IMAGE DECONVOLUTION USING TREE-STRUCTURED BAYESIAN GROUP SPARSE MODELING

Ganchi Zhang, Timothy D. Roberts, Nick Kingsbury

Signal Processing Group, Depart. of Engineering, University of Cambridge, UK

ABSTRACT

In this paper, we propose to incorporate wavelet tree structures into a recently developed wavelet modeling method, called VBMM. We show that, using overlapped groups, tree-structured modeling can be integrated into the high-performance non-convex sparsity-inducing VBMM method, and can achieve significant performance gains over the coefficient-sparse version of the algorithm.

Index Terms— Image deconvolution, wavelet tree modeling, variational Bayesian, dual-tree complex wavelets.

1. INTRODUCTION

Image deconvolution appears in many applications of image processing. The object is to estimate the clean image $\mathbf{x}$ from a blurred image $\mathbf{y}$ usually based on a linear observation model:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{H}$ is a $M \times M$ matrix which approximates the convolution, and $\mathbf{n}$ is additive Gaussian noise with variance ν^2. In general, this inverse problem is highly ill-posed, i.e., the direct operator does not have an inverse or it is nearly singular so that its inverse is very sensitive to noise [1]. In previous works, it is found that wavelet-based tools, such as the Discrete Wavelet Transform (DWT), are powerful for handling this ill-posed nature [2, 3, 4]. Most of them are based on regularization or Bayesian frameworks, which largely rely on the sparsity assumptions of wavelet-based priors/regularizers due to the fact that natural images can be represented by relatively few coefficients in the wavelet domain [2]. In general, wavelet coefficients can often be modeled by heavy-tailed priors belonging to the Gaussian scale mixture (GSM) class that captures the local dependencies among different wavelet coefficients [3, 5].

It is also well-established that there is a strong persistence of large/small wavelet coefficients across scales [6, 7]. Such patterns can be well represented using a tree structure where parent-child coefficients at a certain location and adjacent scales are both large or small [6]. Fig. 1 depicts an example of quadtree structure that corresponds to an 8×8 image with 3-level 2D DWT decomposition. There are many methods to model this wavelet tree structure such as bivariate shrinkage [8], Hidden Markov Tree (HMT) [9, 10] and overlapping-group penalty [6, 11, 12]. By integrating such a tree-approximation, it has been shown to significantly improve the recovery performance [13].

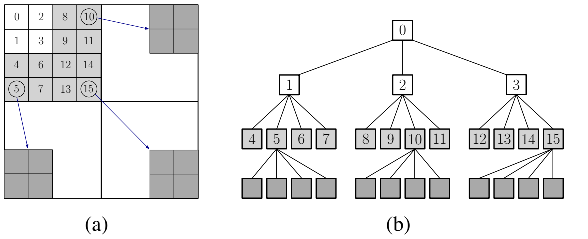

Fig. 1. (a) 8×8 image with 3-level 2D DWT decomposition. (b) quadtree structure of wavelet coefficients.

This paper builds on a hierarchical Bayesian modeling of wavelet coefficients proposed in [14], which is derived from a group-sparse GSM model. Based on a combination of variational Bayesian (VB) inference with a subband-adaptive Majorization Minimization (MM) method, the VBMM algorithm in [14] effectively simplifies computation of the posterior distribution and finds good solutions in the non-convex search space. In addition, VBMM has demonstrated the potential of group-sparse modeling. For instance, the real and imaginary parts of the dual-tree complex wavelet transform (DT $\mathbb{C}$WT) coefficients are clustered into single groups for Bayesian inference [14]. However, tree-structured dependencies among wavelet coefficients were not fully utilized in VBMM in [14].

To achieve the goal of a fully group sparse solution, in this paper we propose a new image deconvolution algorithm which incorporates the VBMM model with wavelet tree structure. The grouping strategies “parent+1child” and “parent+4children” are explored. The experimental results show that both strategies result in significantly improved performance compared with VBMM without an imposed group structure. One important contribution of the paper is to provide a new framework which incorporates a wavelet tree structure in an empirical Bayesian derivation.

The paper is organized as follows. Section 2 describes the key formulations of our model with grouping strategies. Section 3 shows the continuation strategy of the proposed algorithm. Experimental results are shown in Section 4.

