# Efficient Belief Propagation with Learned Higher-Order Markov Random Fields

Xiangyang Lan[1,⋆], Stefan Roth[2,⋆], Daniel Huttenlocher[1], and Michael J. Black[2]

[1] Computer Science Department, Cornell University, Ithaca, NY, USA
{xylan, dph}@cs.cornell.edu
[2] Department of Computer Science, Brown University, Providence, RI, USA
{roth, black}@cs.brown.edu

**Abstract.** Belief propagation (BP) has become widely used for low-level vision problems and various inference techniques have been proposed for loopy graphs. These methods typically rely on ad hoc spatial priors such as the Potts model. In this paper we investigate the use of learned models of image structure, and demonstrate the improvements obtained over previous ad hoc models for the image denoising problem. In particular, we show how both pairwise and higher-order Markov random fields with learned clique potentials capture rich image structures that better represent the properties of natural images. These models are learned using the recently proposed Fields-of-Experts framework. For such models, however, traditional BP is computationally expensive. Consequently we propose some approximation methods that make BP with learned potentials practical. In the case of pairwise models we propose a novel approximation of robust potentials using a finite family of quadratics. In the case of higher order MRFs, with $2 \times 2$ cliques, we use an adaptive state space to handle the increased complexity. Extensive experiments demonstrate the power of learned models, the benefits of higher-order MRFs and the practicality of BP for these problems with the use of simple principled approximations.

## 1 Introduction

There are two current threads of research that are modernizing Markov random fields (MRFs) for machine vision. The first involves new algorithms based on belief propagation (BP) and graph cuts for performing approximate probabilistic (e. g., maximum *a posteriori*) inference on MRFs [1, 2, 3, 4, 5, 6]. These methods have extended the usefulness of MRFs by making inference tractable, but have often relied on ad hoc or hand-tuned models of spatial image structure with a limited spatial neighborhood structure (e. g., pairwise models). Such approaches have lacked the representational power needed to capture the rich statistics of natural scenes. The second line of research involves improving the expressive power of MRFs with higher-order models that are learned from data [7, 8, 9]. These approaches better capture the rich statistics of the natural world and

⋆ The first two authors contributed equally to this work, authorship order was determined randomly.

A. Leonardis, H. Bischof, and A. Prinz (Eds.): ECCV 2006, Part II, LNCS 3952, pp. 269–282, 2006.

provide a principled way of learning the model. Our goal is to combine these two lines of research to provide efficient algorithms for inference with rich, higher-order MRFs.

To that end we develop a series of principled approximations to the learned MRF models and to belief propagation. Throughout the paper we develop and test our solutions in the context of image denoising to illustrate the power of learned MRFs and the applicability of BP to these models. In particular, we exploit the recently proposed Field-of-Experts (FoE) model for learning MRFs from example data [9]. We start with the case of pairwise MRFs, where previous work on efficient inference schemes has relied on ad hoc potentials such as the Potts model [1] or the truncated quadratic [4]. While the FoE models exploit robust potentials that better match the image statistics, these potentials do not readily admit efficient inference. We develop an approximation method that, for a pairwise MRF, represents such robust potentials as a finite family of quadratics. With such a representation, the distance transform method of [4] can be employed for efficient inference. We apply the method to image denoising and find that the resulting algorithm is several times faster than regular BP, achieves a lower energy state, and is considerably more accurate than the ad hoc model proposed in [4]. We also note that in loopy graphs such as this, convergence of BP depends on the message passing scheme employed. We show that a randomized scheme helps achieve a lower energy state than synchronous updates.

It is often observed that maximum *a posteriori* (MAP) estimates using MRF models produce piecewise constant results. This is true in the case of pairwise cliques where the potential function is robust (i. e., it downweights outliers). Such results are due to the representational weakness of pairwise models, which are too local to capture the richness of natural image statistics. To alleviate these effects we use the FoE framework to learn higher-order models of images; in particular we learn an MRF with $2 \times 2$ cliques. While such a model produces much more natural results that are no longer piecewise constant, inference becomes much harder. Applying standard BP to MRFs with $2 \times 2$ cliques requires $\mathcal{O}(N^4)$ operations to compute each message, where $N$ is the number of labels for each pixel. In case of image denoising, $N = 256$ making traditional BP algorithms impractical. Consequently we propose an approximate BP algorithm that uses an adaptive state space to reduce the number of states for each pixel, as well as a further state quantization that speeds up the message computations. Despite this approximation, the learned higher-order model outperforms learned pairwise MRF models, both visually and quantitatively.

In the following sections we introduce Markov random fields and loopy belief propagation along with our proposed approximations. We will review the related work in the context of our methods and their applicability. We present the results of experiments on image denoising that compare different MRF models as well as different BP methods.

## 2 Learning Markov Random Field Models of Images

In this paper we use two different types of Markov random fields to model the prior probability of images: pairwise MRFs and higher-order MRFs with larger, square-shaped cliques. The pairwise MRFs employed here are very similar to models that have been popular for a long time [10]; the higher-order MRFs follow the recently proposed Fields-of-Experts (FoE) approach [9]. Richer models of natural images have also been proposed on the basis of MRFs with multiple pairwise pixel interactions [11, 12]. We are not following this approach here, but a comparison of the benefits of these approaches deserves further study.

