Int J Comput Vis (2014) 110:290–307
DOI 10.1007/s11263-014-0708-6

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

**Vibhav Vineet · Jonathan Warrell · Philip H. S. Torr**

Received: 6 June 2013 / Accepted: 25 February 2014 / Published online: 19 March 2014


**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 labelling problems, as well as object class segmentation, showing in addition for joint object-stereo labelling 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 labelling.

**Keywords** Object class segmentation · Dense stereo reconstruction · Mean-field methods · Higher order potentials · Bilateral filters · CRF

Communicated by Carlo Colombo.

Vibhav Vineet and Jonathan Warrell have contributed to this work equally as joint first author.

V. Vineet (✉)
Oxford Brookes University, Oxford, UK
e-mail: vibhav.vineet-2010@brookes.ac.uk;
vibhav.vineet@gmail.com

J. Warrell
MIAS (CSIR), Pretoria , South Africa
e-mail: jwarrell@csir.co.za

P. H. S. Torr
Department of Engineering Science, University of Oxford, Oxford, UK
e-mail: philip.torr@eng.ox.ac.uk

## 1 Introduction

Many computer vision problems, such as object class segmentation, stereo and optical flow, can be formulated as multi-labelling problems within a Markov Random Field (MRF) or Conditional Random Field (CRF) framework. 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 ( Kolmogorov 2006; Komodakis et al. 2011; Boykov et al. 2001). Recently, a number of cross bilateral Gaussian filter-based methods have been proposed for problems such as object class segmentation ( Krahenbuhl and Koltun 2011), denoising (Kornprobst et al. 2009), stereo and optical flow (Rhemann et al. 2011), 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.

A problem with filter-based methods as currently formulated is that they can only be applied to models with limited types of structure. In Rhemann et al. (2011), dependencies between output labels are abandoned, and the filtering step is used to generate unary costs which are treated independently. In Krahenbuhl and Koltun (2011), filtering is used to per-

form 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 labelling problems mentioned. For object class segmentation, the importance of enforcing label consistency over homogeneous regions has been demonstrated using $P^n$-Potts models (Kohli et al. 2007), and co-occurrence relations between classes at the image level have also been shown to provide important priors for segmentation (Ladickỳ et al. 2010). For stereo and optical flow, second-order priors have proved to be effective (Woodford et al. 2009), as have higher-order image priors for denoising (Potetz and Lee 2008).

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 Komodakis and Paragios (2009), Rother et al. (2009)), which include $P^n$-Potts models and second-order smoothness priors, and (b) global potentials representing co-occurrence relationships between labels as in Ladickỳ et al. (2010); Gonfaus et al. (2010). We assume a base-layer MRF with full connectivity and weighted Gaussian edge potentials as in Krahenbuhl and Koltun (2011). 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 labelling as in Ladickỳ et al. (2010), and second a model for object class segmentation with co-occurrence priors as in Ladickỳ et al. (2010). In the case of joint stereo and object labelling, 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 labelling of applying recent *domain transform filtering* techniques (Gastla and Oliveira 2011) in our framework. In both joint stereo-object labelling 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 labelling, again permitting efficient inference,
- An investigation of the advantages/disadvantages of alternative filtering methods recently proposed (Kornprobst et al. 2009; Gastla and Oliveira 2011; Adams et al. 2010) within our framework.

We briefly give details about some of the related work in Sect. 2. In Sect. 3 we review the method of Krahenbuhl and Koltun (2011). Sections 4 and 5 provide details on how we encode higher-order terms and product label spaces respectively, Sect. 6 gives experimentation on joint stereo and object labelling, and object class segmentation. Finally Sect. 7 analyses the mean-field method and Sect. 8 concludes with a discussion.

## 2 Related Work

Over the last few years many different methods have been proposed for the problem of object class segmentation, which assigns an object label such as *road* or *building* to every pixel in the image.

First we briefly review some of the interactive algorithms before going into the automatic algorithms. In the interactive segmentation case, the algorithms are guided by the user-defined seed pixels corresponding to different labels for segmenting out the objects of interest. Rother et al. (2004)) and Boykov and Jolly (2001) proposed graph-cuts based approaches to do interactive segmentation. But graph-cuts based methods generally suffer from the problem of the shrinkage bias (bias towards shorter boundaries). In order to overcome this issue, Leo Grady proposed a random walk based method (Grady 2006) for multilabel interactive image segmentation. They analytically determine the probability that an unlabelled pixel would reach one of the labelled pixel which helps them to decide the label of each unlabelled pixel. But the random walk approach suffers from the problem of sensitivity to location of pixels labelled by the users. Singaraju et al. (2008) proposed a continuous MRF based formulation, a hybrid of the graph-cuts based approach and the random walker, to recover from these issues. Another interesting approach for segmentation is based on the geodesic distance of each pixel to the user-provided seed pixels (Criminisi et al. 2008; Bai and Sapiro 2007). While these approaches efficiently solved the problem of interactive segmentation, in this work we focus on automatic segmentation.

Over the years several interesting algorithms have also been developed for automatic object class segmentation. Many of these algorithms integrate information from various sources such as top-down object-specific knowledge and bottom-up pixel level knowledge to improve the accuracy

(Borestein and Malik 2006; Kumar et al. 2005). Though these approaches work well on some challenging dataset, they generally fail while dealing with large number of classes. Shotton et al. (2009) proposed *TextonBoost* approach to overcome this issue. While this approach produce good results for the object class segmentation problem, they fail to capture/enforce higher order constraints on the output label space.

