Int J Comput Vis (2012) 99:319–337
DOI 10.1007/s11263-010-0385-z

On Learning Conditional Random Fields for Stereo

Exploring Model Structures and Approximate Inference

Christopher J. Pal · Jerod J. Weinman · Lam C. Tran · Daniel Scharstein

Received: 22 March 2010 / Accepted: 15 September 2010 / Published online: 20 October 2010

Abstract Until recently, the lack of ground truth data has hindered the application of discriminative structured prediction techniques to the stereo problem. In this paper we use ground truth data sets that we have recently constructed to explore different model structures and parameter learning techniques. To estimate parameters in Markov random fields (MRFs) via maximum likelihood one usually needs to perform approximate probabilistic inference. Conditional random fields (CRFs) are discriminative versions of traditional MRFs. We explore a number of novel CRF model structures including a CRF for stereo matching with an explicit occlusion model. CRFs require expensive inference steps for each iteration of optimization and inference is particularly slow when there are many discrete states. We explore belief propagation, variational message passing and graph cuts as inference methods during learning and compare with learning via pseudolikelihood. To accelerate approximate inference we have developed a new method called sparse variational message passing which can reduce inference time by an order of magnitude with negligible loss in quality. Learning using sparse variational message passing improves upon previous approaches using graph cuts and allows efficient learning over large data sets when energy functions violate the constraints imposed by graph cuts.

Keywords Stereo · Learning · Structured prediction · Approximate inference

This research was supported in part by: a Google Research award, Microsoft Research through awards under the eScience and Memex funding programs, a gift from Kodak Research and an NSERC discovery award to C.P. Support was also provided in part by NSF grant 0413169 to D.S.

C.J. Pal (✉)
École Polytechnique de Montréal, Montréal, QC, Canada
e-mail: christopher.pal@polymtl.ca

J.J. Weinman
Dept. of Computer Science, Grinnell College, Grinnell, IA, USA
e-mail: weinman@grinnell.edu

L.C. Tran
Dept. of Electrical and Computer Engineering, University of California San Diego, San Diego, CA, USA
e-mail: lat003@ucsd.edu

D. Scharstein
Middlebury College, Middlebury, VT, USA
e-mail: schar@middlebury.edu

1 Introduction

In recent years, machine learning methods have been successfully applied to a large number of computer vision problems, including recognition, super-resolution, inpainting, texture segmentation, denoising, and context labeling. Stereo vision has remained somewhat of an exception because of the lack of sufficient training data with ground-truth disparities. While a few data sets with known disparities are available, until recently they had been mainly been used for benchmarking of stereo methods (e.g., Scharstein and Szeliski 2002). Our earlier work in this line of research (Scharstein and Pal 2007) sought to remedy this situation by replacing the heuristic cues used in previous approaches with probabilistic models for structured prediction derived from learning using real images and ground truth stereo imagery. To obtain a sufficient amount of training data, we used the structured-lighting approach of Scharstein and Szeliski (2003) to construct a database of 30 stereo pairs

with ground-truth disparities, which we have made available for use by other researchers.[1]

By addressing the need for greater quantities of ground truth data, we are now able to take a machine learning approach using a classical structured prediction model, the conditional random field (CRF). We derive a gradient-based learning approach that leverages efficient graph-cut minimization methods and our ground-truth database. We then explore the characteristics and properties of a number of different models when learning model parameters. Using graph-cut minimization techniques for gradient-based learning in CRFs corresponds to an aggressive approximation of the underlying probabilities needed for expectations of key quantities. In this work we further explore the issues of learning and its interaction with different inference techniques under richer model structures.

Among the few existing learning approaches for stereo, one of the most prominent is the work by Zhang and Seitz (2005), who iteratively estimate the global parameters of an MRF stereo method from the previous disparity estimates and thus do not rely on ground-truth data. Kong and Tao (2004) learn to categorize matching errors of local methods using the Middlebury images. Kolmogorov et al. (2006) construct MRF models for binary segmentation using locally learned Gaussian Mixture Models (GMMs) for foreground and background colors. Some interesting recent work has explored learning in a hidden variable CRF-like model for stereo (Trinh and McAllester 2009). They formulate stereo as the problem of modeling the probability of the right image given the left. Thus, they are able to construct a conditional model with hidden variables for depth information. As such, they do not use ground-truth depth information and cast the approach as an instance of unsupervised learning. They use monocular texture cues, define potential functions on segments from image segmentation and construct an energy function based on a slanted-plane model. They perform learning using a variation of hard assignment conditional expectation maximization.

Learning aside, there has been growing interest in simply creating richer models for stereo vision in which more parameters are introduced to produce more accurate results. In particular, recent activity has focused on explicitly accounting for occlusions in stereo vision models. For example, Kolmogorov and Zabih (2001) have directly incorporated occlusion models in an energy function and graph-cut minimization framework. Sun et al. (2005) 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. (2006) have achieved impressive results building on models that estimate depth in both left and right images 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. All of these methods involve creating richer models to obtain greater disparity accuracy. Thus, we see a growing need to learn or estimate model parameters in an efficient and principled way.

While learning for stereo is growing in interest, much recent progress in stereo vision has been achieved along two other avenues. First, global optimization methods have become practical with the emergence of powerful optimization techniques. Considered too slow when first proposed by Barnard (1989), global methods currently dominate the top of the Middlebury stereo rankings. In particular, MRF models for stereo have become popular since high-quality approximate solutions can be obtained efficiently using graph cuts (Boykov et al. 2001; Kolmogorov and Zabih 2001, 2002b) and belief propagation (Sun et al. 2003, 2005; Felzenszwalb and Huttenlocher 2006). Tappen and Freeman (2003) have compared graph cuts and belief propagation for stereo and Szeliski et al. (2008) have compared a larger set of MRF energy minimization techniques, providing software that we use in our implementation.

A second breakthrough has been the realization of the importance of intensity changes as a cue for object boundaries (i.e., disparity discontinuities). Taken to an extreme, this translates into the assumption that disparity jumps always coincide with color edges, which is the basis of a large number of recent segment-based stereo methods (Tao et al. 2001; Zhang and Kambhamettu 2002; Bleyer and Gelautz 2004; Hong and Chen 2004; Wei and Quan 2004; Zitnick et al. 2004; Sun et al. 2005). Such methods start with a color segmentation and then assume that disparities are constant, planar, or vary smoothly within each segment. This assumption works surprisingly well if the segments are small enough. Alternatively, color segmentations can also be employed as smoothness priors in pixel-based approaches (Sun et al. 2003). Using color segmentations is not the only way to utilize this monocular cue; many pixel-based global methods also change the smoothness cost (i.e., penalty for a disparity change) if the local intensity gradient is high (Boykov et al. 2001; Kolmogorov and Zabih 2002a; Scharstein and Szeliski 2002). This is the approach we take. The relationship between intensity gradient and smoothness cost is learned from real images.

We have focused our discussion so far on discrete formulations for stereo. However, continuous formulations also exist and a number of groups have formulated the continuous depth stereo problem as an energy functional representing an underlying partial differential equation (PDE) (Alvarez et al. 2002; Strecha et al. 2003). More recent work along these lines has cast occlusions as unobserved hidden variables and used expectation maximization (Strecha et

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

al. 2004). This work also draws closer together PDE-based methods and maximum a posteriori (MAP) estimation. As noted by Yang et al. (2006), more studies are needed to understand the behavior of algorithms for optimizing parameters in stereo models. This work addresses that need.

Some of the discrete stereo models discussed above have formulated the problem directly as an energy function without an explicit probabilistic model. When a probabilistic model has been used, it has been a joint or generative random field. However, there are well-known performance advantages to using discriminative as opposed to generative modeling techniques (Ng and Jordan 2002). One of our contributions is the development of a completely probabilistic and discriminative discrete formulation for stereo. We explicitly model occlusions using additional states in the variables of a conditional random field (CRF). 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.

1.1 Conditional Random Fields, Learning and Inference

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, since we cast the stereo problem as a conditional probability model, we are able to view learning as an instance of maximum conditional likelihood. We can also draw from recent insights in the machine learning community to deal with learning in intractable probability models. In this light, learning also becomes a task closely linked to the quality of approximate inference in the model. From this formulation we are able to develop a probabilistic variational method in the sense of Jordan et al. (1999). While we focus on approximate inference and learning in lattice-structured conditional random fields applied to stereo vision, our theoretical results and some experimental insights are applicable to CRFs, MRFs and Bayesian networks with arbitrary structures.

The CRF approach to modeling and learning a random field was first presented for sequence processing problems by Lafferty et al. (2001). Sutton and McCallum (2006) give a good review of modeling and learning techniques for CRFs focusing on natural language processing problems and optimization methods that exploit second order information. Lafferty et al. (2001) originally proposed the use of a method known as improved iterative scaling (Della Pietra et al. 1997) for learning CRFs. However, optimization of the conditional log likelihood using classical gradient-descent-based methods is a natural strategy for learning in CRFs. Indeed, several authors have reported significant increases in learning speed using second-order gradient-based optimization techniques. Quasi-Newton methods such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method or limited-memory versions of the BFGS method have been particularly successful (Sutton and McCallum 2006). More recently, Vishwanathan et al. (2006) have reported even faster convergence with large data sets using Stochastic Meta-Descent, a stochastic gradient optimization method with adaptation of a gain vector.

Model expectations are needed for gradient-based learning. To efficiently compute these in the linear chain structured models commonly used in language processing, a subtle variation of the well-known forward-backward algorithm for hidden Markov models can be used. However, approximate inference methods must be used for many graphical models with more complex structure. The dynamic conditional random fields (DCRFs) of Sutton et al. (2004) use a factorized set of variables at each segment of a linear-chain CRF. This leads to a shallow but dynamically-sized lattice-structured model. Sutton et al. (2004) explore several methods for approximate inference and learning, including tree-based reparameterization (TRP) or the tree-based message passing schedules of Wainwright et al. (2002, 2003) and the loopy belief propagation strategy discussed in Murphy et al. (1999) under a random schedule. Other work such as Weinman et al. (2004) also explores TRP methods but not in a large lattice structured model. For random fields with a hidden layer in the form of a Boltzmann machine, He et al. (2004) have used sampling methods for inference based on contrastive divergence. Contrastive divergence initializes Markov chain Monte Carlo (MCMC) sampling using the data and then takes a few steps of a Gibbs sampler. This approach is faster than traditional MCMC which require convergence to equilibrium. However it can lead to crude approximations to model likelihood gradients used for learning. Kumar and Hebert (2006) optimize the parameters of lattice-structured binary CRFs using pseudolikelihood and perform inference using iterated conditional modes ICM, which is fast but well know to get caught in local minima. Other work by Blake et al. (2004) has investigated the discriminative optimization of lattice-structured joint random field models using autoregression over the pseudolikelihood.