We assume that pixels in an image are represented as nodes $V$ in a graph $G = (V, E)$. In the pairwise case, the set of edges $E$ connects all nodes that are either horizontal or vertical neighbors. In the higher-order case, the set of edges fully connects all nodes in all possible square $m \times m$ image regions. The probability of an image $\mathbf{x}$ under such a Markov random field can be written as a product over all the maximal cliques $C$:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C} \Psi(\mathbf{x}_C), \tag{1}$$

where $\mathbf{x}_C$ is the image region corresponding to clique $C$, $\Psi$ is a positive potential function, and $Z$ is a normalization term.

In the pairwise case, the potentials are typically defined as a function of the grayvalue difference of the two neighboring pixels. The grayvalue difference can be interpreted as a local approximation of the horizontal or vertical image gradient. The MRF model penalizes large gradients and so models the fact that images are often locally smooth. In the natural image statistics literature it has been observed that the marginal distribution of the image gradient is highly kurtotic [13]; marginal gradient histograms show substantial probability mass in the tails. This results from the fact that images occasionally show significant jumps in intensity that for example arise from object boundaries. In order to model this behavior, the pairwise MRF we use here relies on robust potentials based on Student t-distributions, which resemble the marginal statistics of the image gradient. If $x_{C,1}$ and $x_{C,2}$ are the two pixels for the pairwise clique $\mathbf{x}_C$, then we use the potential

$$\Psi_{\mathrm{pw}}(\mathbf{x}_C) = \left(1 + \frac{1}{2}\left(\frac{x_{C,1} - x_{C,2}}{\sigma}\right)^2\right)^{-\alpha}. \tag{2}$$

We will learn two separate parameter sets $(\sigma_{\mathrm{H}}, \alpha_{\mathrm{H}})$ and $(\sigma_{\mathrm{V}}, \alpha_{\mathrm{V}})$ for horizontal and vertical edges respectively, yielding a pairwise image prior $p_{\mathrm{pw}}(\mathbf{x})$.

The Fields-of-Experts framework [9] used in the higher-order MRF case models the clique potentials using a so-called Product of Experts (PoE) [14]. The idea behind the PoE is to model complex distributions as the product of several simpler expert distributions that each work on a low-dimensional subspace, in this case a linear 1D subspace. In the context of images, these linear 1D subspaces can be interpreted as linear filters $\mathbf{J}_i$ applied to the image patch $\mathbf{x}_C$.

It has been observed that, for a wide variety of linear filters, the statistics of the filter responses are highly kurtotic[13]. Consequently, following [9] we take the experts to be Student t-distributions. Assuming that we use $K$ experts, we can write the prior probability of an image under the FoE model as

$$p_{m \times m}(\mathbf{x}) = \frac{1}{Z} \prod_{C} \prod_{i=1}^{K} \phi(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_C; \alpha_i), \tag{3}$$

where $\phi$ is an unnormalized t-distribution with parameter $\alpha_i$:

$$\phi(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_C; \alpha_i) = \left(1 + \frac{1}{2}(\mathbf{J}_i^{\mathrm{T}} \mathbf{x}_C)^2\right)^{-\alpha_i}. \tag{4}$$

Following [9], we trained both types of MRF models using a database of natural images [15]. In the case of the pairwise model we learn the parameters $\alpha_{\mathrm{H}}, \alpha_{\mathrm{V}}, \sigma_{\mathrm{H}}$, and $\sigma_{\mathrm{V}}$, while in the FoE case we learn the filters $\mathbf{J}_i$ as well as the expert parameters $\alpha_i$. To make belief propagation inference tractable as detailed in Section 3, we restrict ourselves to $2 \times 2$ models and use 3 experts. We randomly cropped 2000 patches of $9 \times 9$ pixels out of the training database and found suitable parameters by (approximately) maximizing the likelihood of the data. The learning algorithm is based on stochastic gradient ascent, and uses the idea of contrastive divergence [16] to make it more efficient. Since the proposed pairwise MRF can be treated as special case of the FoE model, they can both be trained in essentially the same way. The learning procedure follows the description in [9], to which we refer the reader for more details.

