the VQ clusters. To not overfit the classification model, we perform optimal tree pruning[45, 46] using a second data set for cross-validation. The pruning set $\{\tilde{v}_s, \tilde{w}_s\}$ is classified into the tree by assigning each data sample to the closest VQ cluster and to all of its parents in the tree. The prediction error for the pruning samples in each node is computed using the node filters computed on the training set. The tree is then pruned in a bottom-up fashion by considering all nodes at each level before moving up by one level. If the prediction error for a node filter is lower than the combined errors in the leaves of the subtree originating at this node, the model is considered overparameterized and the subtree is effectively removed by marking its nodes as pruned.

The covariance parameter R_k for each tree node is computed as a linear combination of the conditional covariance of the pruning samples in class k and the expected conditional covariance over the entire pruning set. Let $R_{1,k}$ be the conditional covariance matrix of the pruning samples in class k

$$R_{1,k} = \frac{1}{|c_k|} \sum_{\{s:\tilde{v}_s \in c_k\}} (\tilde{w}_s - A_{c_k}\tilde{v}_s - b_{c_k})(\tilde{w}_s - A_{c_k}\tilde{v}_s - b_{c_k})^t \tag{69}$$

where $|c_k|$ is the number of pruning samples that fall into node k. Further, let use define R_2 as the expected conditional covariance over the entire pruning set $R_2 = \frac{1}{N_p}\sum_{\{k:c_k \in \mathcal{L}\}} |c_k| R_{1,k}$ where $\mathcal{L}$ is the set

of tree leaves after discarding the pruned nodes and N_p is the total number of samples in the pruning set. We then compute the covariance parameter R_k for node k as

$$R_k = (1 - \delta)R_{1,k} + \delta R_2 \tag{70}$$

where δ is a small constant. The term δR_2 is added to impose a lower limit on the R_k. For our experimentation, we use a fixed value $\delta = 0.001$ except for one case where the training set of discrete images requires a larger value of $\delta = 0.3$.

In order to perform the classification (46) at runtime, we first find the VQ cluster with minimum Euclidean distance to $V(\hat{x}_{\partial s}^{(n)} - \hat{x}_s^{(n)})/\sigma_n$. We then follow the tree upwards until we reach the first node that is not marked pruned. This node corresponds to the class $k_{n,s}$ and contains the associated parameter vector $\theta_{k_{n,s}}^{(n)}$.

References

[1] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741, November 1984.

[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, 48(3):259–302, 1986.

[3] Rama Chellappa and Anil Jain, editors. *Markov Random Fields: Theory and Applications*. Academic Press, Inc., Boston, 1993.

[4] E. Levitan and G. Herman. A maximum a *posteriori* probability expectation maximization algorithm for image reconstruction in emission tomography. *IEEE Trans. on Medical Imaging*, MI-6:185–192, September 1987.

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

[6] T. Hebert and R. Leahy. A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans. on Medical Imaging*, 8(2):194–202, June 1989.

[7] Peter J. Green. Bayesian reconstruction from emission tomography data using a modified EM algorithm. *IEEE Trans. on Medical Imaging*, 9(1):84–93, March 1990.

[8] R. Stevenson and E. Delp. Fitting curves with discontinuities. *Proc. of the first international workshop on robust computer vision*, pages 127–136, October 1-3 1990.

[9] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(3):367–383, March 1992.

[10] C. A. Bouman and K. Sauer. A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. on Image Processing*, 2(3):296–310, July 1993.

[11] S. Brette, J. Idier, and Ali Mohammad-Djafari. Scale invariant Markov models for Bayesian inversion of linear inverse problems. In J. Skilling and S. Sibisi, editors, *Maximum Entropy and Bayesian Methods*, pages 199–212. Kluwer Academic Publishers, Netherlands, 1996.

[12] M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R. Nikoukhah, and A. S. Willsky. Modeling and estimation of multiresolution stochastic processes. *IEEE Trans. on Information Theory*, 38(2):766–784, March 1992.