Pseudolikelihood-based techniques are equivalent to using spatially localized and independent probabilistic models as a substitute for the original global model of a complete joint distribution. Pseudolikelihood can yield a convex optimization problem that often leads to relatively fast optimization times. However, Liang and Jordan (2008) have shown that pseudolikelihood can give poorer estimates of interaction parameters in random fields when interactions are strong. In Sect. 5.6 we compare learning with pseudolikelihood, graph cuts and our new inference method (sparse

mean field), and we find that pseudolikelihood indeed results in the lowest performance. We are therefore motivated to use a learning procedure that accounts for the global view of the underlying structured prediction problem.

1.2 Graph Cuts and Learning

The primary challenge with learning a CRF using a standard gradient-descent-based optimization of the conditional likelihood is that one must compute intractable model expectations of features in the energy function (see Sect. 4). One solution to this problem is to replace distributions needed for the expectation with a single point estimate and compute gradients in a manner reminiscent of the classical perceptron algorithm. For the types of models we explore here, graph-cut-based methods are typically the fastest choice for energy minimization Szeliski et al. (2008). Thus, one particularly attractive solution to the learning problem is to take advantage of the extremely fast and high-quality performance of graph cuts. In more precise terms, this energy corresponds to a *most-probable-explanation* (MPE) (Cowell et al. 2003) estimate for the corresponding CRF. Although it does have important limitations, we use this fast and effective strategy for our initial explorations of model structures.

The maximum conditional likelihood formulation for gradient-based learning in a CRF requires one to compute model *expectations*, not MPE estimates. Furthermore, the graph cut algorithm only works if energy functions satisfy certain conditions. While the original energy function of a random field can have negative weights, the secondary graph constructed when performing graph-cut inference must have non-negative edge weights. This transformation leads to inequality constraints on the original energy function. These constraints also imply that graph-cut inference may cease to be possible during the course of learning for some models—something we have observed during our experiments. These factors have motivated us to explore a second class of inference techniques, based on quickly computing approximate marginal distributions during learning. Thus, in the second broad area of our exploration we compare the efficiency and quality of global inference techniques during learning.

1.3 Other Alternatives for Inference

It is well known that belief propagation (BP) (Yedidia et al. 2003) in tree structured graphs is exact. However, in stereo and many other problems in computer vision one frequently uses graphs defined on a 2D grid. It is possible to apply BP on graphs with loops using loopy belief propagation (LBP), and there have been a number of reports of success using this strategy for applications ranging from error-correcting codes (Frey and MacKay 1997) and inference in large Bayesian networks (Murphy et al. 1999) to low level vision (Felzenszwalb and Huttenlocher 2006). Two important variations of belief propagation consist of the sum-product algorithm and max-product algorithm (Kschischang et al. 2001). The sum-product algorithm is used to compute marginal distributions while the max-product algorithm is used to give the most probable configuration or MPE under a model. In a generative model this is also equivalent to the MAP configuration. The max-product variation of BP is equivalent to the celebrated "Viterbi" algorithm used for decoding in hidden Markov models. In a 2D lattice, loopy variants of max-product can be used to find configurations that correspond to approximate energy minima. The evaluation of Szeliski et al. (2008) includes comparisons with variations of loopy max-product BP but typically finds superior minima using methods based on either graph cuts or BP variants that involve more sophisticated tree-based approximations (Wainwright et al. 2005), which can also be viewed as linear programming relaxations.

Tree-based approximations can be used to improve both the quality of marginal inference and the quality of MAP or MPE estimates. A class of algorithms known as tree-based reparameterization (TRP) (Wainwright et al. 2003) can be used to obtain approximate marginals in a graph with cycles. This class of algorithms can be formulated as a series of reparameterization updates to the original loopy graph. Tree-reweighted message passing (TRW) (Wainwright et al. 2005) is an approach whereby one reweights the usual messages of LBP. This family of algorithms involves reparameterizing a collection of tree-structured distributions in terms of a common set of pseudo-max-marginals on the nodes and edges of the graph with cycles. When it is possible to find a configuration that is locally optimal with respect to every single node and edge pseudo-max-marginal, then the upper bound is tight, and the MAP configuration can be obtained. Recent work (Kolmogorov 2006) has shown that more sophisticated algorithms, such as sequential tree-reweighted max-product message passing (TRW-S), have the ability to produce even better minimum energy solutions than graph cuts.

Belief propagation (Yedidia et al. 2003) and variational methods (Jordan et al. 1999) are both widely used techniques for inference in probabilistic graphical models and are known for being reasonably fast and easy to implement in a memory-efficient manner. Both techniques have been used for inference and learning in models with applications ranging from text processing (Blei et al. 2003) to computer vision (Frey and Jojic 2005). Winn and Bishop (2005) proposed Variational Message Passing (VMP) 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, unlike BP, VMP optimizes a lower bound on the log probability of observed variables

in a generative model. Variational inference thus has a more direct connection to the probability of data under a model when an underlying graphical structure contains cycles.

Experimental and theoretical analysis of variational methods has shown that while the asymptotic performance of other methods such as sampling (Andrieu et al. 2003) can be superior, frequently variational methods are faster for approximate inference (Jordan et al. 1999). 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. Some of our previous work (Pal et al. 2006) has explored sparse methods for approximate inference using BP in chain-structured graphs, in loopy graphs (Weinman et al. 2009), and 2D grids (Weinman et al. 2008).

We focus our explorations of marginal inference for learning in this paper on approximate marginal inference using BP, variational mean field, and sparse variants of these methods. We find that the sparse variants make most tasks dramatically more practical, reducing training times by an order of magnitude. We focus our theoretical analysis on a new method we call sparse variational message passing (SVMP). The method combines the theoretical benefits of variational methods with the time-saving advantages of sparse messages. The state space in stereo models can become quite large if one seeks to account for many possible discretized disparity levels. Thus, we believe that the sparse learning techniques we propose here will be an important contribution. While we do not explore it here, we note that sparse inference methods for tree-based approximation techniques (Wainwright et al. 2005) or structured mean field methods (Jordan et al. 1999) could be a promising direction for future research. This paper explores a broader and richer set of model structures compared to our earlier work (Scharstein and Pal 2007). We also expand upon our earlier experimental analysis in Weinman et al. (2008, 2007) including a comparison with pseudolikelihood.

The remainder of the paper is structured as follows. First, in Sect. 2 we describe the new stereo data sets we use as ground truth for our learning experiments. In Sect. 3, we develop a number of different CRFs architectures with different levels of complexity. We begin with a probabilistic CRF version of a canonical model for stereo vision problem in Sect. 3.1. The canonical model is then augmented to explore: modulation terms that are dependent upon disparity differences (Sect. 3.2), models that use patches to compute local costs (Sect. 3.3), and models that explicitly account for occlusions (Sect. 3.4). In Sect. 4, we present the key challenge in gradient-based learning and motivate how different types of approximate inference can be used to approximate a key intractable expectation. In Sect. 4.2, we then review classical mean field updates and show how sparse variational message passing can accelerate inference. In Sect. 5, we present results using graph cuts for learning in the different model architectures we have discussed. We then present results comparing sparse BP and VMP with graph cuts. Through this we see how using variational distributions for learning improves results over the point estimate given by graph cuts and observe how sparse message passing can lead to an order of magnitude reduction in inference time compared to dense message passing. Finally, we show how learning parameters with our technique allows us to improve the quality of occlusion predictions in more richly structured CRFs.

2 Data Sets

In order to obtain a significant amount of training data for stereo learning approaches, we have created 30 new stereo data sets with ground-truth disparities using an automated version of the structured-lighting technique of Scharstein and Szeliski (2003). Our data sets are available for use by other researchers.[2] Each data set consists of 7 rectified views taken from equidistant points along a line, as well as ground-truth disparity maps for viewpoints 2 and 6. The images are about 1300×1100 pixels (cropped to the overlapping field of view), with about 150 different integer disparities present. Each set of 7 views was taken with three different exposures and under three different lighting conditions. We thus have 9 different images from 7 different viewpoints. These images exhibit significant radiometric differences and can be used to test for robustness to violations of the brightness constancy assumption, which are common in real-world applications.

For the work reported in this paper we only use the six data sets shown in Fig. 1: Art, Books, Dolls, Laundry, Moebius and Reindeer. As input images we use a single image pair (views 2 and 6) taken with the same exposure and lighting. To make the images manageable by the graph-cut stereo matcher, we downsample the original images to one third of their size, resulting in images of roughly 460×370 pixels with a disparity range of 80 pixels. The resulting images are still more challenging than standard stereo benchmarks such as the Middlebury Teddy and Cones images, due to their larger disparity range and higher percentage of untextured surfaces.

3 Stereo Vision and CRFs

The classical formulation of the stereo vision problem is to estimate the *disparity* (horizontal displacement) at each

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

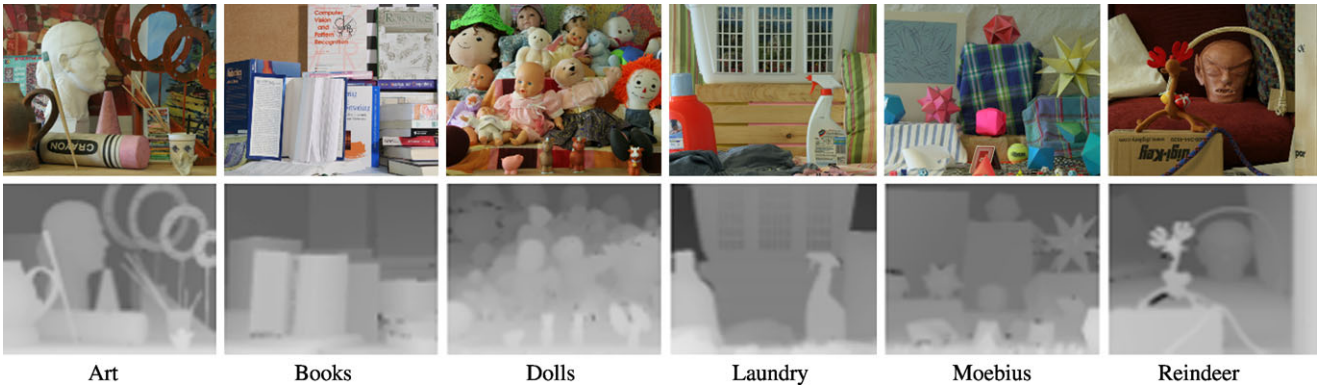

Fig. 1 The six data sets used in this paper. Shown is the left image of each pair and the corresponding ground-truth disparities (©2007 IEEE)

pixel given a rectified pair of images. It is common in MRF-based stereo vision methods to work with energy functions of the form

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

where U is a *data term* that measures the compatibility between a disparity x_i and observed intensities $\mathbf{y}$, and V is a *smoothness term* between disparities at neighboring locations $i \sim j$ (Boykov et al. 2001).