Many works have shown the importance of incorporating higher order constraints on the label space. In order to solve CRF with the higher order constraints, there are two broad classes of methods: graph-cuts based methods and message passing approaches. Lan et al. (2009) proposed approximate belief propagation for efficient inference in higher order MRFs. Following this, Potetz (Potetz and Lee 2008) showed how belief propagation can be efficiently performed in graphical models containing moderately large cliques. However, as these methods were based on BP, they were quite slow and took minutes or hours to converge. Kohli et al. 2007 designed graph-cuts based efficient method to incorporate higher order potential in their CRF framework, and showed how certain classes of higher order potentials can be minimized using move making algorithm such as $\alpha$-expansion method (Boykov et al. 2001). This is followed by the work of Ladickỳ et al. (2009, 2010) who showed how context and detector based higher level knowledge can be incorporated in the CRF framework. These approaches have produced excellent results, but they typically involve high time complexity to be applicable for any real-time object class segmentation and recognition.

The second part of our work deals with efficiently solving the problem of jointly estimating the object-stereo labels at the pixel level. This problem has also been studied in the past and some of the previous research has tried to develop efficient algorithms for this problem. The most related is the work of Ladickỳ et al. (2010) who formulated the problem in a CRF framework and used graph-cuts based range-move approaches (Veksler 2007; Pawan Kumar and Torr 2008) to solve the problem efficiently. Further, Bleyer et al. (2011) also proposed method to jointly estimate the object and disparity labels at the pixel level. The main difference between these two works is that the previous work requires already trained models of different object classes, while the later one performs an unsupervised segmentation approach (Comaniciu and Meer 2002) to extract a set of object segments. Bleyer et al. (2012) further improved the object-stereo output by incorporating the scene-consistent 3D prior knowledge to improve the stereo output. While these approaches work in discrete label space, Goldlucke and Cremers (2010) proposed a convex relaxation approach which allows to cast the joint object-stereo problem in terms of convex optimization problems. Though these approaches have produced good results, they still suffer from high time complexity.

While most of these related works have focussed on limited 4 or 8 connectivity, several works in the past have incorporated dense pairwise connections to capture context information. In this paper, we follow such line of research where we have fully connected CRF that allows to enforce pairwise costs on all pairs of pixels in the image. Such fully connected CRFs have been used for semantic image labeling in the past (Rabinovich et al. 2007; Toyoda and Hasegawa 2008; Galleguillos et al. 2008; Payet and Todorovic 2010). Though these approaches motivate us to incorporate dense pairwise connections, the complexity of their inference in such fully connected models restricted their applications to small sets of pixels/regions/variables. Finally, there is an interesting work by Krahenbuhl and Koltun (2011) who proposed a highly efficient algorithm to perform inference in the fully connected pairwise CRF for certain kind of pairwise potentials. We give detailed description of their approach below.

## 3 Filter-Based Inference in Dense Pairwise CRFs

We begin by reviewing the approach of Krahenbuhl and Koltun (2011), 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 ( Krahenbuhl and Koltun 2011) restrict the pairwise potentials to take the form of a weighted mixture of Gaussian kernels:

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

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

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 ( Krahenbuhl and Koltun 2011) 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, ( Krahenbuhl and Koltun 2011) 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 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 Koller and Friedman (2009) for a derivation):

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

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 (Koller and Friedman 2009). In Krahenbuhl and Koltun (2011), 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 Adams et al. (2010) (which introduces a small approximation). This is achieved by the following transformation:

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

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 (Adams et al. 2010) 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 (Koller and Friedman 2009). However, ( Krahenbuhl and Koltun 2011) 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 Krahenbuhl and Koltun (2011) use the permutohedral lattice (Adams et al. 2010) 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 (Gastla and Oliveira 2011) 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 Adams et al. (2010) the complexity and filter size are inversely related. In Sect. 6, we show that for the filter sizes needed for accurate object/stereo labelling, the domain transform approach can allow us to achieve even faster inference times than using Adams et al. (2010).

## 4 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 Sect. 3, 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:

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

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 $\psi_c$ over settings of the random variables in $c$, $\mathbf{v}_c$. In Sect. 3 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

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

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 Koller and Friedman (2009)) 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 $\nu$ 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 $\psi_c$ over $Q$ given that $V_i$ takes the value $\nu$. 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 Sect. 3 for $v_i = x_{1\ldots N}$ and $\nu = 1 \ldots L$. We describe below how similar updates can be efficiently calculated for each of the higher-order potentials we consider.

### 4.1 Pattern-Based Potentials

In Komodakis and Paragios (2009), a *pattern-based* potential[3] is defined as:

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

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:

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

[3] The class of such sparse higher-order potentials is also considered in Rother et al. (2009).

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 $M^{\text{pat}}$ 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 (Kohli et al. 2007):

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

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

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

which contribute $O(L \max_c(|c|)|\mathcal{C}^{\text{potts}}|)$ to each parallel update. Assuming each pixel belongs to at most $M^{\text{pat}}$ 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 Sect. 3 (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 Komodakis and Paragios (2009), including second-order smoothness priors for stereo, as in Woodford et al. (2009).

### 4.2 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 (Ladickỳ et al. 2010) 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

[4] Equation 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.