## 3 Efficient Belief Propagation

Many low-level vision problems can be posed as problems of Bayesian inference, and can be described in the following common framework: Given some observed image $\mathbf{I}$, the goal is to estimate a hidden state $\mathbf{x}$ according to a posterior distribution $\mathrm{p}(\mathbf{x} \,|\, \mathbf{I})$. The hidden state may, for example, correspond to a smoothed image in the case of image denoising, or to a dense disparity map in the case of stereo (where $\mathbf{I}$ in fact represents two images). Here a set of discrete labels is used to represent the state of each hidden variable. The posterior distribution of the hidden state $\mathbf{x}$ given the input image $\mathbf{I}$ is modeled as $p(\mathbf{x} \,|\, \mathbf{I}) = 1/Z \cdot p(\mathbf{I} \,|\, \mathbf{x}) \cdot p(\mathbf{x})$, where $p(\mathbf{I} \,|\, \mathbf{x})$ is the likelihood of the observed image given a hidden labeling and $p(\mathbf{x})$ is the prior probability over labelings. Rather than relying on ad hoc spatial priors, we use the learned priors introduced above, a pairwise prior $p_{\mathrm{pw}}(\mathbf{x})$ and a higher-order prior $p_{2\mathrm{x}2}(\mathbf{x})$. Because the normalization term $Z$ is unknown and intractable to compute in general, we will sometimes refer to the energy $E(\mathbf{x}; \mathbf{I})$ of a labeling $\mathbf{x}$; that is, the unnormalized log-posterior. The energy is related to the posterior distribution through $p(\mathbf{x} \,|\, \mathbf{I}) = 1/Z \cdot \exp\{-E(\mathbf{x}; \mathbf{I})\}$. Note that maximizing the posterior probability is equivalent to minimizing the energy.

There are two basic ways of estimating this labeling, one of which is to compute the expectation of the posterior $p(\mathbf{x} \,|\, \mathbf{I})$ and the other is to compute the maximum (i. e., the MAP estimate). We consider both of these problems here, but use the former as a running example for discussing the proposed algorithms. In general finding exact solutions to these estimation problems is hard for loopy graphs, but approximation approaches based on graph cuts [1, 3, 17, 18] and loopy belief propagation [6, 18, 19] have been found to often work well in practice. The focus of this paper is the family of loopy belief propagation algorithms. In order to apply them to Bayesian inference problems, the posterior must factor into products over relatively small numbers of variables in order to be computationally feasible. In particular it is customary to require that the prior factor into a product of functions $\Psi_h$ over small subsets of nodes $C_h$ (cliques in the underlying hidden layer) and the likelihood factors into a product of functions $\Psi_o$ over small subsets of nodes $C_o$ (often individual nodes, e. g., in image denoising),

$$p(\mathbf{x} \mid \mathbf{I}) = \frac{1}{Z} \prod_{C_o} \Psi_o(\mathbf{x}_{C_o}; \mathbf{I}) \prod_{C_h} \Psi_h(\mathbf{x}_{C_h}), \tag{5}$$

where $\mathbf{x}_{C_o}$ corresponds to the cliques of the likelihood and $\mathbf{x}_{C_h}$ corresponds to the cliques of the spatial prior. In the description of the message passing algorithm below, we will handle both types of cliques and potentials in a unified way, i. e., $p(\mathbf{x} \mid \mathbf{I}) = 1/Z \cdot \prod_C \Psi_C(\mathbf{x}_C; \mathbf{I})$.

Both pairwise and higher-order models can be considered in a common framework using factor graphs [19]. A factor graph is a bipartite graph with edges connecting two kinds of nodes, variable nodes and factor nodes. A variable node corresponds to an individual random variable $x_i$, while a factor node corresponds to a subset (clique) of random variables $\mathbf{x}_C$, whose potential function $\Psi_C(\mathbf{x}_C; \mathbf{I})$ is a specific term in the factorized form of the posterior distribution. Edges in the factor graph connect each factor node to those variables that are involved in its potential function. For models defined on the image grid, the $x_i$ and the associated variable nodes can be seen as corresponding to image pixels, and the $\mathbf{x}_C$ and the associated factor nodes correspond to local neighborhoods (cliques) in the image. See Figure 3 for examples of factor graph representations for a pairwise MRF and a $2 \times 2$ MRF on an image grid. These graphical illustrations include nodes corresponding to the observed data at each pixel.

Belief propagation operates by passing messages between nodes until convergence (which is generally not guaranteed but is usually observed in practice). All message entries are usually initialized to the same value to represent an uninformative prior. We now turn to the message update rules for the sum-product BP algorithm on a factor graph [6, 19], in which case each iteration contains two types of message updates.

For the first type of message, a variable node $i$ sends a message $n_{i \to C}(x_i)$ to a neighboring factor node $C$. To do so it computes the product of the messages received from its other neighboring factor nodes,

$$n_{i \to C}(x_i) = \prod_{C' \in \mathcal{N}(i) \setminus C} m_{C' \to i}(x_i), \tag{6}$$

where $\mathcal{N}(i) \setminus C$ denotes the neighboring factor nodes of $i$ other than $C$.

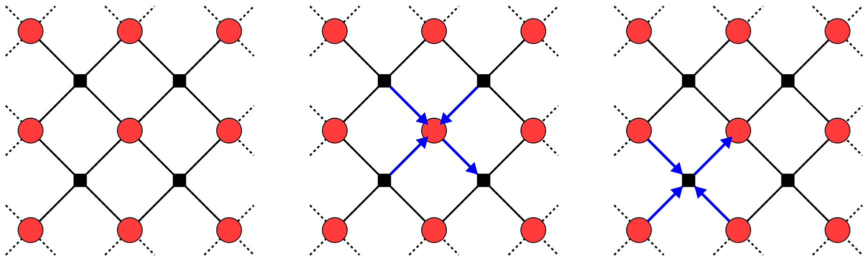


**Fig. 1.** *(a)* Factor graph structure of an image prior with $2 \times 2$ cliques. Red circles correspond to variable nodes (image pixels) and black squares correspond to factor nodes (cliques representing local neighborhood). *(b)* Message passing from a variable node to a factor node (cf. Eq. (6)). *(c)* Message passing from a factor node to a variable node (cf. Eq. (7)).