We construct a CRF for stereo by conditionally normalizing the exponentiated F over all possible values for each x_i and for each pixel location i in the image. More formally, 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 $\mathbf{X}$ takes on values denoted by $\mathbf{x}$. If we denote the conditioning observation in our model as $\mathbf{y}$, we can then express our CRF as

$$P(\mathbf{X}=\mathbf{x}\mid\mathbf{y}) = \frac{1}{Z(\mathbf{y})}\exp\left(-F(\mathbf{x},\mathbf{y})\right), \tag{2}$$

with

$$Z(\mathbf{y}) = \sum_{\mathbf{x}} \exp\left(-F(\mathbf{x},\mathbf{y})\right). \tag{3}$$

The normalizer $Z(\mathbf{y})$ is typically referred to as the partition function. It is useful to note that a key distinction between a CRF and a jointly defined MRF is that the partition function of an MRF does not depend on the observation y and normalizes a joint distribution over the random variables X and a set of random variables Y defined for y. When using our model to create a depth map from a stereo pair, our goal is to find an assignment to $\mathbf{X}$ minimize the negative log probability

$$-\log P(\mathbf{x}|\mathbf{y}) = \log Z(\mathbf{y}) + \sum_i U(x_i,\mathbf{y}) + \sum_{i\sim j} V(x_i,x_j,\mathbf{y}). \tag{4}$$

Note that our formulation, unlike other energy-based stereo approaches, explicitly accounts for a data dependent partition function. Furthermore, following the typical formulation of CRFs, we express cost terms U and pairwise smoothness terms V using a linear combination of feature functions f_u, f_v, which gives us

$$U(\mathbf{x},\mathbf{y}) = \sum_u \theta_u f_u(\mathbf{x},\mathbf{y}), \tag{5}$$

$$V(\mathbf{x}_i,\mathbf{x}_j,\mathbf{y}) = \sum_v \theta_v f_v(\mathbf{x}_i,\mathbf{x}_j,\mathbf{y}), \tag{6}$$

where θ_u, θ_v are the parameters of our model. The notation follows the usual format for specifying the potential functions of CRFs (Lafferty et al. 2001; Sutton and McCallum 2006), and the linear form allows us to derive an intuitive gradient-based minimization procedure for parameter estimation.

3.1 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 Sect. 5, following the model proposed by Scharstein and Pal (2007). The data term U is given by

$$U(x_i,\mathbf{y}) = c(i,x_i,\mathbf{y}), \tag{7}$$

where c simply measures the absolute intensity difference between the corresponding pixels of the images, as indicated by i and x_i. We use the difference measure of Birchfield and Tomasi (1998) summed over all color bands for invariance to image sampling.

The smoothness term V is a gradient-modulated Potts model (Boykov et al. 2001; Scharstein and Pal 2007) with K parameters:

$$V(x_i,x_j,\mathbf{y}) = \begin{cases} 0, & \text{if } x_i = x_j \\ \theta_k, & \text{if } x_i \neq x_j \text{ and } g_{ij} \in B_k. \end{cases} \tag{8}$$

Here, g_{ij} is the color gradient or root mean square color difference between neighboring pixels i and j. The values B_k represent discretized intervals the gradient belongs to for the purposes of modulating the smoothness penalty. Interval breakpoints may be chosen from different sets. For example, in our initial experiments we explore subsets of $\{0, 2, 4, 8, 12, 16, \infty\}$. Let Θ_v denote all the smoothness parameters.

3.2 Disparity Difference Dependent Modulation

Interaction potentials that take into account the difference in disparities between pixels have been of considerable interest in the past. Felzenszwalb and Huttenlocher (2006) have explored parametric forms for this interaction such as $V(x_i, x_j, \mathbf{y}) = c|x_i - x_j|$ or $V(x_i, x_j, \mathbf{y}) = c(x_i - x_j)^2$. However, our framework allows us to *learn* the functional form of such interactions. To explore other aspects of smoothness modulation, we shall investigate models with interaction terms as a more general function of disparity changes, e.g., $V(x_i, x_j, \mathbf{y}) = f(|x_i - x_j|)$. We are able to achieve this in a manner similar to our gradient discretization approach by discretizing the absolute disparity differences $d_{ij} = |x_i - x_j|$ into bins C_l and defining feature functions that are active on the jointly discretized disparity difference bins C_l and gradient bins B_k such that

$$V(x_i, x_j, \mathbf{y}) = \theta_{kl} \quad \text{if } g_{ij} \in B_k \text{ and } d_{ij} \in C_l. \tag{9}$$

3.3 Patch Matching

While pixel to pixel intensity matching is often an effective strategy for stereo matching, modern cameras are usually able to produce images at a resolution much higher than one might need for the corresponding disparity map. We thus explore matching patches in a pair of high resolution images to compute our local cost terms that will be used for inferring disparity at lower resolution. The pair of high resolution images are partitioned into $n \times n$ patches. The new data term U uses the same Birchfield and Tomasi costs (Birchfield and Tomasi 1998) over the color channels as (7), except it must now sum the costs over all corresponding pixels in the high resolution image patches indicated by i and x_i.

We explore this model in cases where the smoothness term V is defined similarly to our simple, canonical stereo model of (8). However the color gradient g_{ij} between neighboring locations i and j is divided by the size of the patch.

In our experiment, we used the full size color images of roughly 1380×1110 pixels and the one-third size ground truth disparity maps of roughly 460×370 to train and test our model. Thus, the resolution we have selected for the rest of our experiments allows us to use a patch size of 3×3.

3.4 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 this extended model we use $x_i \in \{0, \ldots, N-1\} \cup$ "occluded". The local data term U in the extended model has the form:

$$U(x_i, \mathbf{y}) = \begin{cases} c(i, x_i, y), & \text{if } x_i \neq \text{"occluded"} \\ \theta_o, & \text{if } x_i = \text{"occluded"}, \end{cases} \tag{10}$$

where $c_i(x_i, y)$ 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:

$$V(x_i, x_j, \mathbf{y}) = \begin{cases} 0, & \text{if } x_i = x_j \text{ and } x_i \neq \text{"occluded"} \\ \theta_k, & \text{if } x_i \neq x_j,\ g_{ij} \in B_k \text{ and both } x_i, x_j \neq \text{"occluded"} \\ \theta_{o,o}, & \text{if } x_i = x_j \text{ and } x_i = \text{"occluded"} \\ \theta_{o,k}, & \text{if } x_i \neq x_j,\ g_{ij} \in B_k \text{ and } x_i \text{ or } x_j = \text{"occluded"}. \end{cases} \tag{11}$$

4 Parameter Learning

The energy function $F(\mathbf{x}, \mathbf{y})$ in our models is parameterized by $\Theta = (\Theta_u, \Theta_v)$, where Θ_u denotes the data term parameters and Θ_v denotes the smoothness term parameters. These parameters may be learned in a maximum conditional likelihood framework with labeled training pairs. The objective function and gradient for one training pair $(\mathbf{x}, \mathbf{y})$ can be expressed as the minimization of

$$\mathcal{O}(\Theta) = -\log P(\mathbf{x} \mid \mathbf{y}; \Theta) \tag{12}$$

$$= F(\mathbf{x}, \mathbf{y}; \Theta) + \log Z(\mathbf{y}) \quad \text{with} \tag{13}$$

$$\nabla \mathcal{O}(\Theta) = \nabla F(\mathbf{x}, \mathbf{y}; \Theta) - \langle \nabla F(\mathbf{x}, \mathbf{y}; \Theta) \rangle_{P(\mathbf{X} \mid \mathbf{y}; \Theta)}, \tag{14}$$

and where $\langle \cdot \rangle_{P(\mathbf{X} \mid \mathbf{y}; \Theta)}$ denotes an expectation under the models conditional distribution over $\mathbf{X}$. It is known that the CRF loss function is convex for fully observed states (Lafferty et al. 2001). However, in 2D grid lattices such as the ones we consider here we have a critical, but intractable expectation in (14). But, the particular factorization of $F(\mathbf{x}, \mathbf{y})$ in (1) allows the expectation in (14) to be decomposed into a sum of expectations over gradients of each term $U(x_i, \mathbf{y})$ and $V(x_i, x_j, \mathbf{y})$ using the corresponding single node and pairwise marginals $P(X_i \mid \mathbf{y}; \Theta)$ and $P(X_i, X_j \mid \mathbf{y}; \Theta)$, respectively. In this context we can view the main computational challenge in learning as the task of computing good approximations to these marginals. Approximation of these

key expectations through computing approximate marginals is thus the crux of our exploration here.

For our experiments, we use a simple gradient descent approach for learning. Our (approximate) gradients can be very noisy, which can cause problems for second order methods such as BFGS. We also have a small number of training examples, so learning methods such as stochastic gradient, which creates small batches of training data to accelerate learning, are not well-suited for our setting. Our experiments will focus on comparing different approximate inference techniques using a straightforward learning algorithm, minimizing potential interactions between more sophisticated learning approaches and the consequences of using approximate distributions.

In previous work by Scharstein and Pal (2007), graph cuts were used to find the most likely configuration of $\mathbf{X}$. This was taken as a point estimate of $P(\mathbf{X} \mid \mathbf{y}; \Theta)$ and used to approximate the gradient. Such an approach is potentially problematic for learning when the marginals are more uniform or contain a number of solutions with similar probability and look unlike a single delta function. Fortunately, a variational distribution $Q(\mathbf{X})$ 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.

We now derive the equations for *sparse* mean field inference using a variational message passing (VMP) perspective (Winn and Bishop 2005). Sparse VMP iteratively minimizes the KL divergence between an approximation Q and the distribution P. In the context of CRFs, the functional optimized by sparse VMP is an upper bound on the negative log conditional partition function.

