Filter-based Mean-Field Inference for Random Fields with Higher-Order Terms and Product Label-Spaces

Vibhav Vineet*, Jonathan Warrell*, and Philip H.S. Torr

Oxford Brookes University
{vibhav.vineet-2010,jwarrell,philiptorr}@brookes.ac.uk
http://cms.brookes.ac.uk/research/visiongroup/

Abstract. Recently, a number of cross bilateral filtering methods have been proposed for solving multi-label problems in computer vision, such as stereo, optical flow and object class segmentation that show an order of magnitude improvement in speed over previous methods. These methods have achieved good results despite using models with only unary and/or pairwise terms. However, previous work has shown the value of using models with higher-order terms e.g. to represent label consistency over large regions, or global co-occurrence relations. We show how these higher-order terms can be formulated such that filter-based inference remains possible. We demonstrate our techniques on joint stereo and object labeling problems, as well as object class segmentation, showing in addition for joint object-stereo labeling how our method provides an efficient approach to inference in product label-spaces. We show that we are able to speed up inference in these models around 10-30 times with respect to competing graph-cut/move-making methods, as well as maintaining or improving accuracy in all cases. We show results on PascalVOC-10 for object class segmentation, and Leuven for joint object-stereo labeling.

1 Introduction

Many computer vision problems, such as object class segmentation, stereo and optical flow, can be formulated as multi-labeling problems, and expressed within a framework such as Markov Random Fields (MRFs), Conditional Random Fields (CRFs), or other structured models. Although exact inference in such models is in general intractable, much attention has been paid to developing fast approximation algorithms, including variants of belief propagation, dual decomposition methods, and move-making approaches [1–3]. Recently, a number of cross bilateral Gaussian filter-based methods have been proposed for problems such as object class segmentation [4], denoising [5], stereo and optical flow [6],

* The first two authors contributed to this work equally as joint first author. The work was supported by the EPSRC and the IST programme of the European Community, under the PASCAL2 Network of Excellence. Professor Philip H.S. Torr is in receipt of a Royal Society Wolfson Research Merit Award.

which permit substantially faster inference in these problems, as well as offering performance gains over competing methods. Our approach builds on such filter-based approaches and shows them to outperform or perform equally well to the previously dominant graph-cut/move-making approaches on all problems considered. This strongly suggests that mean-field message-passing enhanced with recent filtering techniques should be considered as a general state-of-the-art inference method for a large number of computer vision problems currently of interest.

A problem with filter-based methods as currently formulated is that they can only be applied to models with limited types of structure. In [6], dependencies between output labels are abandoned, and the filtering step is used to generate unary costs which are treated independently. In [4], filtering is used to perform inference in MRF models with dense pairwise dependencies taking the form of a weighted mixture of Gaussian kernels. Although allowing fully connected pairwise models increases expressivity over typical 4 or 8-connected MRF models, the inability to handle higher-order terms is a disadvantage.

The importance of higher-order information has been demonstrated in all of the labeling problems mentioned. For object class segmentation, the importance of enforcing label consistency over homogeneous regions has been demonstrated using P^n-Potts models [7], and co-occurrence relations between classes at the image level have also been shown to provide important priors for segmentation [8]. For stereo and optical flow, second-order priors have proved to be effective [9], as have higher-order image priors for denoising [10].

In this paper, we propose a number of methods by which higher-order information can be incorporated into MRF models for multi-label problems so that, under certain model assumptions, using efficient bilateral filter-based methods for inference remains possible. Specifically, we show how to encode (a) a broad class of local *pattern-based* potentials (as introduced in [11]), which include P^n-Potts models and second-order smoothness priors, and (b) global potentials representing co-occurrence relationships between labels as in [8, 12]. We assume a base-layer MRF with full connectivity and weighted Gaussian edge potentials as in [4]. Our approach allows us to apply bilateral filter-based inference to a wide range of models with complex higher-order structure. We demonstrate the approach on two such models, first a model for joint stereo and object class labeling as in [13], and second a model for object class segmentation with co-occurrence priors as in [8]. In the case of joint stereo and object labeling, in addition to demonstrating fast inference with higher-order terms, we show how cost-volume filtering can be applied in the product label-space to generate informative disparity potentials, and more generally how our method provides an efficient approach to inference in such product label-spaces. Further, we demonstrate the benefits for object-stereo labeling of applying recent *domain transform filtering* techniques [14] in our framework. In both joint stereo-object labeling and object class segmentation, we are able to achieve substantial speed-ups with respect to graph-cut based inference techniques and improvements in accuracy with respect to the baseline methods. In summary, our contributions are:

- A set of efficient techniques for including higher-order terms in random fields with dense connectivity, allowing for mean-field filter-based inference,
- An adaptation of our approach to product label-space models for joint object-stereo labeling, again permitting efficient inference,
- An investigation of the advantages/disadvantages of alternative filtering methods recently proposed [5, 14, 15] within our framework.

In Sec. 2 we review the method of [4]. Sec. 3 provides details on how we encode higher-order terms, Sec. 4 gives experimentation on joint stereo and object labeling and object class segmentation, and Sec. 5 concludes with a discussion.