For the second type of message, a factor node $C$ sends a message $m_{C \to i}$ to a neighboring variable node $i$. To do so it assembles all the messages received from its other neighboring variable nodes weighted with its associated potential function $\Psi_C(\mathbf{x}_C; \mathbf{I})$,

$$m_{C \to i}(x_i) = \sum_{\mathbf{x}_C \setminus x_i} \Psi_C(\mathbf{x}_C; \mathbf{I}) \prod_{i' \in \mathcal{N}(C) \setminus i} n_{i' \to C}(x_{i'}), \tag{7}$$

where $\mathbf{x}_C \setminus x_i$ denotes the variables of $\mathbf{x}_C$ other than $x_i$. That is, $\mathbf{x}_C$ is the cross product space of a set of random variables and the summation is done over all the variables of that cross product space except $x_i$. Recall that $\Psi_C(\mathbf{x}_C; \mathbf{I})$ is the clique potential for clique $C$ in Eq. (5).

We should note that in the pairwise case this factor graph approach results in the same calculations as the loopy belief propagation algorithms on a 4-connected grid that have recently been used by a number of researchers in computer vision (e.g., [4, 5, 20]).

These message updates are iterated until an equilibrium point is reached, at which point the belief of each individual variable node can be computed as

$$b_i(x_i) = \prod_{C \in \mathcal{N}(i)} m_{C \to i}(x_i). \tag{8}$$

Taking the belief as an approximation of the marginal posterior probability, we can then estimate a state for each variable node by taking its expected value.

The sum-product technique that we have presented here approximates the marginal posterior probabilities of the variable nodes. In contrast, the max-product technique is used to approximately compute the MAP estimate. The main differences are the replacement of sums by maximizations in the message update equations, and the replacement of expectation by maximization to compute a final label for each variable node. The max-product formulation has often been used for pixel labeling problems such as stereo and image denoising, whereas the sum-product formulation may be more appropriate to interpolation problems where non-integer solutions may be desired.

The running time for either the sum-product or max-product BP algorithm on a factor graph is $\mathcal{O}(MN^k)$, where $M$ is the number of image pixels, $N$ is the number of possible labels for each pixel, and $k$ is the maximum clique size. For problems like image denoising with $N = 256$ labels corresponding to image intensities, the computational cost is very large. In the next two subsections, we introduce simple but effective techniques to speed up BP algorithms for learned potentials of pairwise and $2 \times 2$ MRFs.

### 3.1 Pairwise MRFs

Standard belief propagation on a 4-connected grid for pairwise MRFs is in general still a computationally demanding task, because it requires $\mathcal{O}(M \cdot N^2)$ steps. It has recently been shown [4] that max-product belief propagation can be carried out more efficiently for pairwise MRFs with certain kinds of potentials by exploiting a distance transform technique. In these cases, the time complexity is linear rather than quadratic in the number of labels, i. e., $\mathcal{O}(MN)$. In particular, if the negative log of the pairwise potentials can be expressed as the lower envelope of (possibly truncated) quadratic or absolute value functions of the pairwise pixel difference then the distance transform technique can be applied.

We extend this work here by applying the distance transform technique to MRFs where the potentials have been learned from data. To that end, we exploit the fact that a large set of robust error functions can be written as the infimum over a family of quadratics as shown by Black and Rangarajan [21]. As discussed earlier, we model the pairwise potentials using Student-t distributions (see Eq. (2)). The t-distribution has the corresponding robust error function $\rho(y) = \alpha \log\left(1 + \frac{1}{2}\left(\frac{y}{\sigma}\right)^2\right)$, where $y$ is the grayvalue difference between neighboring pixels. A derivation similar to the one in [21] reveals that this robust function can be written as $\rho(y) = \inf_z E(y, z)$ with

$$E(y, z) = \frac{y^2}{2\sigma^2} z + z - \alpha + \alpha \log \frac{\alpha}{z}, \tag{9}$$

which is a quadratic function in $y$ and where $z$ is an "outlier process". Instead of writing the negative log of the potential as the infimum over all possible $z$ values in the range $[0, \alpha]$, we approximate it as the minimum (lower envelope) over a fixed, discrete set of $z$ values. Given a fixed number $k$ of quadratics, we find a good approximation by a simple local optimization of the Kullback-Leibler divergence between the learned t-distribution and the probability density corresponding to the lower envelope of the quadratics. We compute the KL divergence using a discrete histogram with range $[-255, 255]$ and 10 bins per gray level. To improve numerical stability we modify the log of the $z$ values and upper bound the $z$ values with a simple penalty function so that $z \leq \alpha$. Figure 2