4.1 Mean Field

Here we briefly review the standard mean field approximation for a conditional distribution like (2). As before we 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 $\mathbf{X}$ takes on values denoted by $\mathbf{x}$, and the conditioning observation is $\mathbf{y}$. Variational techniques, such as mean field, minimize the KL divergence between an approximate distribution $Q(\mathbf{X})$ and the true distribution $P(\mathbf{X} \mid \mathbf{y})$. For the conditional distribution (2), the divergence is

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

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

$$\mathcal{L}\left(Q\left(\mathbf{X}\right)\right) = \left\langle F\left(\mathbf{x}, \mathbf{y}\right)\right\rangle_{Q(\mathbf{X})} - H\left(Q\left(\mathbf{X}\right)\right). \tag{15}$$

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

$$\mathrm{KL}\left(Q\left(\mathbf{X}\right) \parallel P\left(\mathbf{X} \mid \mathbf{y}\right)\right) = \mathcal{L}\left(Q\left(\mathbf{X}\right)\right) + \log Z\left(\mathbf{y}\right). \tag{16}$$

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

$$\mathcal{L}\left(Q\left(\mathbf{X}\right)\right) \geq -\log Z\left(\mathbf{y}\right), \tag{17}$$

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

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

$$\begin{aligned}
\mathcal{L}\left(Q\left(\mathbf{X}\right)\right) &= \sum_{\mathbf{x}} \prod_i Q\left(x_i\right) F\left(\mathbf{x}, \mathbf{y}\right) \\
&\quad + \sum_i \sum_{x_i} Q\left(x_i\right) \log Q\left(x_i\right) \\
&= \sum_{\mathbf{x}} Q(x_j) \left\langle F\left(\mathbf{x}, \mathbf{y}\right)\right\rangle_{\prod_{i:i\neq j} Q(X_i)} \\
&\quad - H(Q(X_j)) - \sum_{i:i\neq j} H\left(Q\left(X_i\right)\right),
\end{aligned} \tag{18}$$

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:

$$Q^*(X_j = x_j) = \frac{1}{Z_j} \exp(-\langle F(\mathbf{x}, \mathbf{y})\rangle_{\prod_{i:i\neq j} Q(X_i)}), \tag{19}$$

where Z_j is a normalization constant computed for each update so that $Q^*(x_j)$ sums to one. See Weinman et al. (2007) for the complete derivation of (19). Iteratively updating $Q(X_j)$ in this manner for each variable X_j will monotonically decrease the free energy $\mathcal{L}(Q(\mathbf{X}))$, thus minimizing the KL divergence.

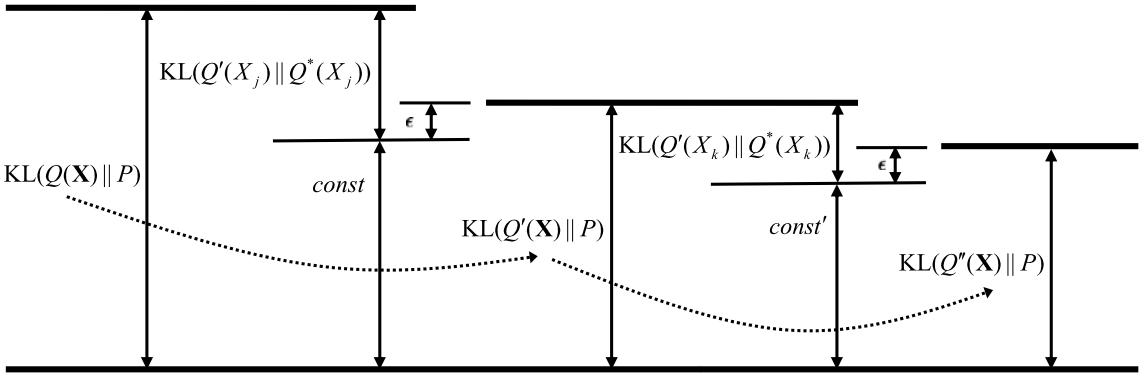

Fig. 2 Minimizing the global KL divergence via two different sparse local updates. The *global* divergence $\mathrm{KL}(Q(\mathbf{X}) \parallel P)$ can be decomposed into a *local* update plus a constant: $\mathrm{KL}(Q'(X_j) \parallel Q^*(X_j)) + const$. Consequently, at each step of sparse variational message passing we may minimize different local divergences to within some ϵ and when updating different local Qs, we minimize the global KL divergence

4.2 Sparse Variational Message Passing

Variational marginals can be more valuable than graph-cut-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 (19) can be computationally burdensome. Here we show how to dramatically reduce the computational load of calculating updates when many states have a very low probability under the variational distribution. The sparse methods presented here represent 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 $Q(x_j)$ 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 $Q(X_j)$.

Given (16), (18), and (19) $\mathrm{KL}(Q(\mathbf{X}) \parallel P(\mathbf{X} \mid \mathbf{y}))$ can be expressed as a function of a *sparse* update $Q'(X_j)$, the original mean field update $Q^*(X_j)$ and the other $Q(X_i)$'s, where $i \neq j$:

$$\begin{aligned}&\mathrm{KL}\left(Q\left(\mathbf{X}\right) \parallel P\left(\mathbf{X} \mid \mathbf{y}\right)\right)\\ &\quad = \mathrm{KL}(Q'(X_j) \parallel Q^*(X_j))\\ &\qquad + \log Z_j + \log Z\left(\mathbf{y}\right) - \sum_{i:i\neq j} H\left(Q\left(X_i\right)\right). \end{aligned} \tag{20}$$

Since the last three terms of (20) are constant with respect to sparse our update $Q'(X_j)$, $\mathrm{KL}(Q(\mathbf{X}) \parallel P(\mathbf{X} \mid \mathbf{y}))$ is minimized when $Q'(X_j) = Q^*(X_j)$. At each step of *sparse* variational message passing, we will minimize $\mathrm{KL}(Q'(\mathbf{X}_j) \parallel Q^*(\mathbf{X}_j))$ to within some small ϵ. As a result, each update to a different $Q(\mathbf{X}_j)$ yields further reduction of the *global* KL divergence. These relationships are illustrated in Fig. 2.

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 $Q'(X_j)$ in terms of the original update $Q^*(X_j)$ and the characteristic / indicator function $\mathbf{1}_{\mathcal{X}_j}(x_j)$ for the restricted range:

$$Q'(x_j) = \frac{\mathbf{1}_{\mathcal{X}_j}(x_j)}{Z'_j} Q^*(x_j), \tag{21}$$

where the new normalization constant is

$$Z'_j = \sum_{x_j} Q'(x_j) = \sum_{x_j \in \mathcal{X}_j} Q^*(x_j). \tag{22}$$

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

$$\begin{aligned}&\mathrm{KL}(Q'(X_j) \parallel Q^*(X_j))\\ &\quad = \sum_x \frac{\mathbf{1}_{\mathcal{X}_j}(x_j)}{Z'_j} Q^*(x_j)\\ &\qquad \times \log\left(\left(\frac{\mathbf{1}_{\mathcal{X}_j}(x_j)}{Z'_j} Q^*(x_j)\right) \Big/ Q^*(x_j)\right) \end{aligned} \tag{23}$$

$$\begin{aligned}&\quad = -\log Z'_j \frac{1}{Z'_j} \sum_{x \in \mathcal{X}_j} Q^*(x_j)\\ &\quad = -\log Z'_j. \end{aligned} \tag{24}$$

As a consequence, it is straightforward and efficient to compute a maximally sparse $Q'(X_j)$ such that

$$\mathrm{KL}(Q'(X_j) \parallel Q^*(X_j)) \leq \varepsilon \tag{25}$$

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 $\varepsilon = -\log 0.99 \approx .01$. We have thus created a global KL divergence minimization scheme using the local divergence analysis given for belief propagation in Pal et al. (2006) (23)–(24).

Figure 2 illustrates the way in which sparse VMP iteratively minimizes the $\mathrm{KL}(Q(\mathbf{X}) \parallel P(\mathbf{X} \mid \mathbf{y}))$ after each iteration of message passing. In short, sparse mean field updates result in a slightly lesser reduction of the KL divergence at each iteration. Thus it is possible that sparse updates may require more iterations to achieve the same approximation. However, the dramatic speedup easily recoups the suboptimal step size by allowing multiple iterations to be completed very quickly. In Sect. 5 we show how using sparse messages can yield a dramatic increase in inference speed.

Concretely speaking, for a model with N label states, when messages are passed between variables and their neighbor through a pairwise interaction potential function, sparsification reduces the computation from $O(N \times N)$ to $O(K \times N)$ for $K \ll N$. Importantly, this speedup is gained for each iteration of variational message passing and one typically needs to perform many iterations, re-visiting each variable multiple times. Indeed, to propagate information across the image one needs to have as many message passing iterations for each variable as the longest path from one variable to another in the lattice. Additionally, after the sparse message passing phase of the algorithm we compute parameter updates. For this step we have single node and pairwise variable expectation which can be performed using $O(K_i)$ vs. $O(N)$ operations and $O(K_i \times K_j)$ vs. $O(N \times N)$ operations for pixels i and pairs of pixels i and j in the image. However this savings is small compared to savings during inference so we use a full distribution for the final expectation and approximate marginal.

5 Experiments

In this section we present the results of a number of experiments. The first batch of experiments examine learning and generalization using a simple model. In Sect. 5.1 we first examine the convergence when learning simple models with only gradient modulation terms. We train models having different numbers of discretized bins with graph cuts for approximate marginals and all six data sets as our training set. Then in Sect. 5.2, we use a leave-one-out approach to evaluate the performance of the learned parameters on a new data set. In Sect. 5.3 we then examine how the learned parameters generalize to other data sets.

Our second batch of experiments in Sect. 5.4 examines the impact of extending our simple model in a variety of ways. These experiments explore extensions of the canonical model of Sect. 3.1 with the disparity difference dependent modulation terms of Sect. 3.2, the patch matching strategy of Sect. 3.3, and the occlusion models developed in Sect. 3.4.

Our third batch of experiments compare inference and learning using different approximate inference techniques for marginals. The first experiment of this batch in Sect. 5.5 compares sparse and traditional mean field methods for approximate inference, showing how sparse message passing can greatly accelerate free energy minimization. The second experiment in Sect. 5.6 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.

For all our experiments we use a straightforward gradient-based optimization procedure: we start with a small learning rate (10^{-4}) and increase it by a small factor unless the norm of the gradient increases dramatically, in which case we backtrack and decrease the learning rate.