[13] M. Basseville, A. Benveniste, and A. Willsky. Multiscale autoregressive processes, part i: Schur-Levinson parametrizations. *IEEE Trans. on Signal Processing*, 40(8):1915–1934, August 1992.

[14] M. Basseville, A. Benveniste, and A. Willsky. Multiscale autoregressive processes, part ii: Lattice structures for whitening and modeling. *IEEE Trans. on Signal Processing*, 40(8):1935–1954, August 1992.

[15] K. C. Chou, S. Golden, and A. S. Willsky. Multiresolution stochastic models, data fusion and wavelet transforms. *Signal Processing*, 34(3):257–282, December 1993.

[16] R. D. Nowak. Multiscale hidden Markov models for Bayesian image analysis. Technical Report 004-98, Michigan State University, Ann Arbor, MI, 1998.

[17] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. on Signal Processing*, 46(4):886–902, April 1998.

[18] K. Chou, A. Willsky, A. Benveniste, and M. Basseville. Recursive and iterative estimation algorithms for multiresolution stochastic processes. In *Proceedings of the 28^{t}h Conference on Decision and Control*, volume 2, pages 1184–1189, Tampa, Florida, December 13-15 1989.

[19] Bernhard Claus and Ghislaine Chartier. Multiscale signal processing: Isotropic random fields on homogeneous trees. *IEEE Trans. on Circuits and Systems:Analog and Digital Signal Proc.*, 41(8):506–517, August 1994.

[20] Mark R. Luettgen, W. Clem Karl, and Alan S. Willsky. Efficient multiscale regularization with applications to the computation of optical flow. *IEEE Trans. on Image Processing*, 3(1):41–64, January 1994.

[21] P. W. Fieguth, W. C. Karl, and A. S. Willsky. Efficient multiresolution counterparts to variational methods for surface reconstruction. *Computer Vision and Image Understanding*, 70(2):157–176, May 1998.

[22] A. B. Frakt and A. S. Willsky. Efficient multiscale stochastic realization. In *Proc. of IEEE Int'l Conf. on Acoust., Speech and Sig. Proc.*, volume 4, pages 2249–2252, Seattle, WA, May 12-15 1998.

[23] Khalid Daoudi, Austin B. Frakt, and Alan S. Willsky. Multiscale autoregressive models and wavelets. *IEEE Trans. on Information Theory*, 45(3):828–845, April 1999.

[24] H-C. Yang and R. Wilson. Adaptive image restoration using a multiresolution Hopfield neural network. In *Fifth International conference on Image Processing and its applications, (IEE Conference Publication No.410)*, pages 198–202, Edinburgh, UK, July 4-6 1995.

[25] Robert D. Nowak and Eric D. Kolaczyk. A multiscale MAP estimation method for Poisson inverse problems. In *Proceedings of the 32nd Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1682–1686, Pacific Grove, CA, November 1-4 1998.

[26] Robert D. Nowak and Eric D. Kolaczyk. A statistical multiscale framework for Poisson inverse problems. *IEEE Trans. on Information Theory*, 46(5):1811–1825, August 2000. Special Issue on Information-Theoretic Imaging.

[27] Zoltan Kato, Marc Berthod, and Josiane Zerubia. Multiscale Markov random field models for parallel image classification. In *Proceedings International Conference on Computer Vision*, pages 253–237, Berlin, Germany, May 1993.

[28] C. A. Bouman and M. Shapiro. A multiscale random field model for Bayesian image segmentation. *IEEE Trans. on Image Processing*, 3(2):162–177, March 1994.

[29] M. L. Comer and E. J. Delp. Segmentation of textured images using a multiresolution Gaussian autoregressive model. *IEEE Trans. on Image Processing*, 8(3):408–420, March 1999.