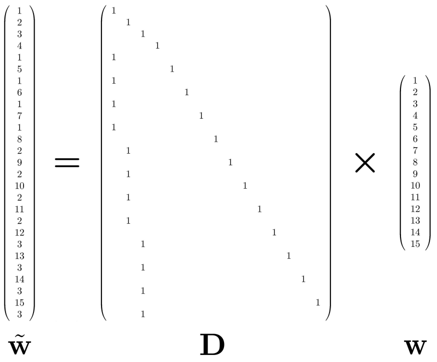

Fig. 2. Simple example of the non-overlapping transformation corresponding to the quadtree in Fig. 1(b), using the parent+1child grouping of Fig. 3(a).

2. MODEL FORMULATION

In this section, we describe the formulations of our tree-structured statistical model. Assume we can represent the image $\mathbf{x}$ by wavelet expansion as $\mathbf{x} = \mathbf{M}\mathbf{w}$ where $\mathbf{M}$ is the inverse wavelet transform, and $\mathbf{w}$ is an $N \times 1$ vector which contains all wavelet coefficients. This results in a wavelet-based formulation of (1):

$$\mathbf{y} = \mathbf{H}\mathbf{M}\mathbf{w} + \mathbf{n} \tag{2}$$

It is noted that for an orthogonal basis, $\mathbf{M}$ is a square orthogonal matrix, whereas for an over-complete dictionary (e.g. a tight frame), $\mathbf{M}$ has N columns and M rows, with $N > M$ [4]. Recent works have shown that modeling wavelet parent-child relationships can be viewed as an overlapping group regularization [6, 11]. Inspired from [11], we adopt a non-overlapping redundant transformation such as $\tilde{\mathbf{w}} = \mathbf{D}\mathbf{w}$ to ensure the persistence of large/small coefficients across scales, where $\tilde{\mathbf{w}}$ is a $P \times 1$ vector that forms wavelet coefficients in the non-overlapping space, and the transformation matrix $\mathbf{D}$ indicates the presence (1) or absence (0) of correspondence between the overlapping and non-overlapping spaces. An example of this non-overlapping redundant transformation is shown in Fig. 2. Note that although $\mathbf{D}^T\mathbf{D} \neq \mathbf{I}$, $\mathbf{D}^T\mathbf{D}$ is a diagonal matrix, with each entry representing the number of groups to which a coefficient belongs. For instance, the entry in $\mathbf{D}^T\mathbf{D}$ corresponding to a parent coefficient with a "parent+1child" grouping scheme, such as coefficient 1 in Fig. 3 (a), will be '5' (one for each of 1's four children, plus one singleton group). As a result, we define a diagonal matrix $\mathbf{A}$ such that $\mathbf{D}^T\mathbf{D}\mathbf{A} = \mathbf{I}$, and the likelihood of the data can be shown to be

$$p(\mathbf{y}|\hat{\mathbf{w}},\nu^2) = \left(2\pi\nu^2\right)^{-\frac{M}{2}} \exp\{-\frac{1}{2\nu^2}\|\mathbf{y} - \mathbf{H}\mathbf{M}\mathbf{D}^T\hat{\mathbf{w}}\|^2\} \tag{3}$$

where $\hat{\mathbf{w}} = \mathbf{D}\mathbf{A}\mathbf{w}$. In this paper, we propose to model $\hat{\mathbf{w}}$ using a group sparse GSM model as described in [14]:

$$p\left(\hat{\mathbf{w}}|\mathbf{S}\right) = \prod_{i=1}^{G} \mathcal{N}\left(\hat{\mathbf{w}}_i|0,\sigma_i^2\right) = \mathcal{N}\left(\hat{\mathbf{w}}|0,\mathbf{S}^{-1}\right) \tag{4}$$

where the i^{th} group $\hat{\mathbf{w}}_i$ is a vector of size g_i whose elements are drawn from a zero-mean Gaussian distribution with a signal variance σ_i^2 (as yet unknown), and where G is the number of groups, and $\mathbf{S}$ is a diagonal matrix formed from the vector $\mathbf{s}$ whose i^{th} entry is $s_i = 1/\sigma_i^2$. Because $\mathbf{S}$ needs to be of size $P \times P$, when $P > G$, its diagonal is an expanded form of $\mathbf{s}$ where each s_i is repeated g_i times [14].