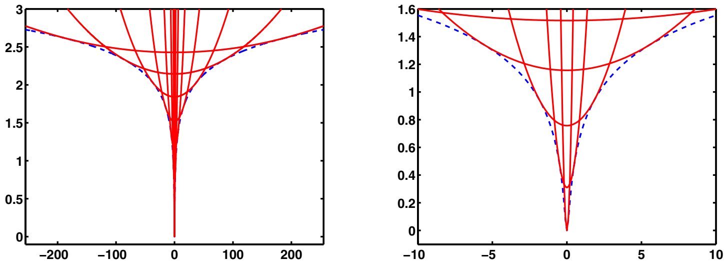


**Fig. 2.** Approximation of the negative log of a Student-t distribution as the lower envelope of 8 quadratics. *(left)* Full grayvalue range $[-255, 255]$. *(right)* Detail of the same approximation over the range $[-10, 10]$.

shows how the negative log of a t-distribution is approximated with 8 quadratics. In the experimental results section, we compare these approximations using 4 and 8 quadratics (using the efficient linear-time distance transform algorithm) with the actual t-distribution (using the conventional quadratic-time algorithm). For details of the distance transform method the reader is referred to [4].

### 3.2 Higher-Order MRFs

Our experiments show that pairwise models as just described suffer from the problem that the optimal solution is piecewise constant (see Figure 4). To overcome this problem, we have to move to using higher-order MRF priors as introduced in Section 2 and illustrated in Figure 1(a). Unfortunately, applying the factor graph belief propagation algorithm directly is infeasible for such models. For $m \times m$ maximal cliques the summation (or maximization in the max-product case) in Eq. (7) is taken over $N^{m \cdot m - 1}$ terms, which is prohibitively expensive even in the $2 \times 2$ case with $N = 256$ labels.

In order to alleviate this problem, we devised a simple, but effective adaptive state space procedure. In many applications, we can fairly reliably estimate a grayvalue range for each pixel that will contain the optimal solution as well as most of the probability mass of the belief. To determine the working range for denoising problems, we find the minimal and maximal grayvalue in a $3 \times 3$ search window around each pixel. To avoid overestimating the range in the presence of noise, we preprocess the image for the range determination step with a very small amount of Gaussian smoothing ($\sigma = 0.7$); denoising is carried out on the original image. When performing the sum-product or max-product operation for a specific pixel $i$ within a factor node $C$ with size $2 \times 2$ (see Eq. (7)), we discretize the label set for the other 3 member pixels into $h$ bins over that range, and only consider those $h^3$ different combinations. Furthermore, we can reduce the computation time for the message updates from Eq. (6) and Eq. (7) by restricting them to the determined range. By using this adaptively quantized state space, the time complexity of BP for a $2 \times 2$ MRF model decreases to $\mathcal{O}(M \cdot N \cdot h^3)$.

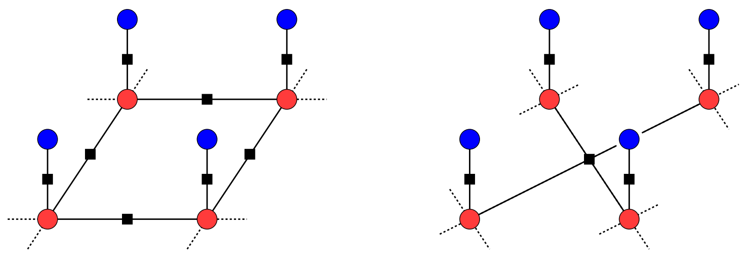

**Fig. 3.** Graphical model structure for image reconstruction. The round nodes represent observed (blue) and hidden (light red) variables; the square nodes are the factor nodes indicating the clique structure. *(left)* Pairwise Markov random field for image denoising. *(right)* Field-of-Experts model for denoising with $2 \times 2$ cliques in the prior.

## 4 Belief Propagation and Image Denoising

To focus on the effects of the different models and our approximations we choose an inference problem with a simple likelihood term: image denoising with known additive noise. As it is common in the denoising literature (e. g., [22]) we assume that images have been contaminated with artificial i. i. d. Gaussian noise, which also facilitates quantitative evaluation. We furthermore assume that the standard deviation $\sigma$ is known; we use $\sigma = 10$ and $\sigma = 20$ here. We can thus write the likelihood of noisy image $\mathbf{I}$ given the true image $\mathbf{x}$ as

$$p(\mathbf{I} \,|\, \mathbf{x}) \propto \prod_{j=1}^{M} e^{-\frac{(x_j - I_j)^2}{2\sigma^2}}. \tag{10}$$

When we combine the Gaussian likelihood with the pairwise prior $p_{\text{pw}}(\mathbf{x})$, the posterior distribution has the form of a pairwise Markov random field, where each observed pixel $I_j$ is connected to a hidden, true pixel $x_j$, and the hidden pixels are all connected to their horizontal and vertical neighbors. When combined with the $2 \times 2$ prior, $2 \times 2$ patches of hidden variables are connected with a single factor node, while the observed pixels $I_j$ are still connected to their hidden counterparts $x_j$. Figure 3 illustrates these two structures.