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 Ladický et al. (2010) 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 $\Lambda^l$ for the indicator $[l \in \Lambda]$, where $[.]$ is 1 for a true condition and 0 otherwise. Equivalently, $\Lambda^l$ is the $l$'th entry of a binary vector of length $|\mathcal{L}|$ which represents $\Lambda$ by its set-indicator function, and $C(\Lambda)$ is a second degree polynomial over these vectors. Equation 13 is the form of $C(.)$ investigated experimentally in Ladický et al. (2010), 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 Sect. 3. 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.

#### 4.2.1 Model 1

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.

With this approximation, the relevant expectations over the latent variables $Y_1, ..., Y_L$ can be calculated as:

$$\begin{aligned}&\sum_{\{\mathbf{V}|Y_l=b\}} Q_{\mathcal{V}-Y_l}(\mathbf{V} - Y_l) \cdot \psi_{c_I}^{\text{cooc-1}}(\mathbf{V}) \\ &= \begin{cases} K \cdot (1 - \prod_i (1 - Q_i(x_i = l))) + \kappa & \text{if } b = 0 \\ C_l + \sum_{l' \neq l} Q_{l'}(Y_{l'} = 1) C_{l,l'} & \\ \quad + K \cdot \prod_i (1 - Q_i(x_i = l)) + \kappa & \text{if } b = 1 \end{cases}\end{aligned} \tag{15}$$

leading to the following mean-field updates for the latent variable distributions:

$$\begin{aligned}Q_l(Y_l = 0) &= \frac{1}{Z_l} \exp\left\{-K \cdot \left(1 - \prod_i (1 - Q_i(x_i = l))\right)\right\} \\ Q_l(Y_l = 1) &= \frac{1}{Z_l} \exp\left\{-C_l - \sum_{l' \neq l} Q_{l'}(Y_{l'} = 1) C_{l,l'} \right. \\ &\qquad \left. - K \cdot \prod_i (1 - Q_i(x_i = l))\right\}\end{aligned} \tag{16}$$

where the expectations can be calculated in $O(N + L)$ time. Further, the expectations for variables $X_i$ can be expressed:

$$\begin{aligned}&\sum_{\{\mathbf{V}|X_i=l\}} Q_{\mathcal{V}-X_i}(\mathbf{V} - X_i) \cdot \psi_{c_I}^{\text{cooc-1}}(\mathbf{V}) \\ &= K \cdot Q_l(Y_l = 0) \\ &\quad + K \cdot \sum_{l' \neq l} Q_{l'}(Y_{l'} = 0) \left(1 - \prod_{j \neq i} (1 - Q_j(x_j = l'))\right) \\ &\quad + K \cdot \sum_{l' \neq l} Q_{l'}(Y_{l'} = 1) \prod_{j \neq i} (1 - Q_j(x_j = l')) + \kappa\end{aligned} \tag{17}$$

which require $O(NL)$ time. This would seem to imply a contribution of $O(NL^2)$ for the cooc-1 terms towards the full parallel update. However, by computing the full products $\prod_i (1 - Q_i(x_i = l))$ once for each $l$, and then dividing by the relevant terms to calculate the partial products in Eq. 21 (we must ensure $Q$ does not take the extreme values 0 and 1 during updates to do this) a complexity of $O(NL + L^2)$ is achieved.

#### 4.2.2 Model 2

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

$$\psi_{c_I}^{\text{cooc-2}}(\mathbf{X}, \mathbf{Y}) = C(\{l | Y_l = 1\}) + K \cdot \sum_{i,l} [Y_l = 0 \wedge x_i = l] \tag{18}$$

using the same latent binary variables $Y_1, ..., Y_L$ introduced in Eq. 14. Setting $K = \infty$ in Eq. 18 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. 18, 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 Ladickỳ et al. (2010)). For large $K$ (i.e. $K \neq \infty$), we can thus expect similar distortions. Thus, for the latent variables $Y_l$ the required expectations are:

$$\sum_{\{\mathbf{V}|Y_l=b\}} Q_{\mathcal{V}-Y_l}(\mathbf{V}-Y_l)\cdot\psi_{c_I}^{\text{cooc-2}}(\mathbf{V}) = \begin{cases} K\cdot\sum_i Q_i(x_i=l)+\kappa & \text{if } b=0 \\ C_l+\sum_{l'\neq l} Q_{l'}(Y_{l'}=1)C_{l,l'}+\kappa & \text{if } b=1 \end{cases} \tag{19}$$

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 $\kappa$ 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:

$$\begin{aligned} Q_l(Y_l=0) &= \frac{1}{Z_l}\exp\{-K\cdot\sum_i Q_i(x_i=l)\} \\ Q_l(Y_l=1) &= \frac{1}{Z_l}\exp\{-C_l-\sum_{l'\neq l} Q_{l'}(Y_{l'}=1)C_{l,l'}\} \end{aligned} \tag{20}$$

For the variables $X_i$, we have the expectations:

$$\sum_{\{\mathbf{V}|X_i=l\}} Q_{\mathcal{V}-X_i}(\mathbf{V}-X_i)\cdot\psi_{c_I}^{\text{cooc-2}}(\mathbf{V}) = K\cdot Q_l(Y_l=0)+\kappa \tag{21}$$

where $\kappa$ is again a common constant. Evaluation of each expectation in Eq. 20 requires $O(N+L)$ time, while each expectation in Eq. 21 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 Sect. 3.