2 Filter-based Inference in Dense Pairwise CRFs

We begin by reviewing the approach of [4], which provides a filter-based method for performing fast approximate maximum posterior marginal (MPM) inference[1] in multi-label CRF models with fully connected pairwise terms, where the pairwise terms have the form of a weighted mixture of Gaussian kernels. We define a random field over random variables $\mathcal{X} = \{X_1, ...X_N\}$ conditioned on an image $\mathbf{I}$. We assume there is a random variable associated with each pixel in the image $\mathcal{N} = \{1...N\}$, and the random variables take values from a label set $\mathcal{L} = \{l_1, ..., l_L\}$. We can then express the fully connected pairwise CRF as:

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp(-E(\mathbf{X}|\mathbf{I})) \tag{1}$$

$$E(\mathbf{X}|\mathbf{I}) = \sum_{i \in \mathcal{N}} \psi_u(x_i) + \sum_{i<j \in \mathcal{N}} \psi_p(x_i, x_j) \tag{2}$$

where $E(\mathbf{X}|\mathbf{I})$ is the energy associated with a configuration $\mathbf{X}$ conditioned on $\mathbf{I}$, $Z(\mathbf{I}) = \sum_{\mathbf{X}'} \exp(-E(\mathbf{X}'|\mathbf{I}))$ is the (image dependent) partition function, and $\psi_u(.)$ and $\psi_p(.,.)$ are unary and pairwise potential functions respectively, both implicitly conditioned on the image $\mathbf{I}$. The unary potentials can take arbitrary form, while [4] restrict the pairwise potentials to take the form of a weighted mixture of Gaussian kernels:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{M} w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) \tag{3}$$

where $\mu(.,.)$ is an arbitrary *label compatibility function*, while the functions $k^{(m)}(.,.)$, $m = 1...M$ are Gaussian kernels defined on feature vectors $\mathbf{f}_i$, $\mathbf{f}_j$ derived from the image data at locations i and j (where [4] form $\mathbf{f}_i$ by concatenating the intensity values at pixel i with the horizontal and vertical positions of pixel i in the image), and $w^{(m)}$, $m = 1...M$ are used to weight the kernels.

Given this form of CRF, [4] show how fast approximate MPM inference can be performed using cross bilateral filtering techniques within a mean-field approximation framework. The mean-field approximation introduces an alternative

[1] For exact MPM inference, the solution satisfies $x_i^{\text{MPM}} \in \text{argmax}_l \sum_{\{\mathbf{x}|x_i=l\}} P(\mathbf{x}|I)$.

distribution over the random variables of the CRF, $Q(\mathbf{X})$, where the marginals are forced to be independent, e.g. $Q(\mathbf{X}) = \prod_i Q_i(x_i)$. The mean-field approximation then attempts to minimize the KL-divergence $\mathbf{D}(Q||P)$ between Q and the true distribution P. By considering the fixed-point equations that must hold at the stationary points of $\mathbf{D}(Q||P)$, the following update may be derived for $Q_i(x_i = l)$ given the settings of $Q_j(x_j)$ for all $j \neq i$ (see [16] for a derivation):

$$Q_i(x_i = l) = \frac{1}{Z_i} \exp\{-\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j(x_j = l')\psi_p(x_i, x_j)\} \tag{4}$$

where $Z_i = \sum_{x_i = l \in \mathcal{L}} \exp\{-\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j(x_j = l')\psi_p(x_i, x_j)\}$ is a constant which normalizes the marginal at pixel i. If the updates in Eq. 4 are made in sequence across pixels $i = 1...N$ (updating and normalizing the L values $Q_i(x_i = l)$, $l = 1...L$ at each step), the KL-divergence is guaranteed to decrease [16]. In [4], it is shown that parallel updates for Eq. 4 can be evaluated by convolution with a high dimensional Gaussian kernel using any efficient bilateral filter, e.g. the permutohedral lattice method of [15] (which introduces a small approximation). This is achieved by the following transformation:

$$\tilde{Q}_i^{(m)}(l) = \sum_{j \neq i} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l) = [G_m \otimes Q(l)](\mathbf{f}_i) - Q_i(l) \tag{5}$$

where G_m is a Gaussian kernel corresponding to the m'th component of Eq. 3, and $\otimes$ is the convolution operator. Since $\sum_{j \neq i} Q_j(x_j = l')\psi_p(x_i, x_j)$ in Eq. 4 can be written as $\sum_m w^{(m)} \tilde{Q}_i^{(m)}(l')$, and approximate Gaussian convolution using [15] is $O(N)$, parallel[2] updates using Eq. 4 can be efficiently approximated in $O(MNL^2)$ time (or $O(MNL)$ time for the Potts model), thus avoiding the need for the $O(MN^2L^2)$ calculations which would be required to calculate these updates individually. Since the method requires the updates to be made in parallel rather than in sequence, the convergence guarantees associated with the sequential algorithm are lost [16]. However, [4] observe good convergence properties in practice. The algorithm is run for a fixed number of iterations, and the MPM solution extracted by choosing $x_i \in \text{argmax}_l\, Q_i(x_i = l)$ at the final iteration.

Although [4] use the permutohedral lattice [15] for their filter-based inference, we note that other filtering methods can also be used for the convolutions in Eq. 5. Particularly, the recently proposed *domain transform* filtering approach [14] has certain advantages over the permutohedral lattice. Domain transform filtering approximates high-dimensional filtering, such as 5-D bilateral filtering in 2-D spatial and 3-D RGB range space, by alternating horizontal and vertical 1-D filtering operations on transformed 1-D signals which are isometric to slices of the original signal. Since it does not sub-sample the original signal, its complexity is independent of the filter size, while in [15] the complexity and filter size are inversely related. In Sec. 4, we show that for the filter sizes needed for accurate object/stereo labeling, the domain transform approach can allow us to achieve even faster inference times than using [15].

[2] Although the updates are conceptually parallel in form, the permutohedral lattice convolution is implemented sequentially.

3 Inference in Models with Higher-order Terms

We now describe how a number of types of higher-order potential may be incorporated in fully connected models of the kind described in Sec. 2, while continuing to permit efficient mean-field updates. The introduction of such higher-order terms not only greatly expands the expressive power of such densely connected models, but also makes efficient filter-based inference possible in a range of models where other techniques are currently used. We show in our experimentation that filter-based inference generally outperforms the best alternative methods in terms of speed and accuracy.