For quantitative evaluation we use a set of 10 images from the test set of the Berkeley segmentation dataset [15]. The images are reduced to half their original size for efficiency reasons, and only the luminance channel is used. The denoising performance is measured using the peak signal-to-noise ratio (PSNR) averaged over all 10 images ($\text{PSNR} = 20 \log_{10}(255/\sigma_e)$, where $\sigma_e$ is the standard deviation of the reconstruction error), as well as a perceptually-based image similarity metric SSIM [23].

*Learned pairwise models.* We first compared the learned pairwise MRF to the hand-defined MRF model from [4], which uses truncated quadratic potentials. In both cases, the denoised image is computed with max-product belief propagation using 20 iterations (equivalently implemented as the min-sum algorithm).

**Table 1.** Average denoising performance of various inference techniques and models on 10 test images

| | | Model from [4] | pairwise MRF t-dist. | pairwise MRF 8 quad. | 2 × 2 MRF | |
|---|---|---|---|---|---|---|
| | | max-pr. | max-pr. | max-pr. | max-pr. | sum-pr. |
| $\sigma = 10$ | PSNR | 21.98dB | 30.73dB | 29.56dB | **30.89dB** | 30.42dB |
| | SSIM [23] | 0.772 | 0.876 | 0.844 | **0.881** | 0.876 |
| $\sigma = 20$ | PSNR | 20.82dB | 26.66dB | 25.92dB | 26.85dB | **27.29dB** |
| | SSIM | 0.630 | 0.754 | 0.711 | 0.755 | **0.772** |

On a 3GHz Xeon, one BP iteration on a $256 \times 256$ image takes about 30 seconds. We find that the model proposed here substantially outperforms the model from [4] using the suggested parameters, both visually and quantitatively. As detailed in Table 1, the PSNR of the learned model is better by more than 5dB. Figure 4 shows one of the 10 test images, in which we can see that the denoising results from the learned model show characteristic piecewise constant patches, whereas the results from the hand-defined model are overly smooth in many places. Even though the performance of the truncated quadratic model could potentially be increased by hand-tuning its parameters, we refrained from doing so to demonstrate how learned MRFs can lead to competitive denoising results without requiring any manual parameter tuning. Nevertheless, we should note that BP inference is several times slower in the learned MRF case.

*Random message updates.* Based on our observation that the beliefs would not converge in case of the learned model and synchronous message updates (even though the energy seemingly converged), we also applied asynchronous message update schemes. A fixed, checkerboard-like update scheme led to some improvement in the behavior, but we found that random message updates led to the most reliable convergence behavior. At every iteration, each message is updated with a fixed probability, otherwise the previous state is kept. Table 2 shows that random updates led to a dramatic decrease in energy, but no considerable change in PSNR. Moreover, faint checkerboard-like artifacts that were visible before disappear after applying random updates. The update probability does not seem to have any substantial effect on the results (as long as it is not 100%).

*Approximate potentials.* We then investigated how the approximations of the learned potentials as a lower envelope of quadratics affect the denoising results as well as the running time. We found that max-product BP with 8 quadratics is about 6 times faster in practice than when the Student-t potentials are used. The approximation with only 4 quadratics is even faster by a factor of 2. Table 2 shows that the PSNR deteriorates by about 1dB when the approximate potentials are used (both with and without random updates); nevertheless, this still considerably outperforms the hand-designed model from [4]. We also report the average energy $\bar{E}$ of the reconstructed images in all cases computed using the original model and normalized by the number of pixels. Surprisingly, the

