Efficiently Learning Random Fields for Stereo Vision with Sparse Message Passing

Jerod J. Weinman[1], Lam Tran[2], and Christopher J. Pal[2]

[1]Dept. of Computer Science
Grinnell College
Grinnell, IA 50112
weinman@grinnell.edu

[2]Dept. of Computer Science
University of Rochester
Rochester, NY 14627
{ltran,cpal}@cs.rochester.edu

Abstract. As richer models for stereo vision are constructed, there is a growing interest in learning model parameters. To estimate parameters in Markov Random Field (MRF) based stereo formulations, one usually needs to perform approximate probabilistic inference. Message passing algorithms based on variational methods and belief propagation are widely used for approximate inference in MRFs. Conditional Random Fields (CRFs) are discriminative versions of traditional MRFs and have recently been applied to the problem of stereo vision. However, CRF parameter training typically requires expensive inference steps for each iteration of optimization. Inference is particularly slow when there are many discrete disparity levels, due to high state space cardinality. We present a novel CRF for stereo matching with an explicit occlusion model and propose sparse message passing to dramatically accelerate the approximate inference needed for parameter optimization. We show that sparse variational message passing iteratively minimizes the KL divergence between the approximation and model distributions by optimizing a lower bound on the partition function. Our experimental results show reductions in inference time of one order of magnitude with no loss in approximation quality. Learning using sparse variational message passing improves results over prior work using graph cuts.

1 Introduction

There has been growing interest in creating richer models for stereo vision in which more parameters are introduced to create more accurate models. In particular, recent activity has focused on explicitly accounting for occlusions in stereo vision models. For example, Kolmogorov and Zabih [1] have directly incorporated occlusion models in an energy function and graph cuts minimization framework. Sun et al. [2] explored a symmetric stereo matching approach whereby they: (1) infer the disparity map in one view considering the occlusion map of the other view and (2) infer the occlusion map in one view given the disparity map of the other view. More recently, Yang et al. [3] have achieved impressive results building on the dual image set-up and using color weighted correlations for patch matching. They found that this approach made match scores less sensitive to occlusion boundaries as occlusions often cause color discontinuities. As all of these methods involve creating richer models to obtain greater disparity accuracy, there is a growing need to learn or estimate model parameters in an efficient and principled way.

In contrast to previous work [1][3], we are interested in developing a completely probabilistic formulation for stereo with occlusions modeled as additional states of random variables in a conditional random field (CRF). As noted by Yang et al. [3], more studies are needed to understand the behavior of algorithms for optimizing parameters in stereo models. For example, they note that their approach might be re-formulated in an expectation maximization framework. One goal of this paper is to address these types of questions in a general way. As we will show, when traditional stereo techniques are augmented with an occlusion model and cast in a CRF framework, learning can be achieved via maximum (conditional) likelihood estimation. However, learning becomes more challenging as the stereo images and probabilistic models become more realistic.

Belief propagation (BP) [4] and variational methods [5] are widely used techniques for inference in probabilistic graphical models. Both techniques have been used for inference and learning in models with applications ranging from text processing to computer vision [6,7]. Winn and Bishop proposed Variational Message Passing (VMP) [8] as a way to view many variational inference techniques, and it represents a general purpose algorithm for approximate inference. The approach is similar in nature to BP in that messages propagate local information throughout a graph, and the message computation is similar. However, VMP optimizes a lower bound on the log probability of observed variables in a generative model.

Experimental and theoretical analysis of variational methods has shown that while the asymptotic performance of other methods such as sampling [9] can be superior, frequently variational methods are faster for approximate inference. However, many real world problems require models with variables having very large state spaces. Under these conditions, inference with variational methods becomes very slow, diminishing any gains. We address this by proposing *sparse variational methods*. These methods also provide theoretical guarantees that the Kullback-Leibler (KL) divergence between approximate distributions and true distributions are iteratively minimized. Previous work by Pal et al. [10] explored sparse methods for approximate inference using BP in chain-structured graphs. Unlike varational inference, in loopy models BP does not have a direct connection to the probability of data under a model. The method we propose here combines the theoretical benefits of variational methods with the time-saving advantages of sparse messages.

In this work, we use a lattice-structured CRF for stereo vision. This leads to energy functions with a traditional form—single variable terms and pairwise terms. Importantly, unlike purely energy-based formulations [1], since we cast the stereo problem as a probability distribution, we are able to view approximate inference and learning in the model from the perspective of variational analysis. While we focus upon approximate inference and learning in lattice-structured conditional random fields [11] applied to stereo vision, our theoretical results and some experimental insights are applicable to CRFs, MRFs and Bayesian Networks with arbitrary structures.

Many techniques have been used for parameter learning in CRFs used for image labeling, such as pseudo-likelihood [12], tree-based reparameterization (TRP) [13], and contrastive divergence [14]. Pseudo-likelihood is known to give poor estimates of interaction parameters, especially in conditional models. TRP is a variant of BP and has the same potential drawbacks. Contrastive divergence uses MCMC but does not require convergence to equilibrium for approximating the model likelihood gradients used for learning. However, models for image labeling usually only have a few states, whereas the state space in stereo models is much larger, for the many possible disparities. Thus, we believe that the sparse learning techniques we propose here will be an important contribution, providing the additional theoretical guarantees of variational methods.