However, how best to group coefficients is not clear and becomes an important question. In this paper, we consider two grouping schemes: "parent+1child" and "parent+4children" as illustrated in Fig. 3. In the case of "parent+1child" scheme, the parent coefficient is grouped separately with each child coefficient, whereas for "parent+4children" scheme the parent coefficient is grouped with all 4 of its children. Note that in both cases, we group the root-level coefficients individually since they do not have a parent.

Based on Bayes' rule, the posterior distribution can be calculated via:

$$p\left(\hat{\mathbf{w}}|\mathbf{y},\mathbf{S},\nu^2\right) = \frac{p\left(\mathbf{y}|\hat{\mathbf{w}},\nu^2\right) \times p\left(\hat{\mathbf{w}}|\mathbf{S}\right)}{p\left(\mathbf{y}|\mathbf{S},\nu^2\right)} \tag{5}$$

Because both $p\left(\mathbf{y}|\hat{\mathbf{w}},\nu^2\right)$ and $p\left(\hat{\mathbf{w}}|\mathbf{S}\right)$ are Gaussian functions of $\hat{\mathbf{w}}$, the posterior can be rearranged as

$$p\left(\hat{\mathbf{w}}|\mathbf{y},\mathbf{S},\nu^2\right) = \mathcal{N}\left(\hat{\mathbf{w}}|\mu,\Sigma\right) \tag{6}$$

where

$$\mu = \nu^{-2}\Sigma\mathbf{D}\mathbf{M}^T\mathbf{H}^T\mathbf{y} \tag{7}$$

$$\Sigma = \left(\nu^{-2}\mathbf{D}\mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M}\mathbf{D}^T + \mathbf{S}\right)^{-1} \tag{8}$$

However, Σ requires the inversion of a $P \times P$ square matrix $\left(\nu^{-2}\mathbf{D}\mathbf{M}^T\mathbf{H}^T\mathbf{H}\mathbf{M}\mathbf{D}^T + \mathbf{S}\right)$, which is not computationally feasible for big data sets and images. To derive a fast algorithm, we adopt a subband-adaptive MM technique proposed in [15]. Here we introduce a hidden variable $\mathbf{z}$ and the following approximation model for its posterior distribution:

$$\overline{p}\left(\hat{\mathbf{w}},\mathbf{z}|\mathbf{y},\mathbf{S},\nu^2\right) = p\left(\mathbf{z}|\hat{\mathbf{w}}\right) \times p\left(\hat{\mathbf{w}}|\mathbf{y},\mathbf{S},\nu^2\right) \tag{9}$$

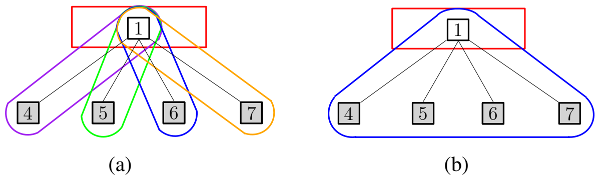

Fig. 3. Illustration of different grouping strategies: (a) parent+1child, (b) parent+4children. The root-level coefficients are grouped individually shown as the red rectangles which represent singleton groups.

and

$$p\left(\mathbf{z}|\hat{\mathbf{w}}\right) \propto \exp\{-(\hat{\mathbf{w}} - \mathbf{DAz})^T \mathbf{Q}(\hat{\mathbf{w}} - \mathbf{DAz})\} \tag{10}$$

where

$$\mathbf{Q} = \frac{\Lambda_\alpha - \mathbf{DM}^T\mathbf{H}^T\mathbf{HMD}^T}{2\nu^2} \tag{11}$$

Similar to the argument in [14], we find that $\mathbf{Q}$ should be positive definite to ensure convergence and hence:

$$\Lambda_\alpha \succ \mathbf{DM}^T\mathbf{H}^T\mathbf{HMD}^T \tag{12}$$

which is equivalent to requiring that:

$$\begin{aligned}\mathbf{D}^T\Lambda_\alpha\mathbf{DA} &\succ \mathbf{D}^T\mathbf{DM}^T\mathbf{H}^T\mathbf{HMD}^T\mathbf{DA} \\ &= \mathbf{A}^{-1}\mathbf{M}^T\mathbf{H}^T\mathbf{HM}\end{aligned} \tag{13}$$

where we assume $\tilde{\Lambda}_\alpha = \mathbf{D}^T\Lambda_\alpha\mathbf{DA}$. It is known that we need $\mathbf{A}\tilde{\Lambda}_\alpha \succ \rho(\mathbf{M}^T\mathbf{H}^T\mathbf{HM})$ in order to fulfill the condition. Various subband-adaptive methods can be applied to determine the diagonal matrix $\mathbf{A}\tilde{\Lambda}_\alpha$ [15, 16]. We can easily compute the Λ_α once $\tilde{\Lambda}_\alpha$ is found. In practice, Λ_α is equivalent to applying non-overlapping transformation to $\tilde{\Lambda}_\alpha$.