## 5 Inference in Models with Product Label Spaces

Now we discuss how we provide an efficient inference method for jointly estimating per-pixel object class and disparity labels. Before going into the details of the joint inference, we briefly describe the specific forms of the energy functions we use, which are based on the model of Ladicky et.al. (Ladickỳ et al. 2010) for joint object and stereo labelling.

For object class segmentation, we define a CRF defined over a set of random variables $\mathcal{X} = \{X_1 \ldots X_N\}$ ranging over pixels $i = 1 \ldots N$ in image $\mathbf{I}_1$, where $X_i$ takes values in $\mathcal{L} = \{1 \ldots L\}$ representing the object present at each pixel. The energy function for the object variables includes the unary, pairwise and higher order terms as described in Sect. 4 as follows:

$$E^O(\mathbf{x}) = \sum_i \psi_u^O(x_i) + \sum_{ij} \psi_p^O(x_i, x_j) + \sum_c \psi_c^O(\mathbf{x}_c|\mathbf{I}) \tag{22}$$

Similarly, we express the stereo CRF by a set of variables $\mathcal{U} = \{U_1 \ldots U_N\}$ ranging over pixels $i = 1 \ldots N$ in the image $\mathbf{I}_1$ and each random variable $U_i$ takes a label in $\mathcal{D} = \{1 \ldots D\}$ representing the disparity between pixel $i$ in $\mathbf{I}_1$ at a fixed resolution, and a proposed match in $\mathbf{I}_2$. We define a multi-class CRF framework for disparity labels using the unary and pairwise energy function as:

$$E^D(\mathbf{u}) = \sum_i \psi_u^D(u_i) + \sum_{ij} \psi_p^D(u_i, u_j) \tag{23}$$

### 5.1 Joint Formulation for Object and Stereo Labelling

Now we describe our model for jointly estimating per-pixel object and stereo labels. In this model, we define a CRF over two sets of variables $\mathcal{V} = \{\mathcal{X}, \mathcal{U}\}$ conditioned on the images, $P(\mathbf{V}|\mathbf{I}_1, \mathbf{I}_2)$. Each random variable $V_i = [X_i, U_i]$ takes a label $v_i = [x_i, u_i]$ from the product label space of object and stereo labels $\mathcal{L} \times \mathcal{D}$ corresponding to the variable $V_i$ taking object label $x_i$ and disparity label $u_i$. In this framework, we define our joint energy function as:

$$E^J(\mathbf{v}) = \sum_i \psi_u^J(v_i) + \sum_{ij} \psi_p^J(v_i, v_j) + \sum_c \psi_c^J(\mathbf{v}_c|\mathbf{I}) \tag{24}$$

where $\psi_u^J$ and $\psi_p^J$ are the joint unary and pairwise terms. We represent the joint unary potential as sum of the object and disparity unary terms, and a connecting pairwise term as:

$$\psi_u^J(v_i) = \psi_u^O(x_i) + \psi_u^D(u_i) + \psi_p(x_i = l, u_i = d) \tag{25}$$

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 Ladickỳ et al. (2010). Further, we define a set of $P^n$-Potts higher order potentials over $\mathcal{X}$, as described in Sect. 4.

### 5.2 Mean-Field Updates

Within this model, the mean-field updates for the object variables, $Q_i^O(x_i = l)$ are calculated as in Eq. 4, with additional terms for the $P^n$-Potts model expectation (Eq. 11) and pairwise expectations for the joint potentials $\psi_p(x_i, u_i)$ as follows:

$$\begin{aligned} Q_i^O(x_i = l) = \frac{1}{Z_i} \exp\{ & -\psi_u^O(x_i) - \\ & \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j^O(x_j = l') \psi_p^O(x_i, x_j) \\ & - \sum_{\mathbf{x}_c | x_i = l} Q_{c-i}^O(\mathbf{x}_{c-i}).\psi_c^{potts}(\mathbf{x}_c) \\ & - \sum_{d' \in \mathcal{D}} Q_i^D(u_i = d').\psi_p(x_i, u_i)\} \end{aligned} \tag{26}$$

The updates for $Q_i^D(u_i = d)$ are similar, but without higher-order terms, take following form:

$$\begin{aligned} Q_i^D(u_i = d) = \frac{1}{Z_i} \exp\{ & -\psi_u^D(u_i) \\ & - \sum_{d' \in \mathcal{D}} \sum_{j \neq i}^{D} Q_j(u_j = d') \psi_p^D(u_i, u_j) \\ & - \sum_{l' \in \mathcal{L}} Q_i^O(x_i = l').\psi_p(u_i, x_i)\} \end{aligned} \tag{27}$$

### 5.3 Cost Volume Filtering

In addition to the model as described above, we also investigate an approach to updating the unary potentials for the disparity variables based on the cost-volume filtering framework of Rhemann et al. (2011). This approach involves building a cost-volume of labels, performing edge-preserving filtering in each of the label slices, and then finally estimating the per-pixel labels based on winner-take all label selection strategy. They achieve good speed-ups without loosing much accuracy on challenging problems such as stereo correspondence and optical flow. We leverage cost-volume filtering techniques to improve our stereo unary potentials by extending this work to operate in the product label space $L \times D$. First, we define a CRF at each of the disparity label slices $d \in \mathcal{D} = \{1...D\}$ in the cost volume including variables by $\mathcal{V}^d = \{V_1^d ... V_N^d\}$, where each variable $V_i^d$ takes a disparity label $d$ and object labels in $\mathcal{L} = \{1...L\}$. The energy function at each of the disparity label slice in the cost volume takes following form:

$$\begin{aligned} E^d(v^d) = & \sum_i \psi_u^O(x_i = l) + \sum_i \psi_u^D(u_i = d) \\ & + \sum_i \psi_p(x_i = l, u_i = d) + \sum_{ij} \psi_p^O(x_i, x_j) \end{aligned} \tag{28}$$