Previous efforts at learning parameters for stereo models have used graph cuts to provide point estimates [15]. While recent work has shown that sequential tree-reweighted max-product message passing (TRW-S) has the ability to produce even better minimum energy solutions than graph cuts [16], max-product TRW-S also produces point estimates as opposed to approximate marginals.

The remainder of the paper is structured as follows. In section 2, we present a canonical conditional random field for the stereo vision problem. The canonical model is then augmented to explicitly account for occlusions. Next, we show how approximate inference is used for learning and to infer depth in an image. Section 3 shows how sparse variational message passing minimizes the KL divergence between a variational approximation and a distribution of interest. Results comparing sparse BP and VMP with graph cuts are given in section in section 4. Using variational distributions for learning improves results over the point estimate given by graph cuts, and sparse message passing can lead to an order of magnitude reduction in inference time. Furthermore, we show how learning parameters with our technique allows us to improve the quality of occlusion predictions in more richly structured CRFs.

2 Stereo vision and CRFs

The stereo vision problem is to estimate the *disparity* (horizontal displacement) at each pixel given a rectified pair of images. It is common in MRF-based stereo vision methods to work with energy functions of the form

$$F(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i} U(x_i, \boldsymbol{y}) + \sum_{i \sim j} V(x_i, x_j, \boldsymbol{y}) \tag{1}$$

where U is a *data term* that measures the compatibility between a disparity x_i and observed intensities $\boldsymbol{y}$, and V is a *smoothness term* between disparities at neighboring locations $i \sim j$ [17].

We construct a formal CRF probability model for stereo by normalizing the exponentiated F over all values for $\boldsymbol{X}$,

$$P\left(\boldsymbol{X} \mid \boldsymbol{y}\right) = \frac{1}{Z\left(\boldsymbol{y}\right)} \exp\left(-F(\boldsymbol{X},\boldsymbol{y})\right) \quad \text{with} \quad Z\left(\boldsymbol{y}\right) = \sum_{\boldsymbol{x}} \exp\left(-F(\boldsymbol{x},\boldsymbol{y})\right). \tag{2}$$

The normalizer $Z\left(\boldsymbol{y}\right)$ is typically referred to as the partition function.

2.1 A Canonical Stereo Model

The CRF of (2) is a general form. Here we present the specific CRF used for our experiments on stereo disparity estimation in section 4, following the model proposed by Scharstein and Pal [15]. The data term U simply measures the absolute intensity difference between corresponding pixels, summed over all color bands. We use the measure of Birchfield and Tomasi [18] for invariance to image sampling. The smoothness term V is a gradient-modulated Potts model [17,15] with $K = 3$ parameters:

$$V\left(x_i, x_j, \boldsymbol{y}\right) = \begin{cases} 0 & \text{if } x_i = x_j \\ \theta_k & \text{if } x_i \neq x_j \text{ and } g_{ij} \in B_k \end{cases} \tag{3}$$

Here g_{ij} is the color gradient between neighboring pixels i and j, and the B_k are three consecutive gradient bins with interval breakpoints from the set $\{0, 4, 8, \infty\}$. Let Θ_v denote all the parameters.

2.2 Occlusion Modeling

To account for occlusion, we create a model with an explicit occlusion state for the random variable associated with each pixel in the image. In our extended model $x_i \in \{0, \ldots, N-1\} \vee$ "occluded". The local data term U in our extended model has the form:

$$U\left(x_i, \boldsymbol{y}\right) = \begin{cases} c_i\left(x_i\right) & \text{if } x_i \neq \text{“occluded”} \\ \theta_o & \text{if } x_i = \text{“occluded”}, \end{cases} \tag{4}$$

where $c_i\left(x_i\right)$ is the Birchfield and Tomasi cost for disparity x_i at pixel i, as before. The new parameter θ_o is a local bias for predicting the pixel to be occluded.

We may also extend the gradient modulated smoothness terms to treat occluded states with a separate set of parameters such that:

$$V\left(x_i, x_j, \boldsymbol{y}\right) = \begin{cases} 0 & \text{if } x_i = x_j \text{ and } x_i \neq \text{“occluded”} \\ \theta_k & \text{if } x_i \neq x_j,\ g_{ij} \in B_k \text{ and both } x_i, x_j \neq \text{“occluded”} \\ \theta_{o,o} & \text{if } x_i = x_j \text{ and } x_i = \text{“occluded”} \\ \theta_{o,k} & \text{if } x_i \neq x_j,\ g_{ij} \in B_k \text{ and } x_i \text{ or } x_j = \text{“occluded”}. \end{cases} \tag{5}$$