Because $p\left(\mathbf{z}|\hat{\mathbf{w}}\right)$ and $p\left(\hat{\mathbf{w}}|\mathbf{y}, \mathbf{S}, \nu^2\right)$ are Gaussian functions of $\mathbf{w}$, when $\mathbf{z}$ is given (typically as a previous estimate for $\mathbf{w}$), the approximation model can be rearranged into a Gaussian form as

$$\overline{p}\left(\hat{\mathbf{w}}|\mathbf{y}, \mathbf{z}, \mathbf{S}, \nu^2\right) = \mathcal{N}\left(\hat{\mathbf{w}}|\overline{\mu}, \overline{\Sigma}\right) \tag{14}$$

with

$$\overline{\mu} = \nu^{-2}\overline{\Sigma}[\Lambda_\alpha\mathbf{DAz} - \mathbf{DM}^T\mathbf{H}^T(\mathbf{HMz} - \mathbf{y})] \tag{15}$$

$$\overline{\Sigma} = (\nu^{-2}\Lambda_\alpha + \mathbf{S})^{-1} \tag{16}$$

where $\overline{\Sigma}^{-1}$ is now a purely diagonal matrix and easy to invert, which gives the subband-adaptive MM technique.

3. CONTINUATION STRATEGIES

In this section, we apply the VB approximation to derive the continuation strategies of our model. To update the variables appearing in (9), we construct a 3-layer hierarchical prior as described in [14]. To be more specific, we impose Gamma distributions for both inverse signal variance $\mathbf{s}$ and its rate parameter $\mathbf{b}$:

$$p(\mathbf{s}|a, \mathbf{b}) = \prod_{i=1}^{G} \frac{b_i^a}{\Gamma(a)} s_i^{a-1} \exp(-b_i s_i) \tag{17}$$

and

$$p(\mathbf{b}|k, \theta) = \prod_{i=1}^{G} \frac{\theta^k}{\Gamma(k)} b_i^{k-1} \exp(-\theta b_i) \tag{18}$$

When k and θ approach zero, the Gamma prior on $p(\mathbf{b})$ becomes a noninformative Jeffreys prior, and therefore the mean of signal variance σ^2 approximately follows a noninformative prior. This means that the posterior depends only on the data and not the prior, which is known to strongly promote sparse estimates. If the prior knowledge is available for the model, we can tune the hyperparameters so that the prior becomes more informative. As a result, we can move between the informative and noninformative prior more flexibly using the 3-layer model.

Now we can represent the posterior of hidden variables as

$$\begin{aligned}p\left(\hat{\mathbf{w}}, \mathbf{z}, \mathbf{s}, \mathbf{b}|\mathbf{y}\right) &= \frac{p(\hat{\mathbf{w}}, \mathbf{z}, \mathbf{s}, \mathbf{b}, \mathbf{y})}{p(\mathbf{y})} \\ &= \frac{p\left(\mathbf{y}|\hat{\mathbf{w}}, \beta\right) p\left(\hat{\mathbf{w}}|\mathbf{S}\right) p(\mathbf{z}|\hat{\mathbf{w}}) p(\mathbf{s}|a, \mathbf{b}) p(\mathbf{b}|k, \theta)}{p\left(\mathbf{y}\right)}\end{aligned} \tag{19}$$

where $\beta \equiv \nu^{-2}$. Because the marginal likelihood $p\left(\mathbf{y}\right)$ is typically intractable to compute, the exact Bayesian inference of (19) cannot be performed [17]. Here we adopt the VB approximation to approximate $p\left(\xi|\mathbf{y}\right)$ using a distribution $q\left(\xi\right)$, where $\xi = \{\hat{\mathbf{w}}, \mathbf{z}, \mathbf{s}, \mathbf{b}\}$. It is known that we can find $q(\xi)$ by minimizing the Kullback-Leibler (KL) divergence between $q(\xi)$ and $p(\xi|\mathbf{y})$ where

$$\mathrm{KL}(q(\xi)\|p(\xi|\mathbf{y})) = -\int q(\xi) \ln\left(\frac{p(\xi|\mathbf{y})}{q(\xi)}\right) d\xi \tag{20}$$