We then introduce mean-field distributions $Q_i^t(l, d)$, which represent the probability of assigning pair of object-disparity combination at pixel $i$ over a series of update steps $t = 0...T$. These updates take following form:

$$\begin{aligned} Q_i^{t+1}(l, d) = \frac{1}{Z_i} \exp\{ & -\psi_u^O(x_i = l) - \psi_u^D(u_i = d) \\ & -\psi_p(x_i = l, u_i = d) \\ & - \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j^t(l, d) \cdot \psi_p^O(x_i, x_j)\} \end{aligned} \tag{29}$$

Further, we set $Q_i^0(l, d) = 1/L$ for all $i, l, d$. At each step, we can derive costs $\lambda_i^t(l, d)$ for each object-disparity assignment to the pixel $i$ which takes the form as:

$$\lambda_i^t(l, d) = -\log(Q_i^{t+1}(l, d)) \tag{30}$$

We update $Q_i^t(l, d)$ and $\lambda_i^t(l, d)$ 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. The final output costs are then given by $\lambda_i^T(l, d)$. We form enhanced disparity unary potentials for the full model by adding the maximum across the output costs to the original potential output: $\psi_u'^D(u_i = d) = \max_l \lambda_i^T(l, d) + \psi_u^D(u_i = d)$.

## 6 Experiments

We demonstrate our approach on two labelling problems including higher-order potentials, joint object-stereo labelling and object class segmentation, adapting 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. In addition, we also evaluate the convergence of our mean-field algorithm after inclusion of Potts and co-occurrence based higher order terms. We show the KL-divergence values between $Q$ and $P$ distributions after each iteration of our mean-field update.

### 6.1 Implementation Details

The parameters of the model are set as follows. As in Ladickỳ et al. (2010), for the joint object-stereo model we use JointBoost classifier responses to form the object unary potentials $\psi_u^O(x_i = l)$ (Torralba et al. 2007). A truncated $l_2$-norm of the intensity differences is used to form the disparity potentials $\psi_u^D(u_i = d)$ (using the interpolation technique described in Boykov et al. (2001)), 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 (see Ladickỳ et al. (2010) for details). For Pascal VOC-10 dataset, we use the

unary potentials provided by Krahenbuhl and Koltun (2011). Further, for both of these datasets, we use densely connected pairwise terms where we use kernels and weightings identical to Krahenbuhl and Koltun (2011) and an Ising model for the label compatibility function, $\mu(l_1, l_2) = [l_1 \neq l_2]$.

For $P^n$-Potts higher-order potentials over $\mathcal{X}$ for the joint object-stereo problem, as described in Sect. 4, we first run meanshift segmentation (Comaniciu and Meer 2002) 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. However, on the PascalVOC dataset, we generate a set of 10 layers of segments where each layer corresponds to one application of unsupervised segmentation with different parameters of mean-shift and KMeans algorithms. This way of generating multiple segments have been found to be useful in dealing with the complex object boundaries (Ladickỳ et al. 2009). Once we have generated these higher order cliques, we train the higher-order potentials in a piecewise manner. We first train a classifier using Jointboost (Torralba et al. 2007) to classify the segments associated with the $P^n$-Potts cliques, and set the parameters $\gamma_l$ in Eq. 10 to be the negative log of the classifier output probabilities, truncated to a fixed value $\gamma_{\max}$ set by cross validation. 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 Ladickỳ et al. (2010)).

A co-occurrence potential is also included for the PascalVOC dataset, which takes the form of either $\psi^{\text{cooc-1}}$ or $\psi^{\text{cooc-2}}$ as in Sect. 4. The parameters of the co-occurrence cost Eq. 13 are set as in Ladickỳ et al. (2010), 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. Finally, individual weights on the potentials are set by cross-validation.

### 6.2 Joint Object and Stereo Labelling

We evaluate the efficiency offered by our mean-field update for joint object-stereo estimation to the Leuven dataset (Ladickỳ et al. 2010). The dataset consists of stereo images of street scenes, with ground truth labelling for 7 object classes, and manually annotated ground truth stereo labellings quantized into 100 disparity labels. We use identical training and test sets to Ladickỳ et al. (2010).

We compare results from the following methods. As our baseline, we use the method of Ladickỳ et al. (2010), 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 Ladickỳ et al. (2010)). 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 Kumar et al. (2011)). We also consider a baseline based on the extended cost-volume filtering