2.3 Parameter Learning

Since the function $F(\boldsymbol{x}, \boldsymbol{y})$ is parameterized by $\Theta = (\theta_o, \Theta_v)$, these parameters may be learned in a maximum-likelihood framework with labeled training pairs. The objective function and gradient for one training pair $(\boldsymbol{x}, \boldsymbol{y})$ is

$$\begin{aligned} \mathcal{O}(\Theta) &= \log P(\boldsymbol{x} \mid \boldsymbol{y}; \Theta) && (6) \\ &= -F(\boldsymbol{x}, \boldsymbol{y}; \Theta) - \log Z(\boldsymbol{y}) && (7) \\ \nabla \mathcal{O}(\Theta) &= -\nabla F(\boldsymbol{x}, \boldsymbol{y}; \Theta) + \langle \nabla F(\boldsymbol{x}, \boldsymbol{y}; \Theta) \rangle_{P(\boldsymbol{X} \mid \boldsymbol{y}; \Theta)} . && (8) \end{aligned}$$

The particular factorization of $F(\boldsymbol{x}, \boldsymbol{y})$ in (1) allows the expectation in (8) to be decomposed into a sum of expectations over gradients of each term $U(x_i, \boldsymbol{y})$ and $V(x_i, x_j, \boldsymbol{y})$ using the corresponding marginals $P(X_i \mid \boldsymbol{y}; \Theta)$ and $P(X_i, X_j \mid \boldsymbol{y}; \Theta)$, respectively.

In previous work [15], graph cuts was used to find the most likely configuration of $\boldsymbol{X}$. This was taken as a point estimate of $P(\boldsymbol{X} \mid \boldsymbol{y}; \Theta_v)$ and used to approximate the gradient. Such an approach is potentially problematic for learning when the marginals are multi-modal or diffuse and unlike a delta function. Fortunately, a variational distribution $Q(\boldsymbol{X})$ can provide more flexible approximate marginals that may be used to approximate the gradient. We show in our experiments that using these marginals for learning is better than using a point estimate in situations when there is greater uncertainty in the model.

3 CRFs and Sparse Mean Field

In this section we derive the equations for *sparse* mean field inference using a variational message passing (VMP) perspective [8]. We show that sparse VMP will iteratively minimize the KL divergence between an approximation Q and the distribution P. Furthermore, we present sparse VMP in the context of CRFs and show that the functional we optimize is an upper bound on the negative log conditional partition function.

3.1 Mean Field

Here we briefly review the standard mean field approximation for a conditional distribution like (2). Let X_i be a discrete random variable taking on values x_i from a finite alphabet $\mathcal{X} = \{0, \ldots, N-1\}$. The concatenation of all random variables $\boldsymbol{X}$ takes on values denoted by $\boldsymbol{x}$, and the conditioning observation is $\boldsymbol{y}$. Variational techniques, such as mean field, minimize the KL divergence between an approximate distribution $Q(\boldsymbol{X})$ and the true distribution $P(\boldsymbol{X} \mid \boldsymbol{y})$. For the conditional distribution (2), the divergence is

$$\begin{aligned} \mathrm{KL}(Q(\boldsymbol{X}) \parallel P(\boldsymbol{X} \mid \boldsymbol{y})) &= \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \log \frac{Q(\boldsymbol{x})}{P(\boldsymbol{x} \mid \boldsymbol{y})} \\ &= \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \log \frac{Q(\boldsymbol{x}) Z(\boldsymbol{y})}{\exp(-F(\boldsymbol{x}, \boldsymbol{y}))} \\ &= \langle F(\boldsymbol{x}, \boldsymbol{y}) \rangle_{Q(\boldsymbol{X})} - H(Q(\boldsymbol{X})) + \log Z(\boldsymbol{y}) . && (9) \end{aligned}$$

The energy of a configuration $\boldsymbol{x}$ is $F(\boldsymbol{x}, \boldsymbol{y})$. We define a "free energy" of the variational distribution to be

$$\mathcal{L}(Q(\boldsymbol{X})) = \langle F(\boldsymbol{x}, \boldsymbol{y}) \rangle_{Q(\boldsymbol{X})} - H(Q(\boldsymbol{X})) . \tag{10}$$

Thus, the free energy is the expected energy under the variational distribution $Q(\boldsymbol{X})$, minus the entropy of Q. The divergence then becomes

$$\mathrm{KL}\left(Q(\boldsymbol{X}) \parallel P(\boldsymbol{X} \mid \boldsymbol{y})\right) = \mathcal{L}(Q(\boldsymbol{X})) + \log Z(\boldsymbol{y}) . \tag{11}$$

Since the KL divergence is always greater than or equal to zero, it holds that

$$\mathcal{L}(Q(\boldsymbol{X})) \geq -\log Z(\boldsymbol{y}) , \tag{12}$$

and the KL divergence is minimized at zero when the free energy equals the negative log partition function. Since $\log Z(\boldsymbol{y})$ is constant for a given observation, minimizing the free energy serves to minimize the KL divergence.