However $q(\xi)$ cannot be obtained simply as we do not know $p(\mathbf{y})$. A general approach is the mean-field approximation where we factorize $q(\xi)$ into disjoint groups:

$$q(\xi) = q(\hat{\mathbf{w}}, \mathbf{z}, \mathbf{s}, \mathbf{b}) \approx q(\hat{\mathbf{w}})q(\mathbf{z})q(\mathbf{s})q(\mathbf{b}) \tag{21}$$

Based on this factorization, the distribution of each variable $q(\lambda)$, $\lambda \in \xi$, which minimizes (20) can be optimized as

$$\begin{aligned}\ln q(\lambda) &= \langle \ln p(\xi|\mathbf{y}) \rangle_{q(\xi\setminus\lambda)} \\ &= \langle \ln p(\xi, \mathbf{y}) \rangle_{q(\xi\setminus\lambda)} + \text{const}\end{aligned} \tag{22}$$

where $\langle\cdot\rangle_q$ denotes expectation over q and $\xi \setminus \lambda$ means the set of ξ with λ removed. The procedure of iteratively updating $q(\lambda)$, for $\lambda = \hat{\mathbf{w}}$, $\mathbf{z}$, $\mathbf{s}$ and $\mathbf{b}$ in turn, results in Algorithm 1.

Algorithm 1 Tree-structured VBMM Image Deconvolution

```
**Inputs**: blur kernel $\mathbf{H}$, blurred image $\mathbf{y}$, $\Lambda_\alpha$, $a$, $\beta$, $k$, $\theta$, initial estimations of $\mathbf{z}^{(0)}$, $\mathbf{s}^{(0)}$ and $\mathbf{b}^{(0)}$.
**while** iterations $t = 0 : t_{\max}$ or $\mathbf{z}$ has converged, **do**
    $\overline{\Sigma}^{(t)} = \left(\beta\Lambda_\alpha + \mathbf{S}^{(t)}\right)^{-1}$
    $\overline{\mu}^{(t)} = \beta\overline{\Sigma}^{(t)}[\Lambda_\alpha\mathbf{DAz}^{(t)} - \mathbf{DM}^T\mathbf{H}^T(\mathbf{HMz}^{(t)} - \mathbf{y})]$
    $\hat{\mathbf{w}}^{(t+1)} = \overline{\mu}^{(t)}$
    $\mathbf{z}^{(t+1)} = \mathbf{D}^T\hat{\mathbf{w}}^{(t+1)}$
    $s_i^{(t+1)} = \frac{g_i + 2a}{\left(\|\overline{\mu}_i^{(t)}\|^2 + \mathrm{tr}[\overline{\Sigma}_i^{(t)}]\right) + 2b_i^{(t)}}$ for $i = 1...G$
    $b_i^{(t+1)} = \frac{a + k}{s_i^{(t+1)} + \theta}$ for $i = 1...G$
**end while**
**Output** deblurred image $\mathbf{x} = \mathbf{Mz}$
```

Table 1. BLUR, Noise Variance and BSNR (dB)

Exp.	BLUR	ν^2	BSNR
1	9×9 uniform	31.10	20
2	9×9 uniform	0.31	40
3	9×9 uniform	0.03	50
4	$h_{ij} = 1/(1+i^2+j^2), i, j = -7, \ldots, 7$	2	31.85
5	$h_{ij} = 1/(1+i^2+j^2), i, j = -7, \ldots, 7$	8	25.85