**Table 1** Quantitative comparison on Leuven dataset

| Algorithm | Time (s) | Object (% corr) | Stereo(1) (% corr) | Stereo(2) (% corr) | Stereo(3) (% corr) | Stereo(4) (% corr) | Stereo(5) (% corr) |
|---|---|---|---|---|---|---|---|
| GC+Range(1) (Ladickỳ et al. 2010) | 24.6 | 95.94 | 43.45 | 56.67 | 65.44 | 72.53 | 76.97 |
| GC+Range(2) (Ladickỳ et al. 2010) | 49.9 | 95.94 | 44.12 | 56.98 | 65.84 | 72.97 | 77.31 |
| GC+Range(3) (Ladickỳ et al. 2010) | 74.4 | 95.94 | 44.14 | 57.06 | 65.94 | 73.03 | 77.46 |
| Extended CostVol ((Adams et al. 2010) filter) | 4.2 | 95.20 | 43.53 | 56.44 | 65.51 | 72.86 | 77.26 |
| Dense+HO ((Adams et al. 2010) filter) | 3.1 | 95.24 | 43.58 | 56.18 | 65.89 | 74.08 | 78.89 |
| Dense+HO ((Gastla and Oliveira 2011) filter) | 2.1 | 95.06 | 43.65 | 56.11 | 65.47 | 73.54 | 78.21 |
| Dense+HO+CostVol ((Gastla and Oliveira 2011) filter) | 6.3 | 94.98 | 43.21 | 56.54 | 66.07 | 73.91 | 79.00 |

The table compares the average time per image and performance (Object and Stereo($\delta$) labelling accuracy) of joint object and stereo labelling algorithms. $\delta$ corresponds to the allowed error such that the disparity for $i^{th}$ pixel is considered correct if it satisfies $\|d_i - d_i^g\| \leq \delta$ where $d_i$ and $d_i^g$ are the disparity label for $i^{th}$ pixel and its corresponding ground truth label respectively. We compare following approaches: graph-cut + range-moves (GC+Range($x$), where range moves to disparity values $d \pm x$ are allowed for fixed $d$ at each iteration) Ladickỳ et al. (2010), 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 Ladickỳ et al. (2010), and is an order of magnitude faster. The best stereo accuracies occur when our model is combined with cost-volume filtered unaries for disparity. Here '% corr' corresponds to the total proportion of correctly labelled pixels

**Table 2** Quantitative comparison on Leuven dataset

| Algorithm | Time (s) | Oveall(% corr) | Av. Recall | Av. I/U |
|---|---|---|---|---|
| GC+Range(1) (Ladickỳ et al. 2010) | 24.6 | 95.94 | 72.79 | 68.72 |
| GC+Range(2) (Ladickỳ et al. 2010) | 49.9 | 95.94 | 72.79 | 68.72 |
| GC+Range(3) (Ladickỳ et al. 2010) | 74.4 | 95.94 | 72.79 | 68.72 |
| Extended CostVol ((Adams et al. 2010) filter) | 4.2 | 95.20 | 70.43 | 65.69 |
| Dense+HO ((Adams et al. 2010) filter) | 3.1 | 95.24 | 70.83 | 66.08 |
| Dense+HO ((Gastla and Oliveira 2011) filter) | 2.1 | 95.06 | 70.62 | 65.75 |
| Dense+HO+CostVol ((Gastla and Oliveira (2011)) filter) | 6.3 | 94.98 | 70.60 | 65.63 |

The table compares the average time per image and performance in terms of '% correct', average recall and intersection-union scores for object labelling task of our joint object and stereo labelling 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) Ladickỳ et al. (2010), 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 Ladickỳ et al. (2010), and is an order of magnitude faster. Here '% correct' measure corresponds to the total proportional of correctly labelled pixels, per class recall measure is defined as $\frac{TP}{TP+FN}$ and intersection versus union (I/U) measure is defined as $\frac{TP}{TP+FN+FP}$

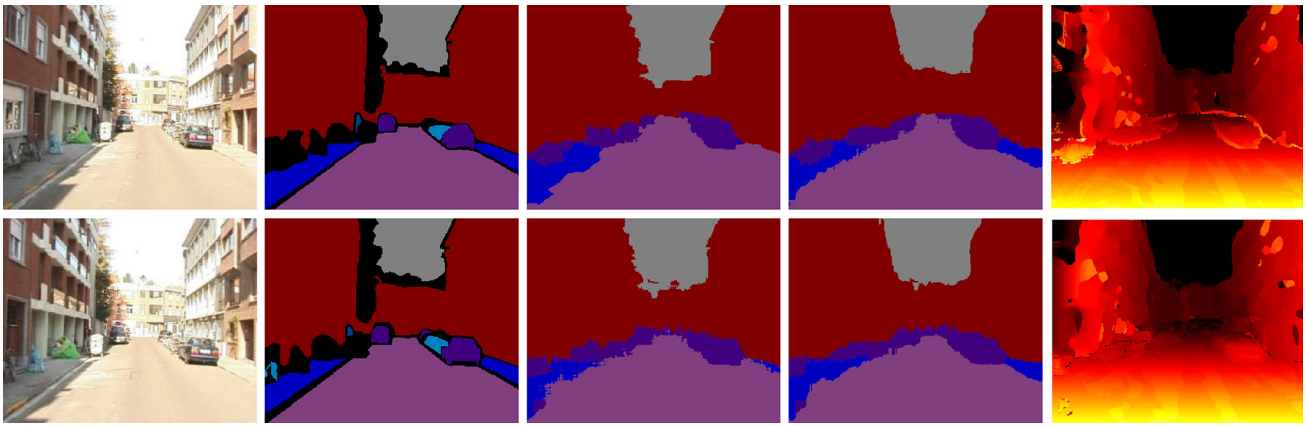

**Fig. 1** Qualitative results on Leuven dataset. From *left* to *right*: input image, ground truth, object labelling from Ladickỳ et al. (2010) (using graph-cut + range-moves for inference), object labelling and stereo outputs from our dense CRF with higher-order terms and extended cost-volume filtering (see text)