Mean field updates will minimize $\mathrm{KL}\left(Q(\boldsymbol{X}) \parallel P(\boldsymbol{X} \mid \boldsymbol{y})\right)$ for a factored distribution $Q(\boldsymbol{X}) = \prod_i Q(X_i)$. Using this factored Q, we can express our objective as

$$\begin{aligned} \mathcal{L}(Q(\boldsymbol{X})) &= \sum_{\boldsymbol{x}} \prod_i Q(x_i) F(\boldsymbol{x}, \boldsymbol{y}) + \sum_i \sum_{x_i} Q(x_i) \log Q(x_i) \\ &= \sum_{\boldsymbol{x}} Q(x_j) \langle F(\boldsymbol{x}, \boldsymbol{y}) \rangle_{\prod_{i:i \neq j} Q(X_i)} - H(Q(X_j)) - \sum_{i:i \neq j} H(Q(X_i)) , \end{aligned} \tag{13}$$

where we have factored out the approximating distribution $Q(X_j)$ for one variable, X_j. We form a new functional by adding Lagrange multipliers to constrain the distribution to sum to unity. This yields an equation for iteratively calculating an updated approximating distribution $Q^*(x_j)$ using the energy F and the distributions $Q(X_i)$ for other i:

$$Q^*(x_j) = \frac{1}{Z_j} \exp\left(- \langle F(\boldsymbol{x}, \boldsymbol{y}) \rangle_{\prod_{i:i \neq j} Q(X_i)} \right) , \tag{14}$$

where Z_j is a normalization constant computed for each update so that $Q^*(x_j)$ sums to one. See Weinman et al. [19] for the complete derivation of (14). Iteratively updating $Q(X_j)$ in this manner for each variable X_j will monotonically decrease the free energy $\mathcal{L}(Q(\boldsymbol{X}))$, thus minimizing the KL divergence.

3.2 Sparse updates

Variational marginals can be more valuable than graph cuts-based point estimates for accurate learning or other predictions. However, when the state space of the X_j is large, calculating the expectations within the mean field update (14) can be computationally burdensome. Here we show how to dramatically reduce the computational load of calculating updates when many states have a very low (approximate) probability. The sparse methods presented here represent

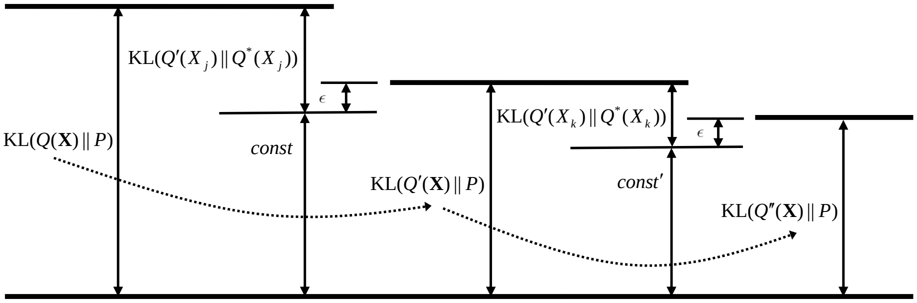

Fig. 1. Minimizing the global KL divergence via two different sparse local updates. The *global* divergence KL($Q(\mathbf{X}) \parallel P$) can be decomposed into a *local* update plus a constant: KL($Q^{'}(X_j) \parallel Q^{*}(X_j)$)+*const*. Consequently, at each step of sparse variational message passing we may minimize different local divergences to within some ϵ and when updating different local Qs, we minimize the global KL divergence.

a middle way between a fully-Bayesian approach and a simple point estimate. While the former considers all possibilities with their corresponding (often small) probabilities, the latter only considers the most likely possibility. Sparse updates provide a principled method for retaining an arbitrary level of uncertainty in the approximation.

The idea behind the sparse variational update is to eliminate certain values of x_j from consideration by making their corresponding variational probabilities $Q\left(x_j\right)$ equal to zero. Such zeros make calculating the expected energy for subsequent updates substantially easier, since only a few states must be included in the expectation. The eliminated states are those with low probabilities to begin with. Next we show how to bound the KL divergence between the original and sparse versions of $Q\left(X_j\right)$.

Given (11), (14), and (14) we can express $\mathrm{KL}\left(Q\left(\boldsymbol{X}\right) \parallel P\left(\boldsymbol{X} \mid \boldsymbol{y}\right)\right)$ as a function of a *sparse* update $Q^{'}\left(X_j\right)$, the original mean field update $Q^{*}\left(X_j\right)$ and the other $Q\left(X_i\right)$, where $i \neq j$:

$$\begin{aligned}\mathrm{KL}\left(Q\left(\boldsymbol{X}\right) \parallel P\left(\boldsymbol{X} \mid \boldsymbol{y}\right)\right) = {} & \mathrm{KL}\left(Q'\left(X_j\right) \parallel Q^{*}\left(X_j\right)\right) \\ & + \log Z_j + \log Z\left(\boldsymbol{y}\right) - \sum_{i: i \neq j} H\left(Q\left(X_i\right)\right). \end{aligned} \quad (15)$$

Since the last three terms of (15) are constant with respect to our update $Q^{'}\left(X_j\right)$, $\mathrm{KL}\left(Q\left(\boldsymbol{X}\right) \parallel P\left(\boldsymbol{X} \mid \boldsymbol{y}\right)\right)$ is minimized when $Q^{'}\left(X_j\right) = Q^{*}\left(X_j\right)$. At each step of *sparse* variational message passing, we will minimize $\mathrm{KL}(Q^{'}(\boldsymbol{X}_j) \parallel Q^{*}(\boldsymbol{X}_j))$ to within some small ϵ. As a result, each update to a different $Q\left(\boldsymbol{X}_j\right)$ yields further minimization of the *global* KL divergence. These relationships are illustrated in Figure 1.