As training and test data we use 6 stereo pair images with ground-truth disparities from the 2005 scenes of the Middlebury stereo database. These images are roughly 460×370 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.

5.1 Convergence

In these experiments we focus on learning the Θ_v parameters of the pairwise V potentials.

It is important to account for the fact that we do not model occlusions in this simple CRF. It is well-known that spurious minimal-cost matches in occluded areas can cause artifacts in the inferred disparity maps. We therefore use our ground-truth data to mask out the contributions of variables in occluded regions to our gradient computation during training. There are a number of more principled ways to address this issue. For example, in the model we developed in Sect. 3.4 we take a more principled approach by creating an additional occlusion state in our model. Another strategy might be to treat the occluded pixel as a hidden variable, then use an expected gradient or expectation maximization approach for learning with these pixels. Techniques for learning CRFs with hidden variables are discussed in more detail in Sutton and McCallum (2006). Indeed, an even better approach might be to use both these strategies, introducing a separate

binary indicator variable for hidden vs. not hidden pixels as well as a hidden value for such pixels.

We experiment with learning models using different numbers of parameters Θ_v, from $K = 1$ (i.e., a single global smoothness weight) to $K = 6$ (i.e., a parameter for each of 6 gradient bins). We first demonstrate the effectiveness of the learning by training on all six datasets. It is useful to visualize the disparities predicted by the model over each iteration of learning. Figures 3 and 4 show how the disparity maps change during training. For clarity we have masked the occluded regions in black in these plots, since our model will assign arbitrary disparities in these areas. Table 1 shows the discretization strategy we use for image gradients as well as the final values of the learned parameters.

Figure 5 (top) shows how the gradient during learning illustrating that our optimization procedure terminates with a near zero gradient. This indicates that the expectation of features under the (approximate) conditional distribution of the model is able to match the empirical expectation of the features. If we did not have an approximate distribution and expectation in our gradient this would indicate a global maximum due to the convexity of the CRF objective.

Note that convergence is faster for fewer parameters. Figure 5 (bottom) shows the disparity errors during learning. Again, models with fewer parameters converge more quickly, thus yielding lower errors faster. However, the models with more parameters eventually outperform the simpler models. In Fig. 5 (top) we observe that there appears to be an initial phase (e.g., during the first 25 iterations) where the norm of the approximate gradient monotonically decreases during the optimization. After this point, models with larger numbers of parameters appear to have less stability. This effect may be as a result of noisy gradient approximations due to our use of graph-cut-derived MPEs for the model expectation term of our gradient.

5.2 Performance of learned parameters

We now use 5 of the 6 datasets for training, and evaluate the disparity error of the remaining dataset using the parameters obtained during training. Figure 6 shows the results for the Moebius dataset. The top plot shows the errors during leave-one-out training. One can observe a sim-

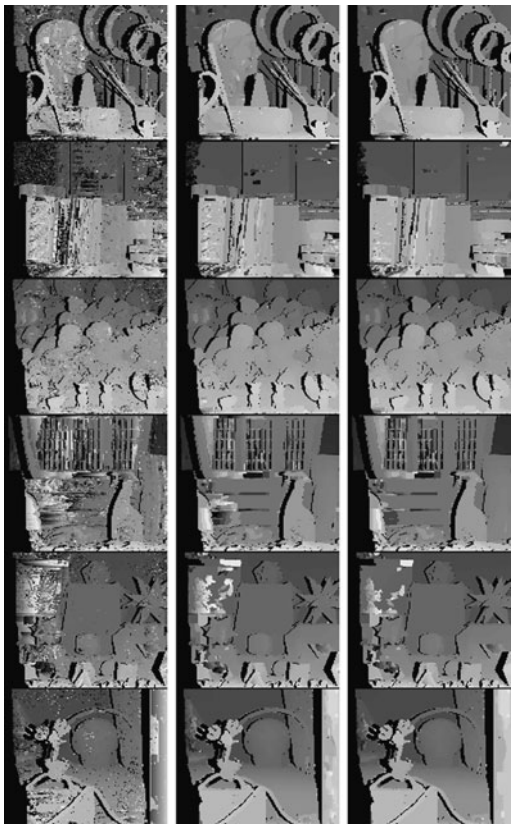

Fig. 3 Disparity maps of the entire training set for $K = 3$ parameters after 0, 10, and 20 iterations. Occluded areas are masked (©2007 IEEE)

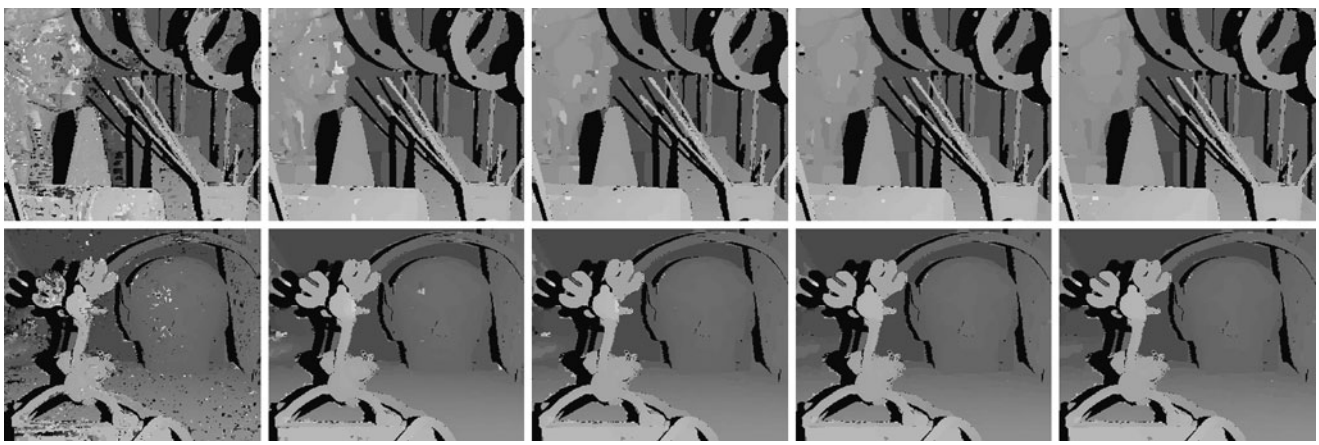

Fig. 4 Two zoomed views of the disparity maps for $K = 3$ parameters and learning on all six data sets after 0, 5, 10, 15, and 20 iterations. Occluded areas are masked (©2007 IEEE)

Table 1 The gradient bins for $K = 1, \ldots, 6$ parameters and the parameter values θ_k learned over all six datasets

Intervals	0-2	2-4	4-8	8-12	12-16	16-∞
$\{\theta_k\}, K=1$	9.8					
$\{\theta_k\}, K=2$	15.3			3.7		
$\{\theta_k\}, K=3$	45.1		0.3	8.7		
$\{\theta_k\}, K=4$	42.2		0.5	5.6		10.4
$\{\theta_k\}, K=5$	42.0		1.6	3.1	5.9	11.3
$\{\theta_k\}, K=6$	104	3.9	11.2	3.8	3.0	13.7

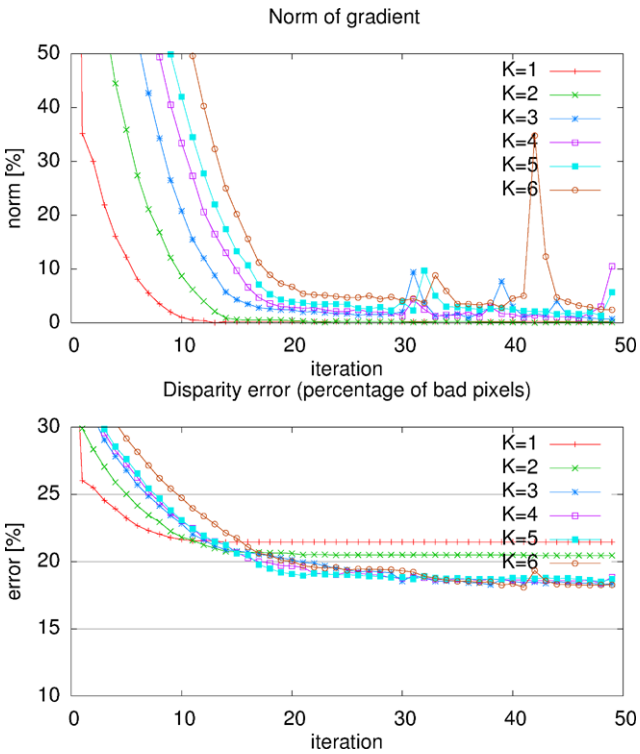

Fig. 5 Gradient norm (*top*) and disparity errors (*bottom*) during learning on all 6 datasets (©2007 IEEE)

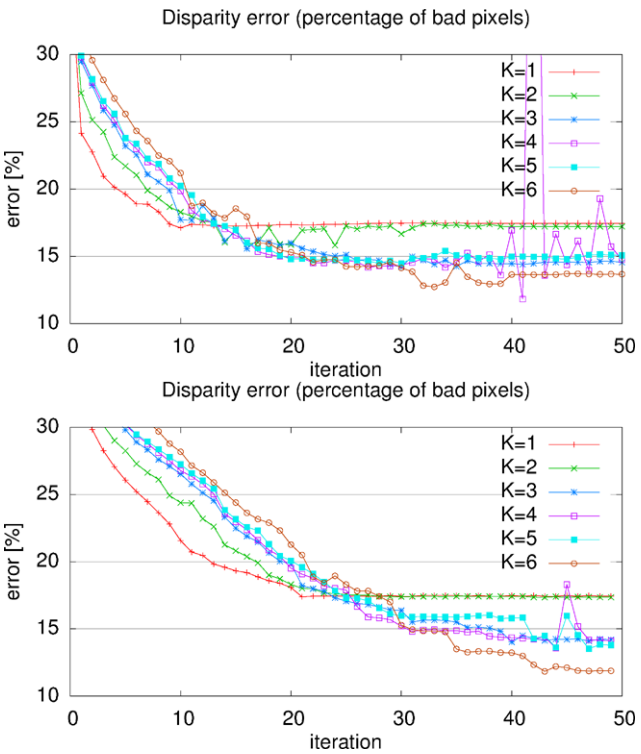

Fig. 6 Results of leave-one-out learning on the Moebius dataset. *Top*: Moebius disparity errors using the parameters obtained during learning from the other 5 datasets. *Bottom*: Moebius disparity errors using the parameters learned from the dataset itself (©2007 IEEE).

ilar trend as in Fig. 5 (bottom), namely that the errors decrease during learning, and that the more complex models eventually outperform the simpler models. For comparison, the bottom plot in Fig. 6 shows the errors when using the Moebius dataset itself for training. In this case finding a low-gradient solution means that we have effectively matched the distribution of disparity changes and associated intensity gradients of the ground-truth image. Not surprisingly, this results in lower errors, but not significantly lower than in the top plot—which indicates that the parameters learned from the other 5 images generalize reasonably well.