We first give a general form of the models we will be dealing with. In place of Eq. 2, we consider the general energy:

$$E(\mathbf{V}|\mathbf{I}) = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{v}_c|\mathbf{I}) \tag{6}$$

where $\mathbf{V}$ is a joint assignment of the random variables $\mathcal{V} = \{V_1, ..., V_{N_V}\}$, $\mathcal{C}$ is a set of cliques each consisting of a subset of random variables $c \subseteq \mathcal{V}$, and associated with a potential function ψ_c over settings of the random variables in c, $\mathbf{v}_c$. In Sec. 2 we have that $\mathcal{V} = \mathcal{X}$, that each X_i takes values in the set $\mathcal{L}$ of object labels, and that $\mathcal{C}$ contains unary and pairwise cliques of the types discussed. In general, in the models discussed below we will have that $\mathcal{X} \subseteq \mathcal{V}$, so that $\mathcal{V}$ may also include other random variables (e.g. latent variables) which may take values in different label sets, and $\mathcal{C}$ may also include higher-order cliques.

The general form of the mean-field update equations (see [16]) is:

$$Q_i(v_i = \nu) = \frac{1}{Z_i} \exp\{-\sum_{c \in \mathcal{C}} \sum_{\{\mathbf{v}_c | v_i = \nu\}} Q_{c-i}(\mathbf{v}_{c-i}) \cdot \psi_c(\mathbf{v}_c)\} \tag{7}$$

where ν is a value in the domain of random variable v_i, $\mathbf{v}_c$ denotes an assignment of all variables in clique c, $\mathbf{v}_{c-i}$ an assignment of all variables apart from V_i, and Q_{c-i} denotes the marginal distribution of all variables in c apart from V_i derived from the joint distribution Q. $Z_i = \sum_\nu \exp\{-\sum_{c \in \mathcal{C}} \sum_{\{\mathbf{v}_c | v_i = \nu\}} Q_{c-i}(\mathbf{v}_{c-i}) \cdot \psi_c(\mathbf{v}_c)\}$ is a normalizing constant for random variable v_i. We note that the summations $\sum_{\{\mathbf{v}_c | v_i = \nu\}} Q_{c-i}(\mathbf{v}_{c-i}) \cdot \psi_c(\mathbf{v}_c)$ in Eq. 7 evaluate the expected value of ψ_c over Q given that V_i takes the value ν. The updates for the densely connected pairwise model in Eq. 4 are derived by evaluating Eq. 7 across the unary and pairwise potentials defined in Sec. 2 for $v_i = x_{1...N}$ and $\nu = 1...L$. We describe below how similar updates can be efficiently calculated for each of the higher-order potentials we consider.

Pattern-based Potentials: In [11], a *pattern-based* potential[3] is defined as:

$$\psi_c^{\text{pat}}(\mathbf{x}_c) = \begin{cases} \gamma_{\mathbf{x}_c} & \text{if } \mathbf{x}_c \in \mathcal{P}_c \\ \gamma_{\max} & \text{otherwise} \end{cases} \tag{8}$$

[3] The class of such sparse higher-order potentials is also considered in [17].

where $\mathcal{P}_c \subset \mathcal{L}^{|c|}$ is a set of recognized *patterns* (i.e. label configurations for the clique) each associated with an individual cost $\gamma_{\mathbf{x}_c}$, while a common cost $\gamma_{\max}$ is applied to all other patterns. We assume $|\mathcal{P}_c| << L^{|c|}$, since when $|\mathcal{P}_c| \approx L^{|c|}$ the representation approaches an exhaustive parametrization of $\psi_c(\mathbf{x}_c)$.

Given higher-order potentials $\psi_c^{\text{pat}}(\mathbf{x}_c)$ of this form, the required expectation for the mean-field updates (Eq. 7) can be calculated:

$$\sum_{\{\mathbf{x}_c|x_i=l\}} Q_{c-i}(\mathbf{x}_{c-i}) \cdot \psi_c^{\text{pat}}(\mathbf{x}_c) = \sum_{p \in \mathcal{P}_{c|i=l}} \Big(\prod_{j \in c, j \neq i} Q_j(x_j = p_j) \Big) \gamma_p + \Big(1 - \Big(\sum_{p \in \mathcal{P}_{c|i=l}} \Big(\prod_{j \in c, j \neq i} Q_j(x_j = p_j) \Big) \Big) \Big) \gamma_{\max} \tag{9}$$

where we write $\mathcal{P}_{c|i=l}$ for the subset of patterns in $\mathcal{P}_c$ for which $x_i = l$. Since the expectation in Eq. 9 can be calculated in $O(|\mathcal{P}_c||c|)$ time, such terms contribute $O(\max_c(|\mathcal{P}_c||c|)|\mathcal{C}^{\text{pat}}|)$ to each parallel update, where $\mathcal{C}^{\text{pat}}$ is the set of pattern-based clique potentials.[4] If we assume each pixel belongs to at most M^{pat} cliques, and each clique has at most $P^{\max}$ patterns, this complexity reduces to $O(M^{\text{pat}} N P^{\max})$.