If each X_j is restricted to a subset of values $x_j \in \mathcal{X}_j \subseteq \mathcal{X}$, we may define sparse updates $Q^{'}\left(X_j\right)$ in terms of the original update $Q^{*}\left(X_j\right)$ and the characeristic/indicator function $\mathbf{1}_{\mathcal{X}_j}(x_j)$ for the restricted range:

$$Q'(x_j) = \frac{\mathbf{1}_{\mathcal{X}_j}(x_j)}{Z'_j} Q^*(x_j)\,, \tag{16}$$

where the new normalization constant is

$$Z'_j = \sum_{x_j} Q'(x_j) = \sum_{x_j \in \mathcal{X}_j} Q^*(x_j)\,. \tag{17}$$

Thus, the divergence between a sparse update and the original is

$$\begin{aligned}
&\mathrm{KL}\left(Q'(X_j) \parallel Q^*(X_j)\right) && (18)\\
&= \sum_x \frac{\mathbf{1}_{\mathcal{X}_j}(x_j)}{Z'_j} Q^*(x_j) \log\left(\left(\frac{\mathbf{1}_{\mathcal{X}_j}(x_j)}{Z'_j} Q^*(x_j)\right) \Big/ \, Q^*(x_j)\right) \\
&= -\log Z'_j \frac{1}{Z'_j} \sum_{x \in \mathcal{X}_j} Q^*(x_j) \\
&= -\log Z_j^{'}. && (19)
\end{aligned}$$

As a consequence, it is straightforward and efficient to compute a maximally sparse $Q^{'}(X_j)$ such that $\mathrm{KL}\left(Q'(X_j) \parallel Q^*(X_j)\right) \leq \epsilon$ by sorting the $Q^*(x_j)$ values and performing a sub-linear search to satisfy the inequality. For example, if we wish to preserve 99% of the probability mass in the sparse approximation we may set $\epsilon = -\log 0.99 \approx .01$. Figure 1 illustrates the way in which sparse VMP iteratively minimizes the $\mathrm{KL}\left(Q(\boldsymbol{X}) \parallel P(\boldsymbol{X} \mid \boldsymbol{y})\right)$ after each iteration of message passing. In section 4 we show how using sparse messages can yield a dramatic increase in inference speed.

4 Experiments

In this section we present the results of two sets of experiments. The first compares sparse and traditional mean field methods for approximate inference, showing how sparse message passing can greatly accelerate free energy minimization. The second compares the performance of models learned using approximate marginals from both sparse mean field and a point estimate of the posterior marginals from graph cuts.

As training and test data we use 6 stereo pair images with ground-truth disparities from the 2005 scenes of the Middlebury stereo database [1]. These images are roughly 450×370 pixels and have discretized disparities with $N = 80$ states. Thus, when there are more than 600,000 messages of length N to send in any round of mean field updates for one image, shortening these to only a few states for most messages can dramatically reduce computation time.

[1] http://vision.middlebury.edu/stereo/data

4.1 Inference

The variational distribution $Q(\boldsymbol{X})$ provides approximate marginals $Q(X_i)$ that may be used for computing an approximate likelihood and gradient for training. These marginals are also used to calculate the mean field updates during free energy minimization. If these marginals have many states with very low probability, discarding them will have minimal effect on the update. First, we examine the need for sparse updates by evaluating the amount of uncertainty in these marginals. Then, we show how much time is saved by using sparse updates.

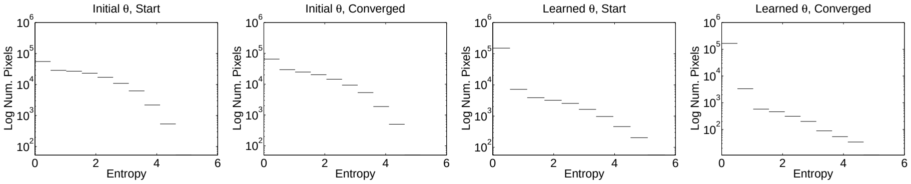

Fig. 2. Histograms of approximate marginal entropies $H(Q(X_i))$ from the variational distributions for each pixel at the start (after the first round) of mean field updates and at their convergence; values using the initial and learned parameters Θ_v of the canonical model are shown.

Our first set of experiments uses the simpler canonical stereo model having the smoothness term V of (3). Figure 2 shows histograms of the marginal entropies $H(Q(X_i))$ during free energy minimization with two sets of parameters, the initial parameters, $\Theta_v = \mathbf{1}$, and the learned Θ_v. We initialize the variational distributions $Q(X_i)$ to uniform and perform one round of VMP updates. Although most pixels have very low entropy, the initial model still has several variables with 2-4 "nats" (about 3-6 bits) of uncertainty. Once the model parameters are learned, the marginal entropies after one round of mean field updates are much lower. By the time the mean field updates converge and free energy is minimized, only a small percentage (less than three percent) have more than a half nat (less than two bits) of uncertainty. However, if point estimates are used, the uncertainty in these marginals will not be well represented. Sparse messages will allow those variables with low entropy to use few states, even a point estimate, while the handful of pixels with larger entropy may use more states.