4. RESULTS

In this section, we test our claim that the VBMM algorithm which incorporates overlapping group sparsity outperforms the coefficient-sparse VBMM algorithm, referred to as Coefficient VBMM (VC) in [14]. We have used both the "parent+1child" grouping (V1) and the "parent+4children" grouping (V4) strategies for these experiments. We have used the DT $\mathbb{C}$WT as our redundant sparsifying transform because it has good sparsity inducing properties. Because DT $\mathbb{C}$WT produces complex coefficients, we assume a pair of real and imaginary coefficients share the same variance and can be clustered into one group. As a result, we have $G = \frac{N}{2}$ groups for VC, $G = \frac{P+6}{4}$ groups for V1 and $G = \frac{P+6}{10}$ groups for V4. Five experiments were performed as shown in Table 1, where we convolved the Cameraman image with two different blur kernels: 9×9 uniform blur (Exp. 1-Exp. 3) and 15×15 circular-symmetric blur as $h_{ij} = \frac{1}{1+i^2+j^2}, i, j = -7, \ldots, 7$ (Exp. 4-Exp. 5). White Gaussian noise was added to the blurred image and the blurred signal-to-noise ratio (BSNR)$=10 \log_{10} \frac{\|\mathbf{Hx}_r - \overline{\mathbf{Hx}_r}\|^2}{M\nu^2}$ was used to define the noise level. $\mathbf{x}_r$ is the original image and $\overline{\mathbf{Hx}_r}$ is the mean of $\mathbf{Hx}_r$. The improvement in signal-to-noise ratio (ISNR) $=10 \log_{10}(\frac{\|\mathbf{y}-\mathbf{x}_r\|^2}{\|\mathbf{Mz}-\mathbf{x}_r\|^2})$ was used to evaluate the relative performance of VC, V1, and V4. A regularized Wiener filter $\mathbf{x}_0$=$(\mathbf{H}^T\mathbf{H}$+$10^{-3}\nu^2\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}$ was used to estimate the initial $\mathbf{x}_r$ and hence $\mathbf{z}^{(0)} = \mathbf{M}^T\mathbf{x}_0$ [16]. In the experiment, we set hyperparameters $a = \theta = 10^{-6}$ and adjusted k to control the sparsity where k should satisfy $0 < k < \frac{g_i}{2}$. We've ensured the matrix $\mathbf{\Lambda}_\alpha$ for the VC, V1, and V4 experiments was the same for each test scenario.

For all test cases, incorporating the group sparse penalty leads to improved deconvolution results in terms of visual quality and final ISNR. Fig. 4 shows the visual and ISNR results of Exp. 2 obtained from applying the VC, V1 and V4 algorithms over 200 iterations. V1 performs similarly to V4, and hence we decided not to display the results due to space constraints. In Table 2, we show average ISNR values obtained from repeating our experiments over 30 noise realizations. It is found that both V1 and V4 have faster convergence and better ISNR results compared with VC. Furthermore, V4 gives better ISNR results than V1. We believe this is because V4 considers both dependencies across scale and also dependencies within scale. We also tested VC, V1 and V4 on the circular-symmetric blur and the ISNR results are shown in Fig. 5[1]. It is observed that both V1 and V4 converge to higher ISNR values than VC, but V4 converges slower than V1 initially because $\mathbf{DM}^T\mathbf{H}^T\mathbf{HMD}^T$ becomes less leading diagonal using V4, which may cause the (Λ_α–$\mathbf{DM}^T\mathbf{H}^T\mathbf{HMD}^T$) term to be larger and slow down the convergence.

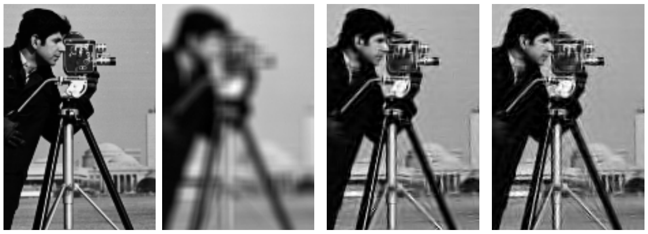
(a) Original (b) Blurred (c) VC strategy, ISNR: 8.09 dB (d) V4 strategy, ISNR: 8.29 dB

Fig. 4. Deconvolution and ISNR (dB) results on Exp. 2 for VC and V4, on Cameraman, BSNR: 40 dB.

Table 2. Average ISNR (dB) results for VC, V1 and V4 over 30 noise realizations using the 9×9 uniform blur.

iters		10	30	50	70	100
Exp.1	VC	2.66	3.15	3.35	3.45	3.52
	V1	2.95	3.55	3.67	3.71	3.74
	V4	**2.96**	**3.62**	**3.72**	**3.74**	**3.76**
Exp.2	VC	7.20	7.66	7.85	7.95	8.04
	V1	**7.63**	7.99	8.11	8.16	8.20
	V4	7.57	**8.01**	**8.14**	**8.20**	**8.24**
Exp.3	VC	10.17	10.66	10.87	10.99	11.08
	V1	10.17	10.75	10.94	11.04	11.14
	V4	**10.19**	**10.86**	**11.06**	**11.16**	**11.26**