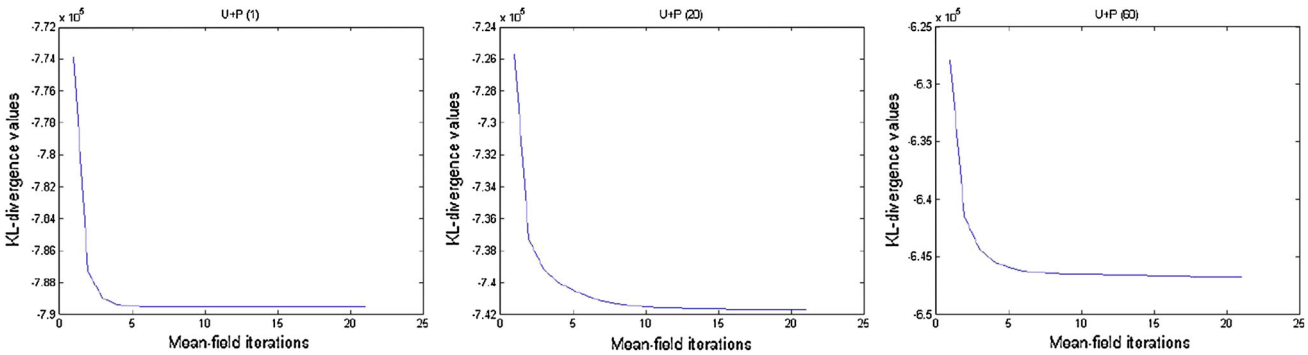


**Fig. 2** Convergence analysis for the joint object-stereo problem: these figures show the KL-divergence values of the mean-field approximation after each iteration when CRF consists of only unary and pairwise terms for the different neighbour-hood sizes. Each column shows the affect of using different neighbour-hood sizes, i.e. when standard spatial deviation is varied from 1 to 20 to 60 pixels. We observe that in practice the KL-divergence values always decrease even though we are using parallel updates

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 $\psi'_u$ as described in Sect. 5. Further, using our basic model we compare two alternative filter-