The variational distribution has many states carrying low probability, even at the outset of training. We may greatly accelerate the update calculations by dropping these states according using (19) and our criterion. Figure 3 shows the free energy after each round of updates for both sparse and dense mean field. In all cases, sparse mean field has nearly reached the free energy minimum before one round of dense mean field updates is done. Importantly, the minimum free energy found with sparse updates is roughly the same as its dense counterpart.

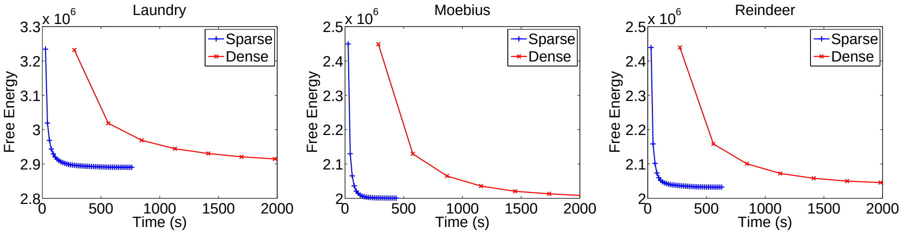

Fig. 3. Comparison of CPU time for free energy minimization with sparse and dense mean field updates using parameters Θ_v learned in the canonical model with three images (Art, Books, Dolls).

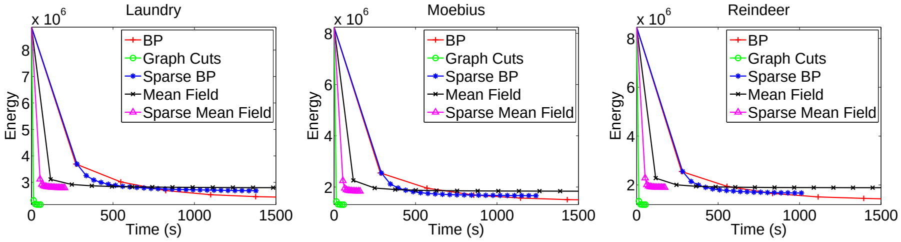

Fig. 4. CPU time versus energy for graph cuts, sum-product belief propagation, and mean field using parameters Θ_v learned with three images (Art, Books, Dolls). *Maximum posterior marginal* (MPM) prediction is used with the approximate marginal at each iteration.

As a comparison, we show in Figure 4 the true energy $F(\boldsymbol{x}, \boldsymbol{y})$ on several images during each iteration of several methods. It is important to note that only graph cuts explicitly minimizes this energy, but it is demonstrative of the relative speed and behavior of the methods.

4.2 Learning

As Figure 4 shows, graph cuts does a very good job of finding a minimum energy configuration. This is useful for making a prediction in a good model. However, maximizing the log likelihood (7) for learning requires marginals on the lattice. When the model is initialized, these marginals have higher entropy (Figure 2) representing the uncertainty in the model. At this stage of learning, the point estimate resulting from an energy minimization may not be a good approximation to the posterior marginals. In fact, using the graph cuts solution as a point estimate distribution having zero entropy, sparse mean field finds a lower *free* energy at the initial parameters $\Theta_v = \mathbf{1}$.

We compare the results of learning using two methods for calculating the gradient: sparse mean field and graph cuts. As demonstrated earlier, the model

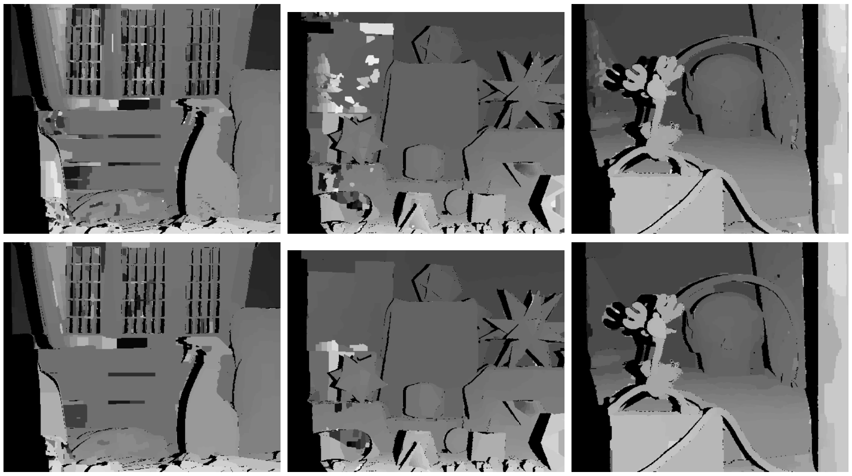

Fig. 5. Test images comparing prediction (using graph cuts) after one round of learning the canonical model with graph cuts (top) or sparse mean field (bottom). Occluded areas are black. Images (l-r): Laundry, Moebius, Reindeer.

has the highest uncertainty at the beginning of learning. It is at this point when sparse mean field has the greatest potential for improvement over graph cuts.