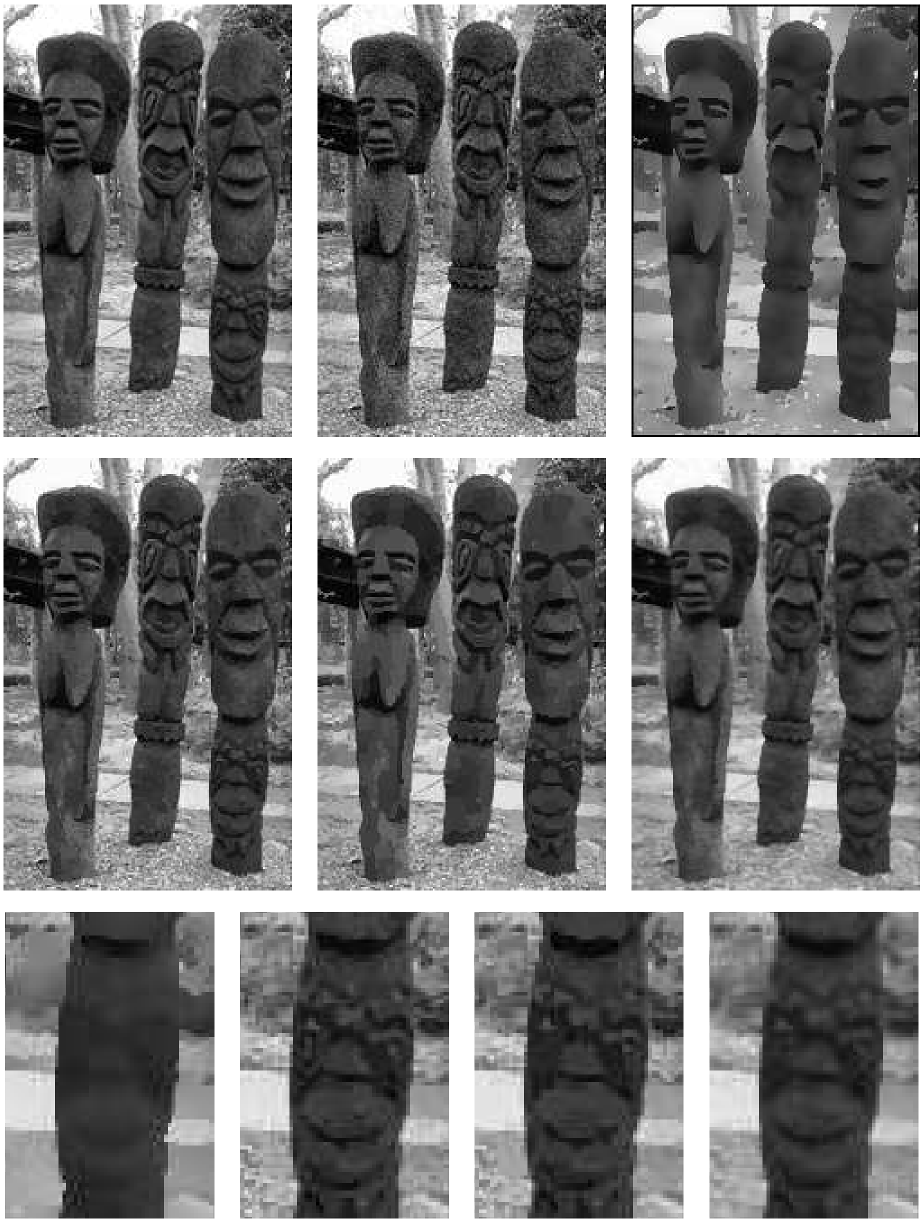

**Fig. 4.** Image denoising. **Top row:** *(left)* Original image. *(middle)* Noisy image ($\sigma = 10$). *(right)* Max-product BP with model from [4]. **Middle row:** *(left)* Max-product BP with t-distribution potentials. *(middle)* Max-product BP with approximate potentials (8 quadratics). *(right)* Max-product BP with learned $2 \times 2$ model. **Bottom row:** Detail view. From left to right: Model from [4], BP with t-distribution potentials, BP with approximate potentials (8 quadratics), BP with learned $2 \times 2$ model.

**Table 2.** Average denoising performance on 10 images for pairwise BP algorithms with and without the use of approximate models. The update percentage denotes the probability of each message being updated during a particular iteration.

| | | Student-t potentials | | | | 8 quadratics | | 4 quadratics | |
|---|---|---|---|---|---|---|---|---|---|
| Update percentage | | 25% | 50% | 75% | 100% | 50% | 100% | 50% | 100% |
| $\sigma = 10$ | $\bar{E}$ | 1.595 | 1.594 | 1.594 | 2.071 | 1.348 | 2.687 | **1.347** | 2.681 |
| | PSNR in dB | 30.73 | 30.73 | 30.73 | **30.74** | 29.56 | 29.60 | 29.54 | 29.57 |
| | SSIM [23] | 0.876 | 0.876 | 0.876 | **0.876** | 0.844 | 0.842 | 0.843 | 0.841 |
| $\sigma = 20$ | $\bar{E}$ | 1.189 | 1.182 | 1.182 | 2.606 | 1.025 | 2.892 | **1.024** | 2.907 |
| | PSNR in dB | 26.64 | 26.66 | **26.67** | 26.67 | 25.92 | 25.96 | 25.90 | 25.95 |
| | SSIM | 0.753 | 0.754 | **0.755** | 0.750 | 0.711 | 0.705 | 0.710 | 0.704 |

reconstructions using the approximate model have a lower energy than the results from the original model. We have not identified any intuitive interpretation of this fact, except that this evidences that BP may not be able to find the global optimum due to the loopiness of the graph.

*Higher-order models.* Next, we applied the learned higher-order MRF model with $2 \times 2$ cliques to the denoising problem. We used the adaptive state space approach as described above, and quantized the maximization with 8 graylevels; the potential functions are not approximated in this case. One iteration takes around 16 minutes for the setup described above. Since this approximation is possible for both max-product and sum-product BP, we report results for both algorithms. Table 1 compares both algorithms to a selection of pairwise MRFs (always with 50% update probability). We can see that the higher-order model outperforms the pairwise priors by about $0.15 - 0.2$dB (with Student-t potentials), and that the sum-product algorithm seems to be more appropriate with large amounts of noise. The perceptual similarity metric exhibits the same relative performance. Visually, the results no longer exhibit any piecewise constancy. Edges are preserved using both types of models, but smoothly varying regions are preserved better using the richer, higher-order prior.

We have also compared the presented results to an implementation of a simple gradient descent inference algorithm as suggested in [9]. This algorithm attempts to locally maximize the posterior density. We found that gradient descent achieves comparable results in terms of PSNR and SSIM, in some cases performing better than BP, in others worse. For both noise levels, the average energy of the max-product BP solution is slightly higher than that of the gradient descent algorithm (possibly due to the state space adaptation).