A particular case of the pattern-based potential is the P^n-Potts model [7]:

$$\psi_c^{\text{potts}}(\mathbf{x}_c) = \begin{cases} \gamma_l & \text{if } \forall i \in c,\ x_i = l \\ \gamma_{\max} & \text{otherwise} \end{cases} \tag{10}$$

where implicitly we have set $\mathcal{P}$ to be the L configurations with constant labelings. The required expectations here can be expressed as:

$$\sum_{\{\mathbf{x}_c|x_i=l\}} Q_{c-i}(\mathbf{x}_{c-i}) \cdot \psi_c^{\text{potts}}(\mathbf{x}_c) = \Big(\prod_{j \in c, j \neq i} Q_j(x_j = l) \Big) \gamma_l + \Big(1 - \Big(\prod_{j \in c, j \neq i} Q_j(x_j = l) \Big) \Big) \gamma_{\max} \tag{11}$$

which contribute $O(L \max_c(|c|)|\mathcal{C}^{\text{potts}}|)$ to each parallel update. Assuming each pixel belongs to at most M^{pat} cliques, we can reexpress this as $O(M^{\text{pat}} NL)$, which effectively preserves the $O(MNL^2)$ complexity of the dense pairwise updates of Sec. 2 (assuming $M^{\text{pat}} \approx M$), and further preserves the $O(MNL)$ complexity when the pairwise terms also use Potts models. Further potentials which can be cast as pattern-based potentials are discussed in [11], including second-order smoothness priors for stereo, as in [9].

[4] Eq. 9 requires evaluation of the joint probability of $c-1$ variable assignments for each of the $|\mathcal{P}_c|$ patterns, leading to the complexity $O(|\mathcal{P}_c||c|)$ for a single evaluation. If Q is prevented from taking the values 0 and 1, the joint pattern probabilities $\prod_{j \in c} Q_j(x_j = p_j)$ can be calculated once for each clique, and the conditional forms $\prod_{j \in c, j \neq i} Q_j(x_j = p_j)$ needed for parallel updates can then be derived by dividing by $Q_i(x_i = p_i)$, leading to the overall $O(\max_c(|\mathcal{P}_c||c|)|\mathcal{C}^{\text{pat}}|)$ complexity.

Co-occurrence Potentials: Co-occurrence relations capture global information about which classes tend to appear together in an image and which do not, for instance that busses tend to co-occur with cars, but tables do not co-occur with aeroplanes. A recent formulation [8] which has been proposed attempts to capture such information in a global *co-occurrence potential* defined over the entire image clique c_I (generalization to arbitrary cliques is also possible) as:

$$\psi_{c_I}^{\text{cooc}}(\mathbf{X}) = C(\Lambda(\mathbf{X})) \tag{12}$$

Here, $\Lambda(\mathbf{X}) \subseteq \mathcal{L}$ returns the subset of labels present in configuration $\mathbf{X}$, and $C(.) : 2^{\mathcal{L}} \rightarrow \mathbb{R}$ associates a cost with each possible subset. In [8] the restriction is placed on $C(.)$ that it should be non-decreasing with respect to the inclusion relation on $2^{\mathcal{L}}$, i.e. $\Lambda_1, \Lambda_2 \subseteq \mathcal{L}$ and $\Lambda_1 \subseteq \Lambda_2$ implies that $C(\Lambda_1) \leq C(\Lambda_2)$. We will place the further restriction that $C(.)$ can be represented in the form:

$$C(\Lambda) = \sum_{l \in \mathcal{L}} C_l \cdot \Lambda^l + \sum_{l_1, l_2 \in \mathcal{L}} C_{l_1, l_2} \cdot \Lambda^{l_1} \cdot \Lambda^{l_2} \tag{13}$$

where we write Λ^l for the indicator $[l \in \Lambda]$, where $[.]$ is 1 for a true condition and 0 otherwise. Equivalently, Λ^l is the l'th entry of a binary vector of length $|\mathcal{L}|$ which represents Λ by its set-indicator function, and $C(\Lambda)$ is a second degree polynomial over these vectors. Eq. 13 is the form of $C(.)$ investigated experimentally in [8], and is shown perform well there on object class segmentation.

We consider below two approximations to Eq. 12 which give rise to efficient mean-field updates when incorporated in fully connected CRFs as discussed in Sec. 2. Both approximations make use of a set of new latent binary variables $\mathcal{Y} = \{Y_1, ..., Y_L\}$, whose intended semantics are that $Y_l = 1$ will indicate that label l is present in a solution, and $Y_l = 0$ that it is absent. As discussed below though, both approximations enforce this only as a soft constraint. In the first, we reformulate Eq. 12 as:

$$\begin{aligned}\psi_{c_I}^{\text{cooc-1}}(\mathbf{X}, \mathbf{Y}) = C(\{l | Y_l = 1\}) + K \cdot \sum_l [Y_l = 1 \wedge (\sum_i [x_i = l]) = 0] \\ + K \cdot \sum_l [Y_l = 0 \wedge (\sum_i [x_i = l]) > 0]\end{aligned} \tag{14}$$

We consider constructing two CRF distributions $P_1(\mathbf{V}_1|\mathbf{I})$ and $P_2(\mathbf{V}_2|\mathbf{I})$ over the variables sets $\mathcal{V}_1 = \mathcal{X}$ and $\mathcal{V}_2 = \{\mathcal{X}, \mathcal{Y}\}$ respectively, where the clique structure is the same in both distributions, except that a potential $\psi_{c_I}^{\text{cooc}}$ in P_1 has been replaced by $\psi_{c_I}^{\text{cooc-1}}$ in P_2. If we set $K = \infty$ in Eq. 14, the marginals across $\mathbf{X}$ in P_2 will match P_1: $P_1(\mathbf{X}|\mathbf{I}) = \sum_{\mathbf{Y}} P_2(\mathbf{X}, \mathbf{Y}|\mathbf{I})$, since the only joint configurations with non-zero probability in P_2 have identical energies. In general this will not be the case; however, for high K, we can expect that these distributions to approximately match, and hence to be able to perform approximate MPM inference using Eq. 14 in place of Eq. 12.