For learning, we use a small initial step size and a simple gradient descent algorithm with an adaptive rate. For prediction evaluation, we use graph cuts to find the most probable labeling, regardless of training method. We use leave-one-out cross validation on the six images.

After just one iteration, the training and test error with sparse mean field is markedly lower than that of the model trained with graph cuts for inference. Figure 5 shows the corresponding depth images after one iteration.

In Table 1, we compare the results of training using point estimates provided by graph cuts, as in previous work [15], and sparse mean field, the method proposed in this paper. We do not present results based on BP or dense mean field as training times are prohibitively long. For each experiment we leave out the image indicated and train on all the others listed. The disparity error is reduced by an average of $4.70 \pm 2.17\%$, and a paired sign test reveals the improvement is significant ($p < 0.05$).

We also test the error of our models' for occlusion predictions. We use the extended smoothness term (5) to handle the interactions between occluded states and the local terms of (4). We show both leave-one-out training and test results as well as the result of training on all the data (as a reference point). Models trained using sparse mean field give more accurate occlusion predictions than the model trained using graph cuts. In the gradient-modulated occlusion model our leave-one-out experiments show that the error in predicting occluded pixels is reduced an average of $4.94 \pm 1.10\%$ and is also significant ($p < 0.05$).

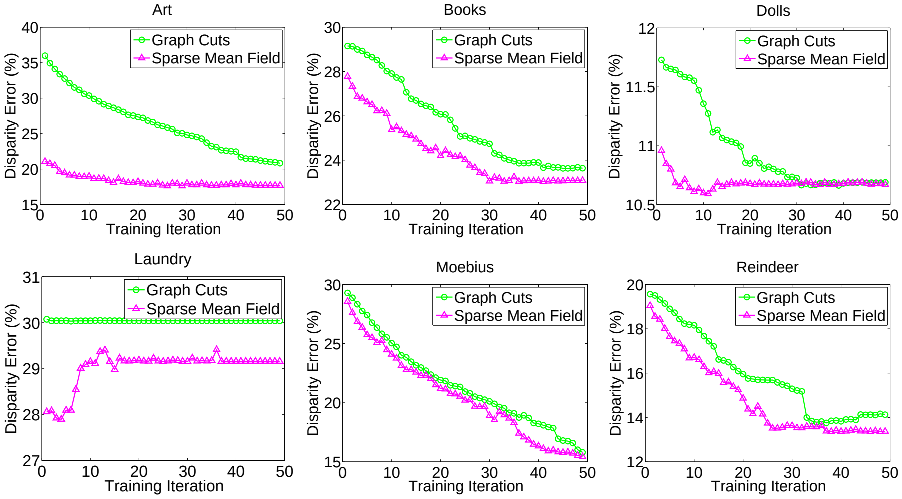

Fig. 6. Disparity error (each image held out in turn) using both graph cuts and mean field for learing the canonical CRF stereo model. The error before learning is omitted from the plots to better highlight performance differences.

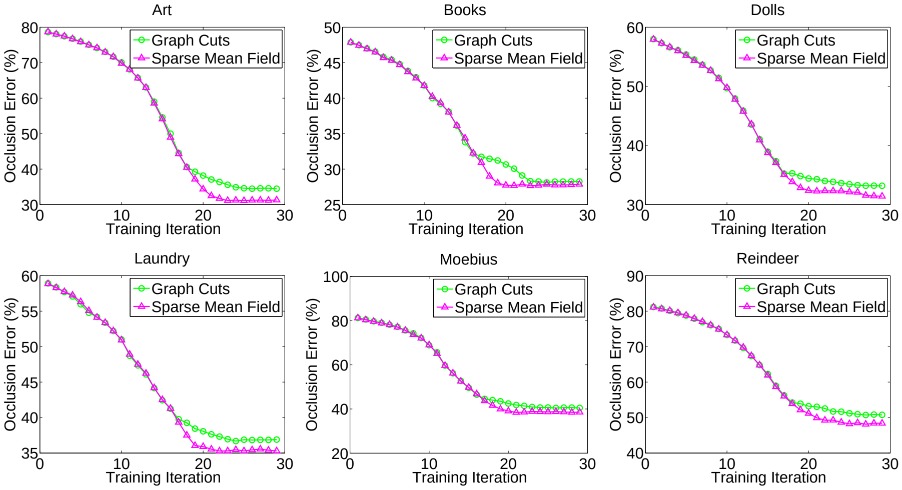

Fig. 7. Comparison of error predicting occluded pixel using graph cuts and sparse mean field for learning in the gradient-modulated occlusion model (5).

Table 1. Comparison of learning with graph cuts and sparse mean field. The disparity error (percentage of incorrectly predicted pixels) given for the canonical stereo model and the gradient-modulated occlusion model (with Eqs. (4) and (5)). For the gradient-modulated occlusion model we show the occlusion prediction error (percentage).