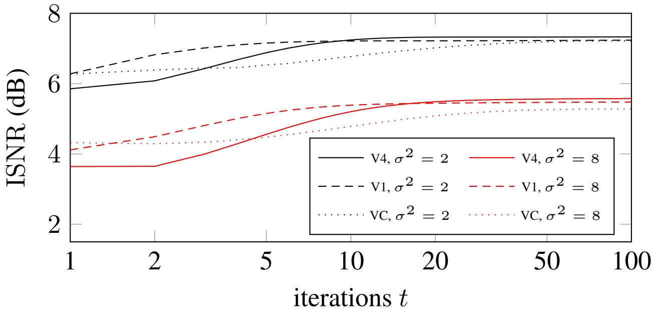

Fig. 5. Average ISNR (dB) on Exp. 4 and Exp. 5 for VC, V1, and V4 over 30 noise realizations.

5. CONCLUSION

Here we have proposed an extension of the VBMM deconvolution algorithm which incorporates tree-structured wavelet modeling and two grouping strategies are discussed. We have shown how to incorporate a wavelet tree structure in an empirical Bayesian derivation. Our model gives some useful improvements over an equivalent method without wavelet group structure, while the computation per iteration increases by about 29 % for V1 or 38 % for V4 group structures, relative to the basic VC scheme which takes 0.09 seconds per iteration.

[1] Numerical results of Exp. 4 and Exp. 5 are available online at http://www-sigproc.eng.cam.ac.uk/Main/GZ243.

6. REFERENCES

[1] J. Oliveira, J. Bioucas-Dias, and M. Figueiredo, "Adaptive total variation image deblurring: A majorization–minimization approach," *Signal Process.*, vol. 89, no. 9, pp. 1683–1693, 2009.

[2] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, pp. 906–916, 2003.

[3] J. Bioucas-Dias, "Bayesian wavelet-based image deconvolution: a GEM algorithm exploiting a class of heavy-tailed priors," *IEEE Trans. Image Process.*, vol. 15, pp. 937–951, 2006.

[4] M. Figueiredo, J. Bioucas-Dias, and R. Nowak, "Majorization–minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, pp. 2980–2991, 2007.

[5] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, pp. 1338–1351, 2003.

[6] N. Rao, R. Nowak, S. Wright, and N. Kingsbury, "Convex approaches to model wavelet sparsity patterns," in *Proc. IEEE ICIP 2011*, 2011, pp. 1917–1920.

[7] H. Choi, J. Romberg, R. Baraniuk, and N. Kingsbury, "Hidden Markov tree modeling of complex wavelet transforms," in *Proc. IEEE ICASSP 2000*, 2000, vol. 1, pp. 133–136.

[8] L. Sendur and I. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2744–2756, 2002.

[9] J. Romberg, M. Wakin, H. Choi, and R. Baraniuk, "A geometric hidden Markov tree wavelet model," in *Optical Science and Technology, SPIE's 48th Annual Meeting*. International Society for Optics and Photonics, 2003, pp. 80–86.

[10] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, 1998.

[11] C. Chen and J. Huang, "Compressive sensing MRI with wavelet tree sparsity," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1124–1132.

[12] L. Baldassarre, N. Bhan, V. Cevher, and A. Kyrillidis, "Group-sparse model selection: Hardness and relaxations," *arXiv preprint arXiv:1303.3207*, 2013.

[13] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inform. Theory*, vol. 56, pp. 1982–2001, 2010.

[14] G. Zhang and N. Kingsbury, "Fast L0-based image deconvolution with variational Bayesian inference and majorization-minimization," in *GlobalSIP 2013 Symposium on Optimization in Machine Learning and Signal Processing*, 2013, pp. 1081–1084.

[15] I. Bayram and I. Selesnick, "A subband adaptive iterative shrinkage/thresholding algorithm," *IEEE Trans. Signal Process.*, vol. 58, pp. 1131–1143, 2010.

[16] Y. Zhang and N. Kingsbury, "Improved bounds for subband-adaptive iterative shrinkage/thresholding algorithms," *IEEE Trans. Image Process.*, vol. 22, pp. 1373–1381, 2013.

[17] D. Tzikas, A. Likas, and N. Galatsanos, "The variational approximation for Bayesian inference: Life after the EM algotrithm," *IEEE Signal Process. Mag.*, vol. 25, pp. 131–146, 2008.