An alternative, looser approximation to Eq. 12 can be given as:

$$\psi_{c_I}^{\text{cooc-2}}(\mathbf{X}, \mathbf{Y}) = C(\{l | Y_l = 1\}) + K \cdot \sum_{i,l} [Y_l = 0 \wedge x_i = l] \tag{15}$$

using the same latent binary variables $Y_1, ..., Y_L$ introduced in Eq. 14. Setting $K = \infty$ in Eq. 15 does not result in matching marginals in the CRF distributions $P_1(\mathbf{V}_1|\mathbf{I})$ and $P_2(\mathbf{V}_2|\mathbf{I})$ (see above) as it did with Eq. 14. Since the constraint $Y_l = 1 \Rightarrow \sum_i [x_i = l] > 0$ is not enforced by Eq. 15, the marginalization for a given $\mathbf{X}$ configuration in P_2 will be across all settings of $\mathbf{Y}$ that include $\Lambda(\mathbf{X})$. Since there are more of these for configurations when $|\Lambda(\mathbf{X})|$ is small than when it is large, this will tend to make configurations with smaller label sets more probable, and those with larger label sets less so, thus accentuating the minimum description length (MDL) regularization implicit in the original cost function, $C(\Lambda(\mathbf{X}))$ (see [8]). For large K (i.e. $K \neq \infty$), we can thus expect similar distortions.

We give below only the expectation calculations for updates based on $\psi_{c_I}^{\text{cooc-2}}$. Those for $\psi_{c_I}^{\text{cooc-1}}$ can be calculated similarly (as shown in the supplementary material), and we compare the empirical performance of both approximations in Sec. 4. First, for the latent variables Y_l the required expectations are:

$$\sum_{\{\mathbf{V}|Y_l=b\}} Q_{\mathcal{V}-Y_l}(\mathbf{V}-Y_l)\cdot\ \psi_{c_I}^{\text{cooc-2}}(\mathbf{V}) = \begin{cases} K \cdot \sum_i Q_i(x_i = l) + \kappa & \text{if } b = 0 \\ C_l + \sum_{l' \neq l} Q_{l'}(Y_{l'} = 1) C_{l,l'} + \kappa & \text{if } b = 1 \end{cases} \quad (16)$$

where we write $\mathbf{V} - Y_l$ for a setting of all random variables $\mathcal{V}$ apart from Y_l (i.e. $\{\mathbf{X}, \mathbf{Y}_{l' \neq l}\}$), $Q_{\mathcal{V}-Y_l}$ for the marginalization of Q across these same variables, $b \in \{0, 1\}$ is a boolean value, and κ is a constant which can be ignored in the mean-field updates since it is common to both settings of Y_l. Substituting these into Eq. 7, we have the following latent variable updates:

$$\begin{aligned} Q_l(Y_l = 0) &= \frac{1}{Z_l} \exp\{-K \cdot \sum_i Q_i(x_i = l)\} \\ Q_l(Y_l = 1) &= \frac{1}{Z_l} \exp\{-C_l - \sum_{l' \neq l} Q_{l'}(Y_{l'} = 1) C_{l,l'}\} \end{aligned} \quad (17)$$

For the variables X_i, we have the expectations:

$$\sum_{\{\mathbf{V}|X_i=l\}} Q_{\mathcal{V}-X_i}(\mathbf{V}-X_i) \cdot \psi_{c_I}^{\text{cooc-2}}(\mathbf{V}) = K \cdot Q_l(Y_l = 0) + \kappa \quad (18)$$

where κ is again a common constant. Evaluation of each expectation in Eq. 17 requires $O(N + L)$ time, while each expectation in Eq. 18 is $O(1)$. The overall contribution to the complexity of parallel updates for $\psi_{c_I}^{\text{cooc-2}}$ is thus $O(NL+L^2)$, as can also be shown for $\psi_{c_I}^{\text{cooc-1}}$. This does not increase on the complexity of $O(MNL^2)$ for fully connected pairwise updates as in Sec. 2.

4 Experiments

We demonstrate our approach on two labeling problems including higher-order potentials, joint stereo and object labeling and object class segmentation, adapt-

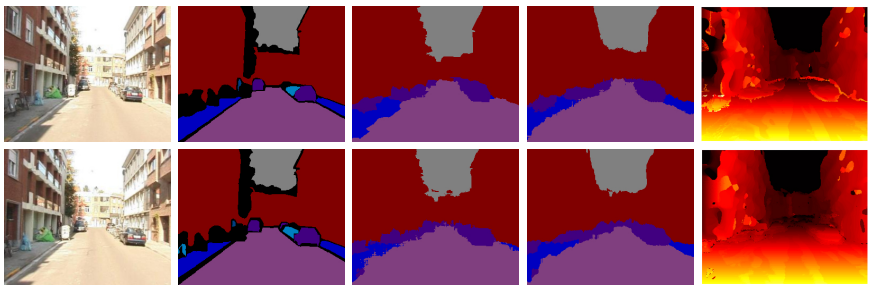

Fig. 1. Qualitative results on Leuven dataset. From left to right: input image, ground truth, object labeling from [13] (using graph-cut + range-moves for inference), object labeling and stereo outputs from our dense CRF with higher-order terms and extended cost-volume filtering (see text).

ing models which have been proposed independently. Details of the experimental set-up and results are provided below. In all experiments, timings are based on code run on an Intel(R) Xeon(R) 3.33 GHz processor, and we fix the number of full mean-field update iterations to 5 for all models.