## 5 Conclusions and Future Work

In this paper we have combined efficient belief propagation inference algorithms with learned MRFs in order to solve low level vision problems. In particular, we demonstrated the use of learned pairwise MRF models and $2 \times 2$ MRF models with robust potential functions that better capture the spatial properties of natural images. In image denoising applications we found that BP based on these learned models substantially outperforms BP based on previous ad hoc MRFs, and that higher-order MRFs lead to both visually and quantitatively superior results.

Naively applying standard BP inference algorithms on these learned MRF models is difficult due to the non-convex functional form of the robust potential function, as well as the exponential explosion of the number of computations for the message updates in the case of $2 \times 2$ cliques. We introduced two effective approximation techniques to address these difficulties. First, for the pairwise case we used a finite family of quadratics to approximate the negative log of the learned robust potential function. This permits the application of distance transform techniques to speed up the running time from quadratic to linear in the number of labels. This approximation technique is quite general and can apply to graphical models in many contexts. Second, in the case of higher-order models such as $2 \times 2$ MRFs, we avoid explicitly searching over the whole state space by determining a plausible small set of configurations for a clique.

We observed that for the pairwise model a random message update scheme can improve the convergence speed as well as result in a significantly lower energy than a standard synchronous message update scheme. We also found that approximating the robust pairwise potential function by a lower envelope of quadratics results in a lower energy state than directly using the robust potential. These results reinforce the need to develop a better understanding of BP in computer vision research.

Comparing the BP results to a simple gradient descent inference technique, we found that belief propagation yields competitive, but not superior results. Our hypothesis is that this may be due to the likelihood being unimodal in the denoising case for which simple inference techniques can perform well. Nevertheless, both the inference and learning techniques developed in this paper are of general use beyond the application to image denoising. In the future, we plan to apply these efficient belief propagation techniques to low-level vision applications with multi-modal likelihoods, such as stereo or optical flow, in which case belief propagation may lead to superior results. Such problems often also have a smaller labeling set, and may thus allow us to use models of even higher-order.

*Acknowledgments.* S.R. and M.J.B. were supported by Intel Research and NSF IIS-0535075. This support is gratefully acknowledged.

## References

1. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE PAMI **23**(11) (2001) 1222–1239
2. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE PAMI **26**(9) (2004) 1124–1137
3. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: ECCV. Volume 2352 of LNCS., Springer (2002) 82–96

4. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. In: CVPR. Volume 1. (2004) 261–268
5. Tappen, M.F., Russell, B.C., Freeman, W.T.: Efficient graphical models for processing images. In: CVPR. Volume 2. (2004) 673–680
6. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Constructing free energy approximations and generalized belief propagation algorithms. Technical Report TR-2004-040, Mitsubishi Electric Research Laboratories, Cambridge, Massachusetts (2004)
7. Zhu, S.C., Wu, Y., Mumford, D.: Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. IJCV **27**(2) (1998) 107–126
8. Paget, R., Longstaff, I.D.: Texture synthesis via a noncausal nonparametric multiscale Markov random field. IEEE T. Image Proc. **7**(6) (1998) 925–931
9. Roth, S., Black, M.J.: Fields of experts: A framework for learning image priors. In: CVPR. Volume 2. (2005) 860–867
10. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE PAMI **6** (1984) 721–741
11. Gimel'farb, G.L.: Texture modeling by multiple pairwise pixel interactions. IEEE PAMI **18**(11) (1996) 1110–1114
12. Zalesny, A., van Gool, L.: A compact model for viewpoint dependent texture synthesis. In: SMILE 2000 Workshop. Volume 2018 of LNCS. (2001) 124–143
13. Huang, J.: Statistics of Natural Images and Models. PhD thesis, Brown University (2000)
14. Hinton, G.E.: Products of experts. In: ICANN. Volume 1. (1999) 1–6
15. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV. Volume 2. (2001) 416–423
16. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neur. Comp. **14**(8) (2002) 1771–1800
17. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D.H., Cohen, M.: Interactive digital photomontage. ACM SIGGRAPH **23**(3) (2004) 294–302
18. Tappen, M.F., Freeman, W.T.: Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In: ICCV. Volume 2. (2003) 900–907
19. Kschischang, F.R., Frey, B.J., Loelinger, H.A.: Fractor graphs and the sum-product algorithm. IEEE T. Info. Th. **47**(2) (2001) 498–519
20. Sun, J., Zhen, N.N., Shum, H.Y.: Stereo matching using belief propagation. IEEE PAMI **25**(7) (2003) 787–800
21. Black, M.J., Rangarajan, A.: On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. IJCV **19**(1) (1996) 57–92
22. Portilla, J., Strela, V., Wainwright, M.J., Simoncelli, E.P.: Image denoising using scale mixtures of Gaussians in the wavelet domain. IEEE T. Image Proc. **12**(11) (2003) 1338–1351
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE T. Image Proc. **13**(4) (2004) 600–612