Figure 7 shows the equivalent plots for a different dataset, Reindeer. Again we show the errors during leave-one-out training at the top and those during training on the dataset itself on the bottom. Here we get slightly different results. First, the leave-one-out results no longer indicate that performance increases with the number of parameters. In fact the model with $K = 2$ does best in the end. But the results in the bottom plot (where we train the parameters on the test data itself) show that this is not necessarily a problem of insufficient generalization, but rather that learning the best parameters (which amounts to matching the smoothness properties of the ground truth) might not always yield to lower matching errors. On the other hand, this could also be due to noisy gradient approximations as mentioned earlier.

5.3 Performance on standard benchmarks

Next, we examine how well the parameters learned from our six datasets generalize to other stereo images. Table 2 shows the disparity errors on the Middlebury benchmark consisting of the Tsukuba, Venus, Teddy, and Cones images. We compare these errors with those of the graph cuts (GC) method in Scharstein and Szeliski (2002), which uses a hand-tuned MRF model with two gradient bins. Our average results for $K = 1$ and $K = 2$ are slightly better than those of GC, and would result in a similar ranking as the GC method in the Middlebury evaluation. We provide this to illustrate that we are able to match and in fact slightly exceed the performance of a canonical model. The fact that the errors for the more

complex models are higher may indicate that the learned parameters of those models are tuned more finely to the characteristics of the training data and generalize less well to datasets that are quite different. We also give a result on the benchmark after learning a more complex disparity difference dependent modulation model as outlined in Sect. 3.2 and further explored in Table 3. Here again we use only the new data for learning and test on the benchmark. In this case we see a more dramatic gain over the canonical model.

5.4 Extending a Canonical CRF for Stereo

In Table 3 we compare the performance of different models and feature types described in Sects. 3.1 to 3.4 when learning across all six of the new stereo image pairs we have created. In all cases here we use graph cuts for inference during learning. Importantly, we observed that our optimization for the experiments with both pixel disparity difference discretization and gradient disparity difference discretization defined in Sect. 3.2 terminates as a result of the energy function violating the constraints imposed by graph cuts. These model configurations perform well despite running into the limitation of graph-cut-based inference. As such, this case serves as an illustrative example motivating our next set of experiments where we use message passing methods for marginal inference during learning.

In Table 4 we show the learned parameters for a model with interaction terms dependent on both the disparity difference between pixels and the magnitude of the gradient between pixels. This table gives us an idea of the shape of the learned interaction potential function. We have used a discretization strategy for our experiments here for two main reasons: First, we have a lot of data in each image, so learning a discretized function is reasonable even with many bins. Second, this strategy allows us to easily visualize the corresponding parameters and gauge their impact. Another reasonable strategy might be to use a functional basis such as one based on polynomials, which may be an interesting avenue for future exploration.

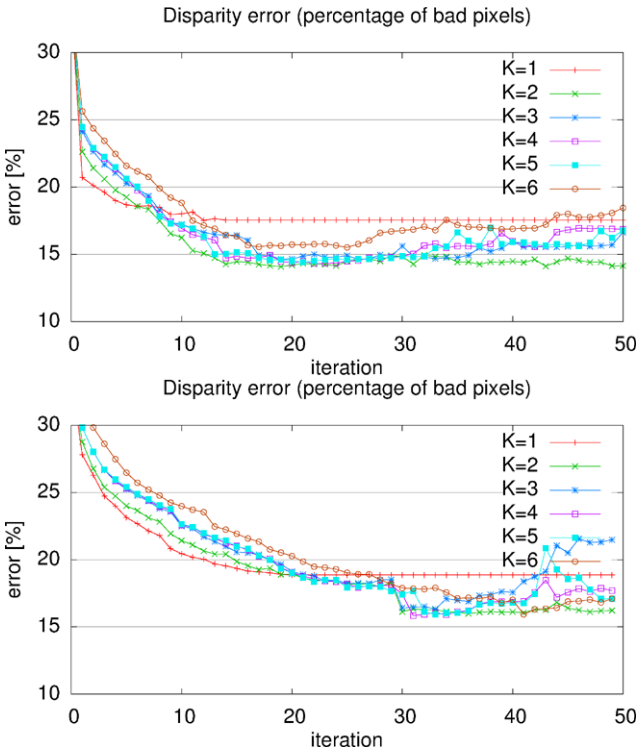

Fig. 7 Results of leave-one-out learning on the Reindeer dataset. *Top*: Disparity errors using the parameters obtained during learning from the other 5 datasets. *Bottom*: Disparity errors using the parameters learned from the dataset itself (©2007 IEEE)

Table 2 A comparison of models with different numbers of parameters K trained on our ground-truth data but evaluated on the Middlebury data set. The last two rows are the performance of the graph cut implementation of Scharstein and Szeliski (2002) and the disparity difference modulation approach of Sect. 3.2

	Tsukuba	Venus	Teddy	Cones	Average
$K=1$	3.0	1.3	11.1	10.8	6.6
$K=2$	2.2	1.6	11.3	10.7	6.5
$K=3$	3.1	2.6	16.4	19.6	10.4
$K=4$	3.0	2.5	17.3	21.5	11.1
$K=5$	2.8	2.1	16.4	21.2	10.6
$K=6$	3.1	2.7	14.5	16.8	9.3
GC	1.9	1.8	16.5	7.7	7.0
Disp. dif.	1.9	1.2	11.0	7.0	5.3

Table 3 Comparison of the training set disparity error (percentage of incorrectly predicted pixels) given: (**a**) for the canonical stereo model with three gradient bins, (**b**) with five, (**c**) modulation terms based on both pixel disparity difference and image gradient information, (**d**) a gradient-modulated occlusion model

Method	Art	Laundry	Books	Dolls	Moebius	Reindeer	Average
Section 3.1: Gradient bins [2, 4]	22.66	30.88	26.17	12.19	18.41	17.92	21.37
Section 3.1: Gradient bins [1, 2, 3, 4]	20.53	22.66	19.03	12.09	13.01	16.06	17.23
Section 3.3: 3×3 patches and gradient bins [1, 2, 3, 4]	14.95	24.29	19.58	10.95	12.39	15.36	16.25
Section 3.2: Disp. dif. bins and grad. bins: both [1, 2, 3, 4]	17.39	19.43	16.89	10.89	12.83	14.10	15.26
Section 3.4: Occlusion model with gradient bins [2, 4]	15.39	21.16	16.53	9.43	12.04	14.27	14.80

5.5 Approximate Inference: Speed, Energy Minimization, and Marginals

The variational distribution $Q(\mathbf{X})$ 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.

Our first set of experiments uses the simpler canonical stereo model having the smoothness term V of (8). Figure 8 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[3] or 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 to (24) and the criterion (25). Figure 9 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.

Table 4 The parameters of the disparity difference and gradient modulated CRF of Sect. 3.2

Disp. difference interval	Grad. difference interval				
	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, inf)
[0, 1)	2.4	1.4	5.4	7.5	0.0
[1, 2)	19.0	20.0	17.7	16.2	21.7
[2, 3)	21.2	21.3	20.9	20.5	22.3
[3, 4)	25.2	25.1	25.0	25.0	25.2
[4, inf)	32.3	32.1	31.1	30.8	32.9

[3]When entropy is computed using the natural logarithm (as opposed to the base 2 logarithm for bits) the *nat* is the implicit and natural unit for information entropy.

5.6 Learning with Message Passing vs. Graph Cuts

Maximizing the log likelihood (13) for learning requires marginals on the lattice. When the model is initialized, these marginals have higher entropy (Fig. 8), 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-cut 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 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 the same 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

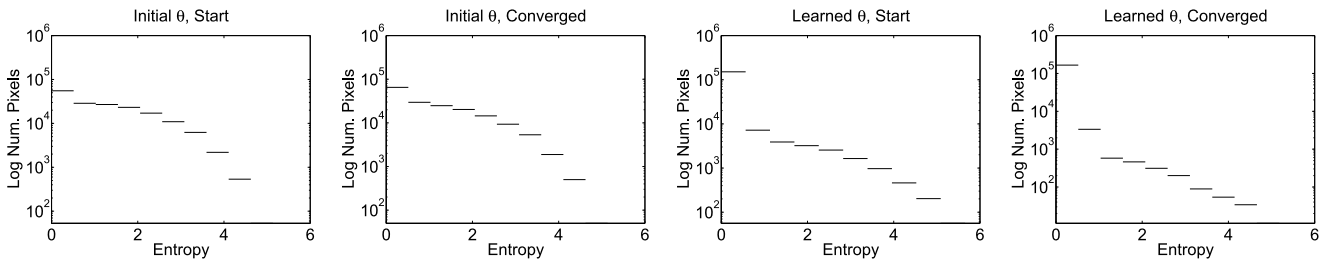

Fig. 8 Histograms of approximate marginal entropies $H(Q(X_i))$ from the variational distributions for each pixel at the start (after the first round) of mean field updates and at their convergence; values using the initial and learned parameters Θ_v of the canonical model are shown

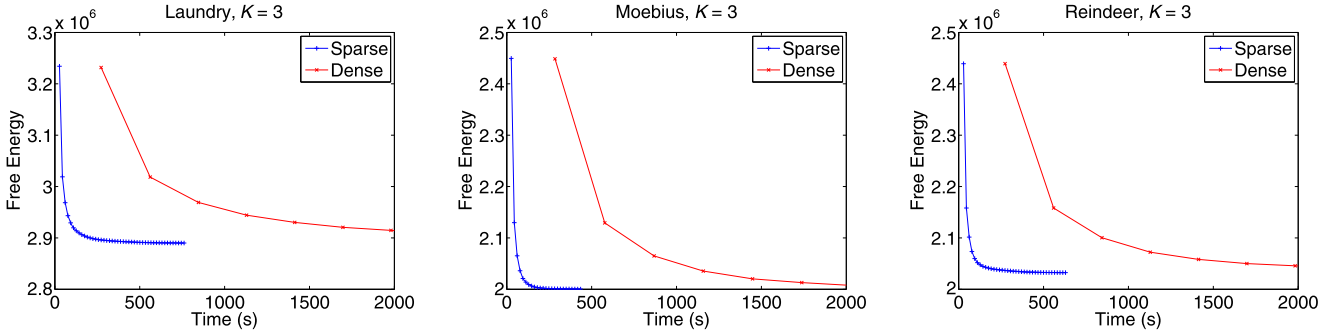

Fig. 9 Comparison of CPU time for free energy minimization with sparse and dense mean field updates using parameters Θ_v learned in the canonical model with three images (Art, Books, Dolls)