Joint Stereo and Object Labeling: We adapt a model for joint stereo and object labeling, as proposed originally in [13]. The model can be expressed in terms of a CRF over two sets of random variables, $\mathcal{V} = \{\mathcal{X},\mathcal{U}\}$, conditioned on a pair of images, $P(\mathbf{V}|\mathbf{I}_1,\mathbf{I}_2)$. $\mathcal{X} = \{X_1,...,X_N\}$ and $\mathcal{U} = \{U_1,...,U_N\}$ each range over pixels $i = 1...N$ in image $\mathbf{I}_1$, where X_i takes values in $\mathcal{L} = \{1...L\}$ representing the object present at each pixel, and U_i takes values in $\mathcal{D} = \{1...D\}$ representing the disparity between pixel i in $\mathbf{I}_1$ and a proposed match in $\mathbf{I}_2$. We introduce dense pairwise connections between the variables $X_{1...N}$ and $U_{1...N}$ as in Sec. 2, as well as a set of P^n-Potts higher-order potentials over $\mathcal{X}$, as described in Sec. 3. The P^n-Potts potentials are set as follows: we run meanshift segmentation [18] over image $\mathbf{I}_1$ at a fixed resolution, and create a clique c from the variables X_i falling within each segment returned by the algorithm. We represent the joint unary potential in [13] by separate unary potentials $\psi_u(x_i = l)$ and $\psi_u(u_i = d)$ over the object labels and the disparity labels respectively and a connecting pairwise potential $\psi_p(x_i = l, u_i = d)$. As discussed, for our mean-field model we replace the 8-connected pairwise structure on $\mathcal{X}$ and $\mathcal{U}$ with dense connectivity. We disregard the joint pairwise term over the product space $\psi_p(x_i = l_1, u_i = d_1, x_j = l_2, u_j = d_2)$ proposed in [13]. The mean-field updates for $Q_i(x_i = l)$ are calculated as in Eq. 4, with additional terms for the P^n-Potts model expectations (Eq. 11) and pairwise expectations for the joint potentials $\psi_p(x_i, u_i)$. Updates for $Q_i(u_i = d)$ are similar, but without higher-order terms.

The model is applied to the Leuven dataset [13], consisting of stereo images of street scenes, with ground truth labeling for 7 object classes, and manually annotated ground truth stereo labelings quantized into 100 disparity labels. We use identical training and test sets to [13]. The parameters of the model are set

Algorithm	Time (s)	Object(% correct)	Stereo(% correct)
GC+Range(1) [13]	24.6	95.94	76.97
GC+Range(2) [13]	49.9	95.94	77.31
GC+Range(3) [13]	74.4	95.94	77.46
Extended CostVol ([15] filter)	4.2	95.20	77.18
Dense+HO ([15] filter)	3.1	95.24	78.89
Dense+HO ([14] filter)	2.1	95.06	78.21
Dense+HO+CostVol ([14] filter)	6.3	94.98	79.00

Table 1. Quantitative comparison on Leuven dataset. The table compares the average time per image and performance (Object and Stereo labeling accuracy) of joint object and stereo labeling algorithms, using graph-cut + range-moves (GC+Range(x), where range moves to disparity values $d \pm x$ are allowed for fixed d at each iteration) [13], an extension of cost-volume filtering (see text), and our dense CRF with higher-order terms and filter-based inference (with and without cost-volume filtered unaries, and using different filtering approaches, see text). Our Dense+HO approach achieves comparable accuracies to [13], and is an order of magnitude faster. The best stereo accuracies occur when our model is combined with cost-volume filtered unary potentials for disparity.

as follows. As in [13], for our basic model we use JointBoost classifier responses to form the object unary potentials $\psi_u(x_i = l)$ [19]. A truncated l_2-norm of the intensity differences is used to form the disparity potentials $\psi_u(u_i = d)$ (using the interpolation technique described in [3]), while the potentials $\psi_p(x_i = l, u_i = d)$ are set according to the observed distributions of object heights in the training set. For the densely connected pairwise terms over $\mathcal{X}$ and $\mathcal{U}$, we use identical kernels and weightings to [4] and an Ising model for the label compatibility function, $\mu(l_1, l_2) = [l_1 \neq l_2]$. For the P^n-Potts potentials, we set $\gamma_l = 0$ for all $l = 1...L$, and set $\gamma_{\max}$ by cross-validation.

In addition to the model as described above, we also investigate an alternative approach to setting the unary potentials for the disparity variables based on the cost-volume filtering framework of [6], which we extend to operate in the product label space $L \times D$, i.e. assigning a cost $\lambda_i^t(l, d)$ for each object-disparity combination at pixel i over a series of update steps $t = 0...T$. We initialize the costs to $\lambda_i^0(l, d) = \psi_u(x_i = l) + \psi_u(u_i = d) + \psi_p(x_i = l, u_i = d) + \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j^0(l, d) \cdot \psi_p(x_i, x_j)$, where $Q_i^0(l, d) = 1/L$ for all i, l, d. We then update the costs at each iteration via independent mean-field updates across the D cost-volumes $\lambda(., d)$, $d = 1...D$, using the same kernel and label compatibility function settings as described above; hence, we set $Q_i^{t+1}(l, d) = (\exp(-\lambda_i^t(l, d)))/(\sum_{l'} \exp(-\lambda_i^t(l', d)))$, and $\lambda_i^{t+1}(l, d) = \psi_u(x_i = l) + \psi_u(u_i = d) + \psi_p(x_i = l, u_i = d) + \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j^{t+1}(l, d) \cdot \psi_p(x_i, x_j)$. The output costs are then given by $\lambda_i^T(l, d)$. We form updated disparity unary potentials for the full model by adding the maximum across the output costs to the original potential output: $\psi'_u(u_i = d) = \max_l \lambda_i^T(l, d) + \psi_u(u_i = d)$.