Metric	Method	Art	Books	Dolls	Laundry	Moebius	Reindeer	Average
Canonical Model - leave-one-out training & testing								
Disparity Error	Graph Cuts	20.83	23.64	10.69	30.04	15.80	14.13	19.17
	Sparse Mean Field	17.70	23.08	10.67	29.16	15.43	13.37	18.22
Gradient-Modulated Occlusion Model - leave-one-out training & testing								
Disparity Error	Graph Cuts	21.82	24.10	11.94	27.54	11.08	16.74	19.30
	Sparse Mean Field	21.05	23.14	11.62	27.37	11.45	16.44	18.93
Occlusion Error	Graph Cuts	34.50	28.27	32.99	36.89	40.65	50.83	37.36
	Sparse Mean Field	31.19	27.84	31.51	35.37	38.68	48.39	35.50
Gradient-Modulated Occlusion Model - trained on all (for comparison)								
Disparity Error	Graph Cuts	10.61	19.2	5.98	20.95	7.15	5.53	12.78
	Sparse Mean Field	8.29	13.41	4.72	19.22	5.11	4.76	10.15
Occlusion Error	Graph Cuts	16.20	10.40	24.88	29.77	27.88	32.97	21.83
	Sparse Mean Field	10.47	8.10	19.43	23.04	21.10	27.31	16.43

Figure 6 shows that sparse mean field reduces the disparity error in the model more quickly than graph cuts during learning on many images. Even when the two methods approach each other as learning progresses, sparse mean field still converges at parameters providing lower errors on both disparity and occlusions (Figure 7).

5 Conclusions

In this paper, we have provided a framework for sparse variational message passing (SVMP). Calculating sparse updates to the approximating variational distribution can greatly reduce the time required for inference in models with large state spaces. For high resolution imagery this reduction in time can be essential for practical inference and learning scenarios. In addition, we have a variational bound on the cost of our approximation. Furthermore, compare to graph cuts, the resulting marginals of SVMP provide better parameter estimates when used for learning in a maximimum likelihood framework. Graph cuts is often the best at finding a low energy solution in a given model. However, for model learning, a distribution over configurations is required. In models where there is more uncertainty (as in the early stages of learning), we fond that sparse mean field provides a lower free energy than graph cuts. As such, our analysis indicates that SVMP can be used as an effective tool for approximating the distributions necessary for accurate learning. Sparse mean field can be seen as a method occupying a middle ground between producing point estimates and creating fuller approximate distribution.

Finally, one of the most important advantages of the sparse mean field technique is that one no longer has strong constraints on the forms of allowable potentials that are required for graph cuts. As such, we see sparse message passing methods a being widely applicable for models where the constraints on potentials imposed by graph cuts are too restrictive.

References

1. Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions using graph cuts. In: Proc. ICCV. (2001) 508–515
2. Sun, J., Li, Y., Kang, S.B., Shum, H.Y.: Symmetric stereo matching for occlusion handling. In: Proc. CVPR. (2005) 399–406
3. Yang, Q., Wang, L., Yang, R., Stewenius, H., Nister, D.: Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling. In: Proc. CVPR. (2006)
4. Yedidia, J., Freeman, W., Weiss, Y.: Understanding belief propagation and its generalizations. In: Exploring Artificial Intelligence in the New Millennium. (January 2003) 239–236
5. Jordan, M.I., Ghahramani, Z., Jaakkola, T., Saul, L.: Introduction to variational methods for graphical models. Machine Learning **37** (1999) 183–233
6. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. of Machine Learning Research **3** (2003) 993–1022
7. Frey, B.J., Jojic, N.: A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE TPAMI **27**(9) (Sept 2005)
8. Winn, J., Bishop, C.: Variational message passing. J. of Machine Learning Research **6** (2005) 661–694
9. Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.: An introduction to MCMC for machine learning. Machine Learning **50** (2003) 5–43
10. Pal, C., Sutton, C., McCallum, A.: Sparse forward-backward using minimum divergence beams for fast training of conditional random fields. In: Proc. ICASSP. (2006) 581–584
11. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. ICML. (2001) 282–289
12. Kumar, S., Hebert, M.: Discriminative random fields. IJCV **68**(2) (2006) 179–201
13. Weinman, J., Hanson, A., McCallum, A.: Sign detection in natural images with conditional random fields. In: IEEE Intl. Workshop on Machine Learning for Signal Processing. (2004) 549–558
14. He, Z., Zemel, R.S., Carreira-Perpin, M..: Multiscale conditional random fields for image labeling. In: Proc. CVPR. (2004) 695–702
15. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: Proc. CVPR. (2007)
16. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE TPAMI **28** (2006) 1568–1583
17. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE TPAMI **23**(11) (2001) 1222–1239
18. Birchfield, S., Tomasi, C.: A pixel dissimilarity measure that is insensitive to image sampling. IEEE TPAMI **20**(4) (1998) 401–406
19. Weinman, J., Pal, C., Scharstein, D.: Sparse message passing and efficiently learning random fields for stereo vision. Technical Report UM-CS-2007-054, U. Massachusetts Amherst (Oct. 2007)