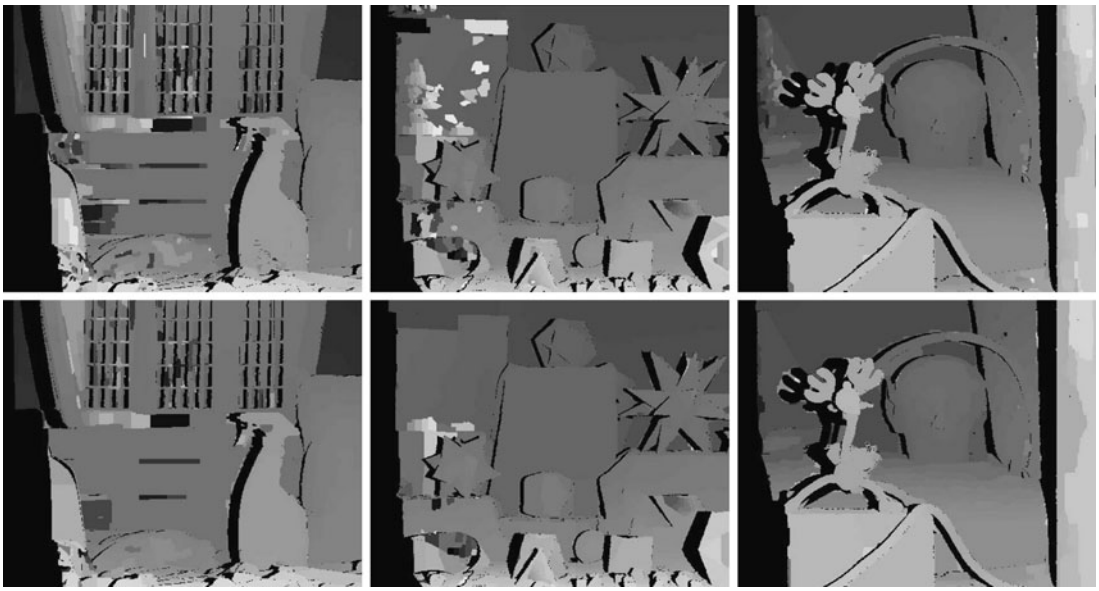

Fig. 10 Test images comparing prediction (using graph cuts) after one round of learning the canonical model with graph cuts (*top*) or sparse mean field (*bottom*). Occluded areas are *black*. Images (l–r): Laundry, Moebius, Reindeer

trained with graph cuts for inference. Figure 10 shows the corresponding depth images after one iteration.

In Table 5, we compare the results of training using pseudolikelihood, sparse mean field, and point estimates provided by graph cuts. 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. Comparing learning with graph cuts and learning with sparse mean field, the disparity error in 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 (11) to handle the interactions between occluded states and the local terms of (10). We show both leave-one-out training and test results as well as the result of training on all the data to serve as a reference point for the capacity of the model. For this last set of experiments we show root mean square (RMS) errors for disparity predictions. 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 ($p < 0.05$).

Figure 11 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 (Fig. 12). We also provide the learning curves for pseudolikelihood for comparison. We see that pseudolikelihood generally has poorer performance, both initially and after many iterations using a similar learning strategy compared to graph cuts and sparse mean field.

Table 5 Comparison of learning with pseudolikelihood, 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 (10) and (11)). For the gradient-modulated occlusion model we show the occlusion prediction error (percentage). In the last block of experiments we show RMS error

Metric	Method	Art	Books	Dolls	Laundry	Moebius	Reindeer	Average
Canonical Model—leave-one-out training & testing								
Disparity	Pseudo Likelihood	22.03	27.32	11.85	29.50	15.93	15.88	20.42
% 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	Graph Cuts	21.82	24.10	11.94	27.54	11.08	16.74	19.30
% error	Sparse Mean Field	21.05	23.14	11.62	27.37	11.45	16.44	18.93
Occlusion	Graph Cuts	34.50	28.27	32.99	36.89	40.65	50.83	37.36
% error	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	Graph Cuts	10.61	19.2	5.98	20.95	7.15	5.53	12.78
RMS error	Sparse Mean Field	8.29	13.41	4.72	19.22	5.11	4.76	10.15
Occlusion	Graph Cuts	16.20	10.40	24.88	29.77	27.88	32.97	21.83
% error	Sparse Mean Field	10.47	8.10	19.43	23.04	21.10	27.31	16.43

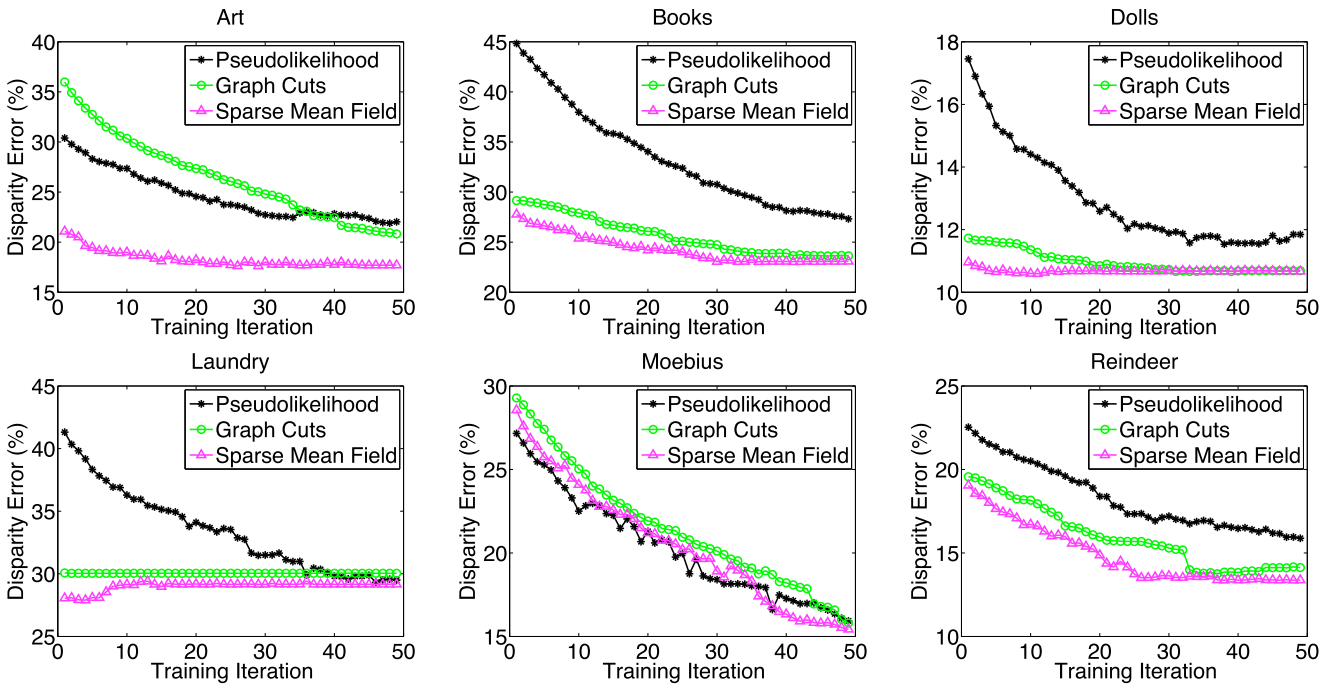

Fig. 11 Disparity error (each image held out in turn) using pseudolikelihood, graph cuts, and mean field for learning the canonical CRF stereo model. The error before learning is omitted from the plots to better highlight performance differences

6 Conclusions

As more evaluation and training data becomes available for stereo it is natural for researchers to desire to create more sophisticated models for stereo. While hand specified stereo models were once widely used, with the increasing availability of ground truth data it is likely that interest will increase in using learning to improve more complex models. By creating ground truth data with structured light we have been able to explore and evaluate a variety of different CRF models and learning techniques for stereo. We have shown advantages to using image patches computed on higher resolution imagery, advantages to interactions potentials that are dependant on disparity differences as well as advantages

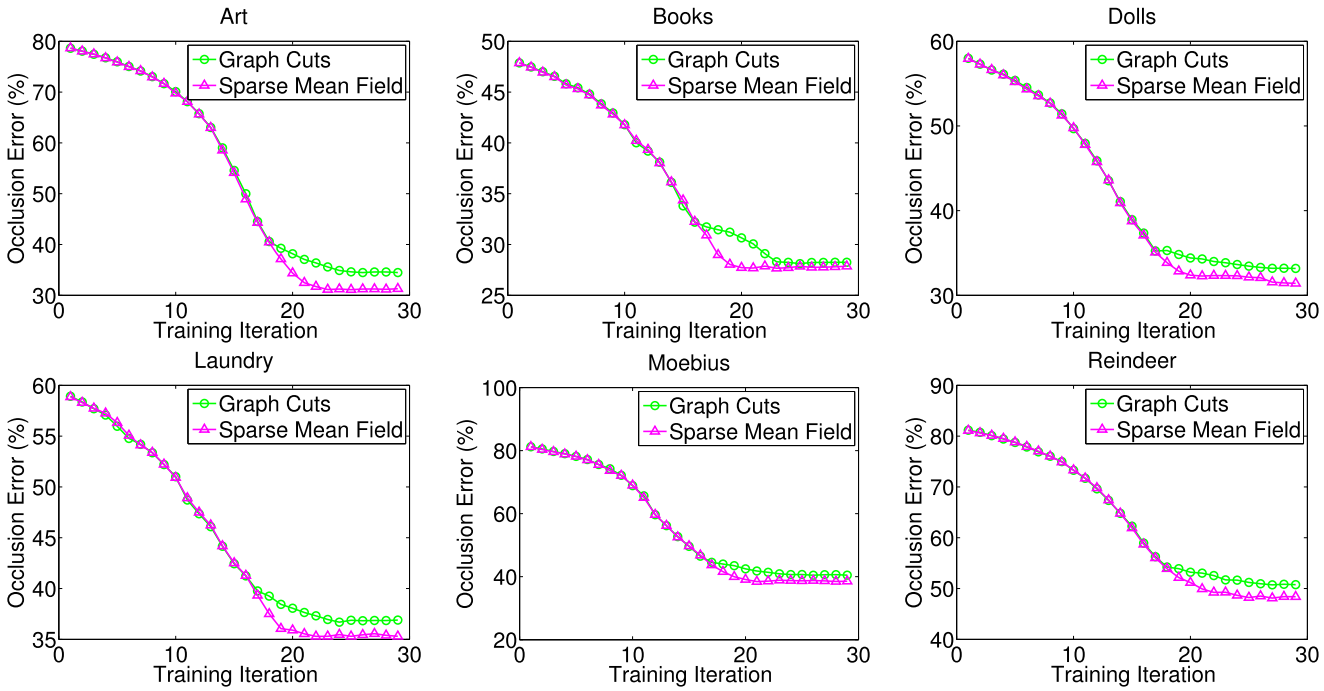

Fig. 12 Comparison of error predicting occluded pixel using graph cuts and sparse mean field for learning in the gradient-modulated occlusion model (11)

to formally modelling occlusions as random variables in a CRF. A natural direction for future work would be to combine these different elements into a more sophisticated CRF. Indeed, one of the top performing methods on the Middlebury benchmark (Yang et al. 2006) is a technique that incorporates many similar elements. More specifically, this work by Yang et al. (2006) is based on an MRF and uses: hierarchical belief propagation, color-weighted correlation and occlusion handling.