We compare results from the following methods. As our baseline, we use the method of [13], whose CRF structure is similar to ours, but without dense

connectivity over $\mathcal{X}$, and with a truncated L_1-prior on the disparity labels $\mathcal{U}$. Inference is performed by alternating alpha-expansion on $\mathcal{X}$ with range moves on $\mathcal{U}$ (forming *projected moves*, see [13]). Since the speed and accuracy are affected by the size of range moves considered, we test 3 settings of the range parameter, corresponding to moves to disparity values $d \pm 1$, $d \pm 2$ and $d \pm 3$, for a fixed d at each iteration (see [20]). We consider also a baseline based on the extended cost-volume filtering approach outlined above where we simply select $(x_i, u_i) = \text{argmax}_{(l,d)} \lambda_i^T(l, d)$ as output. We compare these with our basic higher-order model with full connectivity as described above, and our model combined with extended cost-volume filtered disparity unary terms ψ'_u as described. Further, using our basic model we compare two alternative filtering methods for inference, the first using the permutohedral lattice, as in [4, 15], and the second using the domain transform based filtering method of [14]. We evaluate the average time for inference, and the %-correct pixels for stereo and object labeling, where a disparity is considered correct if it is within 5 pixels of the ground truth.

Qualitative and quantitative results are shown in Fig. 1 and Tab. 1 respectively. We note that the densely connected CRF with higher-order terms (Dense+HO) achieves comparable accuracies to [13], and that the use of domain transform filtering methods [14] permits an extra speed up, with inference being almost 12 times faster than the least accurate setting of [13], and over 35 times faster than the most accurate. The extended cost-volume filtering baseline described above also performs comparably well, and at a small extra cost in speed, the combined approach (Dense+HO+CostVol) achieves the best overall stereo accuracies. We note that although the improved stereo performance appears to generate a small decrease in the object labeling accuracy in our full model, the former remains at an almost saturated level, and the small drop could possibly be recovered through further tuning or weight learning.

Object Class Segmentation: We also test our approach on object class segmentation, adapting the Associative Hierarchical CRF (AHCRF) model with a co-occurrence potential proposed in [8]. As described in Sec. 3, we build a CRF over variables $\mathcal{V} = \{\mathcal{X}, \mathcal{Y}\}$, with X_i denoting the object label at pixels $i = 1...N$, and latent binary variable Y_l indicating the presence/absence of label $l = 1...L$. The model includes dense pairwise connections over the X_i variables, as well 10 layers of P^n-Potts potentials formed by running mean shift and K-means clustering algorithms at 5 parameter settings each (coarse-fine) across the test image, and forming a clique $c \in \mathcal{C}$ across variables from $\mathcal{X}$ falling within each returned segment. An additional set of P^n-Potts potentials is also included based on segments returned by grabcut initialized to the bounding boxes returned from detectors trained on each of the L classes (see [8]). A co-occurrence potential is also included, taking the form of either $\psi^{\text{cooc-1}}$ or $\psi^{\text{cooc-2}}$ as in Sec. 3.

We test the model on the PascalVOC-10 training and validation set. We use the same split as used in [4], who randomly partition the available images into 3 groups: 40% training, 15% validation, and 45% test set. Further, we use the unary potentials provided by [4], along with identical kernel weights and parameters, and an Ising label compatibility function $\mu(l_1, l_2) = [l_1 \neq l_2]$. The

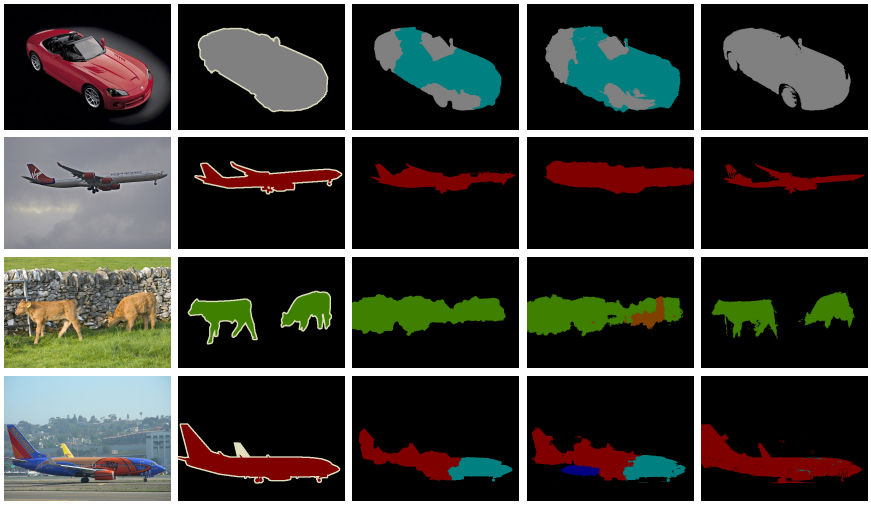

Fig. 2. Qualitative results on PascalVOC-10 dataset. From left to right: input image, ground truth, output from [8] (AHCRF+Cooccurrence), output from [4] (Dense CRF), output from our dense CRF with Potts and Co-occurrence terms.

higher-order potentials are trained piecewise: we train a classifier using Jointboost [19] to classify the segments associated with the P^n-Potts cliques, and set the parameters γ_l in Eq. 10 to be the negative log of the classifier output probabilities, truncated to a fixed value $\gamma_{\max}$ set by cross validation. The parameters of the co-occurrence cost function Eq. 13 are set as in [8], by fitting a second-degree polynomial to the negative logs of the observed frequencies of each subset of labels L occurring in the training data. Individual weights on the potentials are set by cross-validation.

We compare both the timing and performance of four algorithms. As our two baselines, we take the AHCRF with a co-occurrence potential [8], whose model includes all higher-order terms but is not densely connected and uses α-expansion based inference, and the dense CRF [4], which uses filter-based inference but does not include higher-order terms. We compare these with our approach, which adds first P^n-Potts terms to the dense CRF, and then P^n-Potts and co-occurrence terms. We use the permutohedral lattice for filtering in all models. We assess the overall percentage of pixels correctly labeled, the average recall and intersection/union score per class (defined in terms of the true/false positives/negatives for a given class as TP/(TP+FP+FN)).