[30] J-M. Laferte, P. Perez, and F. Heitz. Discrete Markov image modeling and inference on the quadtree. *IEEE Trans. on Image Processing*, 9(3):390–404, March 2000.

[31] Zhenyu Wu, Gabor T. Herman, and Jolyon A. Browne. Edge preserving reconstruction using adaptive smoothing in wavelet domain. In *Proc. of IEEE Nucl. Sci. Symp. and Med. Imaging Conf.*, volume 3, pages 1917–1921, San Francisco, California, October 31 - November 6 1993.

[32] Suhail S. Saquib, Charles A. Bouman, and Ken Sauer. A non-homogeneous MRF model for multiresolution Bayesian estimation. In *Proc. of IEEE Int'l Conf. on Image Proc.*, volume 2, pages 445–448, Lausanne Switzerland, September 16-19 1996.

[33] P. W. Fieguth, W. C. Karl, W. S. Willsky, and C. Wunsch. Multiresolution optimal interpolation and statistical analysis of TOPEX/POSEIDON satellite altimetry. *IEEE Trans. on Geoscience and Remote Sensing*, 33(2):280–292, March 1995.

[34] J.-M. Laferte, F. Heitz, P. Perez, and E. Fabre. Hierarchical statistical models for the fusion of multiresolution image data. In *Proc. of Int'l Conference on Computer Vision*, pages 908–913, Cambridge, MA, June 20-23 1995.

[35] William W. Irving, Paul W. Fieguth, and Alan S. Willsky. An overlapping tree approach to multiscale stochastic modeling and estimation. *IEEE Trans. on Image Processing*, 6(11):1517–1529, June 1997.

[36] R. D. Nowak. Shift invariant wavelet-based statistical models and 1/f processes. In *IEEE DSP Workshop*, 1998.

[37] R. Coifman and D. Donoho. Translation invariant de-noising. In *Lecture Notes in Statistics: Wavelets and Statistics*, pages 125–150. Springer Verlag, New York, 1995.

[38] Zoltan Kato, Marc Berthod, and Josiane Zerubia. A hierarchical Markov random field model and multitemperature annealing for parallel image classification. *Graphical Models and Image Processing*, 58(1):18–37, January 1996.

[39] C. B. Atkins, C. A. Bouman, and J. P. Allebach. Tree-based resolution synthesis. In *Proc. of the Image Proc., Image Quality, Image Capture Systems Conference (PICS '99)*, pages 405–410, Savannah, GA, April 25-28 1999.

[40] E. L. Lehmann. *Theory of Point Estimation*. John Wiley & Sons, New York, 1983.

[41] K. Sauer and C. A. Bouman. A local update strategy for iterative reconstruction from projections. *IEEE Trans. on Signal Processing*, 41(2):534–548, February 1993.

[42] C. A. Bouman and K. Sauer. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Trans. on Image Processing*, 5(3):480–492, March 1996.

[43] J. Zheng, S. Saquib, K. Sauer, and C. A. Bouman. Parallelizable Bayesian tomography algorithms with rapid, guaranteed convergence. *IEEE Trans. on Image Processing*, 9(10):1745–1759, October 2000.

[44] J. Qi, R. M. Leahy, Chinghan Hsu, T. H. Farquhar, and S. R. Cherry. Fully 3D Bayesian image reconstruction for the ECAT EXACT HR+. *IEEE Trans. on Nuclear Science*, 45(3):1096–1103, June 1998.

[45] S. B. Gelfand, C. S. Ravishankar, and E. J. Delp. An iterative growing and pruning algorithm for classification tree design. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(2):163–174, February 1991.

[46] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

[47] N. C. Rouze, W. Winkle, and G. D. Hutchins. IndyPET - a high resolution, high sensitivity dedicated research scanner. In *Proc. of IEEE Nucl. Sci. Symp. and Med. Imaging Conf.*, volume 3, pages 1460–1464, Seattle, WA, October 24-30 1999.