As state of the art models become more complex, both a principled underlying modelling formalism as well as principled, efficient, stable and robust learning techniques become important. We hope that the CRF formulation we have provided in this paper can serve as a good starting point for more sophisticated discriminative random field models for stereo. In practical terms, graph cuts was the fastest algorithm we explored. Pseudolikelihood (PL) based learning can also be fast, especially if one exploits the fact that gradients can be computed exactly and fast second order optimization could thus be used. However, our experiments indicate that the quality of models learned via PL is worse than those learned using GC and that sparse variational message passing (sparse VMP) can produce the highest quality learned models among these three alternatives.

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 models where there is more uncertainty (as in the early stages of learning), we find 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 VMP could be seen as a method occupying a middle ground between producing point estimates and creating fuller approximate distributions. Interestingly, sparse message passing could also be used to speed up state of the art TRW and TRW-S inference techniques. 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.

Finally, with the insights provided in this study, we hope to open up a number of avenues of exploration for learning in generally richer models and learning models suitable for processing more views using the additional data sets we have created.

Acknowledgements We would like to thank Anna Blasiak and Jeff Wehrwein for their help in creating the data sets used in this paper. Figures (©2007 IEEE) also appear in Scharstein and Pal (2007). Support for this work was provided in part by NSF grant 0413169 to D.S.

References

Alvarez, L., Deriche, R., Snchez, J., & Weickert, J. (2002). Dense disparity map estimation respecting image discontinuities: a PDE and scale-space based approach. *Journal of Visual Communication and Image Representation*, *13*(1–2), 3–21.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Barnard, S. (1989). Stochastic stereo matching over scale. *International Journal of Computer Vision*, *3*(1), 17–32.

Birchfield, S., & Tomasi, C. (1998). A pixel dissimilarity measure that is insensitive to image sampling. *IEEE TPAMI*, *20*(4), 401–406.

Blake, A., Rother, C., Brown, M., Perez, P., & Torr, P. (2004). Interactive image segmentation using an adaptive GMMRF model. In *Proc. ECCV* (pp. 428–441)

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bleyer, M., & Gelautz, M. (2004). A layered stereo algorithm using image segmentation and global visibility constraints. In *Proc. ICIP.*

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE TPAMI*, *23*(11), 1222–1239.

Cowell, R., Dawid, A., Lauritzen, S., & Spiegelhalter, D. (2003). *Probabilistic Networks and Expert Systems*. Berlin: Springer.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE TPAMI*, *19*, 380–393.

Felzenszwalb, P., & Huttenlocher, D. (2006). Efficient belief propagation for early vision. *International Journal of Computer Vision*, *70*(1), 41–54.

Frey, B., & Jojic, N. (2005). A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE TPAMI*, *27*(9), 1392–1416.

Frey, B., & MacKay, D. (1997). A revolution: Belief propagation in graphs with cycles. In *Proc NIPS*.

He, Z., Zemel, R., & Carreira-Perpinan, M. (2004). Multiscale conditional random fields for image labeling. In *Proc. CVPR* (pp. 695–702).

Hong, L., & Chen, G. (2004). Segment-based stereo matching using graph cuts. In *Proc. CVPR* (Vol. I, pp. 74–81).

Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, *37*, 183–233.

Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *IEEE TPAMI*, *28*, 1568–1583.

Kolmogorov, V., & Zabih, R. (2001). Computing visual correspondence with occlusions using graph cuts. In *Proc. ICCV* (pp. 508–515).

Kolmogorov, V., & Zabih, R. (2002a). Multi-camera scene reconstruction via graph cuts. In *Proc. ECCV* (Vol. III, pp. 82–96).

Kolmogorov, V., & Zabih, R. (2002b). What energy functions can be minimized via graph cuts? In *Proc. ECCV* (Vol. III, pp. 65–81).

Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., & Rother, C. (2006). Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *IEEE TPAMI*, *28*(9), 1480–1492.

Kong, D., & Tao, H. (2004). A method for learning matching errors in stereo computation. In *Proc. BMVC*.

Kschischang, F., Frey, B., & Loeliger, H.A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions Info Theory*, *47*(2), 498–519.

Kumar, S., & Hebert, M. (2006). Discriminative random fields. *International Journal of Computer Vision*, *68*(2), 179–201.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML* (pp. 282–289).

Liang, P., & Jordan, M. (2008). An asymptotic analysis of generative, discriminative and pseudolikelihood estimators. In *Proc ICML.*

Murphy, K., Weiss, Y., & Jordan, M. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Proc. UAI* (pp. 467–475).

Ng, A. Y., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Proc. NIPS.*

Pal, C., Sutton, C., & McCallum, A. (2006). Sparse forward-backward using minimum divergence beams for fast training of conditional random fields. In *Proc. ICASSP* (pp. 581–584).

Scharstein, D., & Pal, C. (2007). Learning conditional random fields for stereo. In *Proc. CVPR.*

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*(1), 7–42.

Scharstein, D., & Szeliski, R. (2003). High-accuracy stereo depth maps using structured light. In *Proc. CVPR* (Vol. I, pp. 195–202).

Strecha, C., Tuytelaars, T., & Van Gool, L. (2003). Dense matching of multiple wide-baseline views. In *Proc. CVPR* (Vol. 2, p. 1194).

Strecha, C., Fransens, R., & Van Gool, L. (2004). Wide-baseline stereo from multiple views: A probabilistic account. In *Proc. CVPR* (Vol. 1, pp. 552–559).

Sun, J., Zheng, N., & Shum, H. (2003). Stereo matching using belief propagation. *IEEE TPAMI*, *25*(7), 787–800.

Sun, J., Li, Y., Kang, S. B., & Shum, H. Y. (2005). Symmetric stereo matching for occlusion handling. In *Proc. CVPR* (pp. 399–406).

Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning. In L. Getoor & B. Taskar (Eds.), *Introduction to Statistical Relational Learning*. Cambridge: MIT Press.

Sutton, C., Rohanimanesh, K., & McCallum, A. (2004). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *Proc. ICML.*

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., & Rother, C. (2008). A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE TPAMI*, *30*, 1068–1080.

Tao, H., Sawhney, H., & Kumar, R. (2001). A global matching framework for stereo computation. In *Proc. ICCV* (Vol. I, pp. 532–539).

Tappen, M., & Freeman, W. (2003). Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In *Proc. ICCV* (pp. 900–907).

Trinh, H., & McAllester, D. (2009). Unsupervised learning of stereo vision with monocular cues. In *Proc. BMVC.*

Vishwanathan, S., Schraudolph, N., Schmidt, M., & Murphy, K. (2006). Accelerated training of conditional random fields with stochastic gradient methods. In *Proc. ICML* (pp. 969–976). New York: ACM.

Wainwright, M., Jaakkola, T., & Willsky, A. (2002). Tree-based reparameterization for approximate estimation on graphs with cycles. In *Proc. NIPS.*

Wainwright, M., Jaakkola, T., & Willsky, A. (2003). Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transaction on Information Theory*, *45*(9), 1120–1146.

Wainwright, M., Jaakkola, T., & Willsky, A. (2005). Map estimation via agreement on trees: Message-passing and linear programming. *IEEE Transaction on Information Theory*, *51*(11), 3697–3717.

Wei, Y., & Quan, L. (2004). Region-based progressive stereo matching. In *Proc. CVPR* (vol. I, pp. 106–113).

Weinman, J. J., Hanson, A., & McCallum, A. (2004). Sign detection in natural images with conditional random fields. In *IEEE Int. Workshop on Machine Learning for Signal Processing* (pp. 549–558).

Weinman, J. J., Pal, C., & Scharstein, D. (2007). Sparse message passing and efficiently learning random fields for stereo vision. Tech. Rep. UM-CS-2007-054, Univ. of Massachusetts, Amherst.

Weinman, J. J., Tran, L., & Pal, C. (2008). Efficiently learning random fields for stereo vision with sparse message passing. In *Proc. ECCV*.

Weinman, J. J., Learned-Miller, E., & Hanson, A. (2009). Scene text recognition using similarity and a lexicon with sparse belief propagation. *IEEE TPAMI*, *31*(10), 1733–1746.

Winn, J., & Bishop, C. (2005). Variational message passing. *Journal of Machine Learning Research*, *6*, 661–694.

Yang, Q., Wang, L., Yang, R., Stewenius, H., & Nister, D. (2006). Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. In *Proc. CVPR*.

Yedidia, J., Freeman, W., & Weiss, Y. (2003). Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium* (pp. 239–236).

Zhang, L., & Seitz, S. (2005). Parameter estimation for MRF stereo. In *Proc. CVPR* (Vol. II, pp. 288–295).

Zhang, Y., & Kambhamettu, C. (2002). Stereo matching with segmentation-based cooperation. In *Proc. ECCV* (Vol. II, pp. 556–571).

Zitnick, L., Kang, S., Uyttendaele, M., Winder, S., & Szeliski, R. (2004). High-quality video view interpolation using a layered representation. *SIGGRAPH, ACM Transactions on Graphics*, *23*(3), 600–608.