Qualitative and quantitative results are shown in Fig. 2 and Tab. 2 respectively (further per-class quantitative results are provided in the supplementary material). As shown, our approach is able to outperform both of the baseline methods in terms of the class-average metrics, while also reducing the inference time with respect to the AHCRF with a co-occurrence potential almost

Algorithm	Time (s)	Overall (%-corr)	Av. Recall	Av. I/U
AHCRF+Cooc [8]	36	81.43	38.01	30.9
DenseCRF [4]	0.67	71.63	34.53	28.4
Dense+Potts	4.35	79.87	40.71	30.18
Dense+Potts+Cooc	4.4	80.44	43.08	32.35

Table 2. Quantitative results on PascalVOC-10. The table compares timing and performance of our approach (final 2 lines) against two baselines. The importance of higher-order information is confirmed by the better performance of all algorithms compared to the basic dense CRF of [4]. Further, our filter-based inference is both able to improve substantially on the inference time and class-average performance of the AHCRF [8], with P^n-Potts and co-occurrence potentials each giving notable gains.

by a factor of 9. The results shown are only for our approach with the $\psi^{\text{cooc-2}}$ potential, since we found the $\psi^{\text{cooc-1}}$ potential to suffer from poor convergence properties, with performance only marginally better than [4]. We note that our aim here is to assess the relative performance of our approach with respect to our baseline methods, and we expect that our model will need further refinement to compete with the current state-of-the-art on Pascal (our results are $\sim 9\%$ lower for average intersection/union compared to the highest performing method on the 2011 challenge, see [22]). We also note that [4] are able to further improve their average intersection/union score to 30.2% by learning the pairwise label compatibility function, which remains a possibility for our model also.

5 Discussion

We have introduced a set of techniques for incorporating higher-order terms into densely connected multi-label CRF models. As described, using our techniques, bilateral filter-based methods remain possible for inference in such models, effectively retaining the mean-field update complexity $O(MNL^2)$ as in [4] when higher-order P^n-Potts models are used. This both increases the expressivity of existing fully connected CRF models, and opens up the possibility of using powerful filter-based inference in a range of models with higher-order terms. We have shown the value of such techniques for both joint object-stereo labeling and object class segmentation. In each case, we have shown substantial improvements in inference speed with respect to graph-cut based methods, particularly by using recent domain transform filtering techniques, while also observing similar or better accuracies. Further results in the supplementary materials appear to show that the improved accuracy is both due to using dense models and mean-field inference; graph-cut methods are substantially slower, and even when the connectivity of the graph is increased cannot achieve similar accuracies to mean-field in matched energies except in models without higher-order terms. Future directions include investigation of further ways to improve efficiency though parallelization, and learning techniques which can draw on high speed inference for

joint parameter optimization in large-scale models. Code for our method is available for download at `http://cms.brookes.ac.uk/staff/VibhavVineet/`.

References

1. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. In: IEEE PAMI. (2006)
2. Komodakis, N., Paragios, N., Tziritas, G.: MRF energy minimization and beyond via dual decomposition. In: IEEE PAMI. (2011)
3. Boykov, Y., Veksler, O., Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts. In: IEEE PAMI. (2001)
4. Krahenbuhl, P., Koltun, V.: Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In: NIPS. (2011)
5. Kornprobst, P., Tumblin, J., Durand, F.: Bilateral Filtering: Theory and Applications. In: Foundations and Trends in Computer Graphics and Vision. (2009)
6. Rhemann, C., Hosni, A., Bleyer, M., Rother, C., Gelautz, M.: Fast cost-volume filtering for visual correspondence and beyond. In: CVPR. (2011)
7. Kohli, P., Kumar, M.P., Torr, P.H.S.: P3 & beyond: Solving energies with higher order cliques. In: CVPR. (2007)
8. Ladickỳ, L., Russell, C., Kohli, P., Torr, P.H.S.: Graph cut based inference with co-occurrence statistics. In: ECCV. (2010)
9. Woodford, O., Torr, P.H.S., Reid, I., Fitzgibbon, A.: Global stereo reconstruction under second-order smoothness priors. In: IEEE PAMI. (2009)
10. Potetz, B., Lee, T.S.: Efficient belief propagation for higher-order cliques using linear constraint nodes. In: CVIU. (2008)
11. Komodakis, N., Paragios, N.: Beyond pairwise energies: Efficient optimization for higher-order MRFs. In: CVPR. (2009)
12. Gonfaus, J.M., Boix, X., Van De Weijer, J., Bagdanov, A.D., Serrat, J., Gonzalez, J.: Harmony potentials for joint classification and segmentation. In: CVPR. (2010)
13. Ladickỳ, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W.F., Torr, P.H.S.: Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction. In: BMVC (2010)
14. Gastla, E.S.L., Oliveira, M.M.: Domain transform for edge-aware image and video processing. In: ACM Trans. Graph. (2011)
15. Adams, A., Baek, J., Davis, M.A.: Fast High-Dimensional Filtering Using the Permutohedral Lattice. In: Computer Graphics Forum. (2010)
16. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT Press (2009)
17. Rother, C., Kohli, P., Feng, W., Jia, J.: Minimizing sparse higher order energy functions of discrete variables. In: CVPR. (2009)
18. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. In: IEEE PAMI. (2002)
19. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. In: IEEE PAMI. (2007)
20. Kumar, M.P., Veksler, O., Torr, P.H.S.: Improved Moves for Truncated Convex Models. In: JMLR. (2011)
21. Ladickỳ, L., Sturgess, P., Alahari, K., Russell, C., Torr, P.H.S.: What, where and how many? combining object detectors and crfs. In: ECCV. (2010)
22. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2011 (VOC2011). `http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html`.