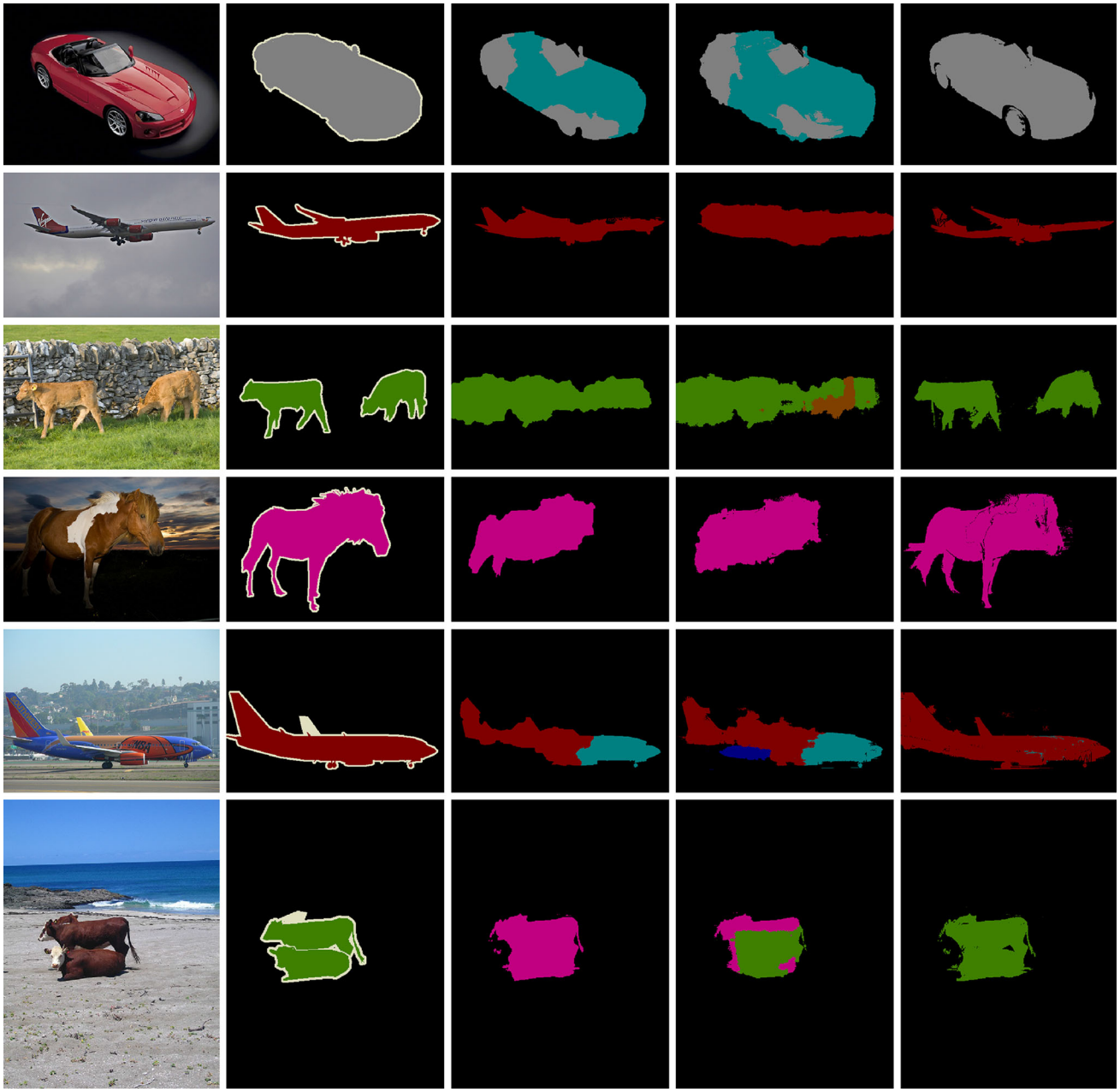

**Fig. 3** Qualitative results on PascalVOC-10 dataset. From *left* to *right*: input image, ground truth, output from Ladickỳ et al. (2010) (AHCRF+Cooccurrence), output from Krahenbuhl and Koltun (2011) (Dense CRF), output from our dense CRF with Potts and Co-occurrence terms

ing methods for inference, the first using the permutohedral lattice, as in Krahenbuhl and Koltun (2011); Adams et al. (2010), and the second using the domain transform based filtering method of Gastla and Oliveira (2011). We evaluate the average time for the joint inference for object and stereo estimation. Further we evaluate the overall percentage of pixels correctly labelled, 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)) over non *void* pixels. For dense stereo reconstruction, we measure the number of pixels satisfying $\|d_i - d_i^g\| \leq \delta$, where $d_i$ is the disparity label for $i^{th}$ pixel, $d_i^g$ is its corresponding ground truth label and $\delta$ is the allowed error. It means a disparity is considered correct if it is within $\delta$ pixels of the ground truth.

In Table 1 we compare the %-correct pixels for object and stereo labelling for different values of the allowed error $\delta$. Further, we also show the average recall and intersec-

**Table 3** Quantitative results on PascalVOC-10

| Algorithm | Time (s) | Overall (%-corr) | Av. Recall | Av. I/U |
|---|---|---|---|---|
| AHCRF+Cooc (Ladický et al. 2010) | 36 | 81.43 | 38.01 | 30.9 |
| DenseCRF ( Krahenbuhl and Koltun 2011) | 0.67 | 80.39 | 35.47 | 28.44 |
| Dense+Potts | 4.35 | 80.13 | 40.49 | 30.27 |
| Dense+Potts+Det | 4.35 | 80.14 | 44.42 | 32.66 |
| Dense+Potts+Cooc | 4.4 | 80.52 | 44.46 | 33.19 |

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 Krahenbuhl and Koltun (2011). Further, our filter-based inference is both able to improve substantially on the inference time and class-average performance of the AHCRF Ladický et al. (2010), with $P^n$-Potts and co-occurrence potentials each giving notable gains. Here '% correct' measure corresponds to the total proportional of correctly labelled pixels, per class recall measure is defined as $\frac{TP}{TP+FN}$ and intersection versus union (I/U) measure is defined as TP/TP+FN+FP

tion/union (I/U) scores for object labelling in Table 2. We note that the densely connected CRF with higher-order terms (Dense+HO) achieves comparable accuracies to Ladický et al. (2010), and that the use of domain transform filtering methods (Gastla and Oliveira 2011) permits an extra speed up, with inference being almost 12 times faster than the least accurate setting of Ladický et al. (2010), 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 labelling 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. Some qualitative results are shown in Fig. 1.

We now highlight the convergence properties of our mean-field algorithm for the joint object-stereo problem. In Fig. 2, we show the KL-divergence values between $Q$ and $P$ distributions after each iteration of our mean-field update under different conditions, specially after varying the neighbourhood size. In practice, we consistently observe that the KL-divergence values always decrease, and in few iterations we reach the local optima even when we vary the density of the CRF.

### 6.3 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 Ladický et al. (2010). We compare both the timing and performance of four algorithms. As our two baselines, we take the AHCRF with a co-occurrence potential (Ladický et al. 2010), whose model includes all higher-order terms but is not densely connected and uses $\alpha$-expansion based inference, and the dense CRF ( Krahenbuhl and Koltun 2011), 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 labelled, 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. 3 and Table 3 respectively. 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 by a factor of 9. Additional per-class quantitative results for object-class segmentation on Pascal-VOC-10 are given in Table 4. We compare the performance of the AHCRF model with co-occurrence potentials of Ladický et al. (2010) with our full model, i.e. a Dense-CRF model with higher-order Potts and co-occurrence potentials, using per-class intersection/union scores. As shown, there is an almost 1.5 %

**Table 4** Per-class Quantitative results on PascalVOC-10 (*bkg* background, *dtb* dining table, *m'bike* motor-bike, *p'son* person, and *av.* average)

| Algorithm | bkg | Plane | Cycle | Bird | Boat | Bottle | Bus | Car | Cat | Chair | Cow |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AHCRF+Cooc (Ladický et al. 2010) | 82.5 | 43.2 | 4.9 | 17.4 | 27.1 | 31.3 | 49.4 | 51.0 | 29.3 | 7.1 | 26.7 |
| Dense+Potts+Cooc | 82.9 | 44.6 | 15.8 | 18.9 | 26.3 | 31.7 | 48.9 | 55.2 | 33.3 | 7.9 | 27.0 |
| | dtb | Dog | Horse | m'bike | p'son | Plant | Sheep | Sofa | Train | TV | av. |
| AHCRF+Cooc (Ladický et al. 2010) | 8.3 | 17.0 | 24.0 | 37.1 | 41.9 | 21.8 | 25.2 | 16.4 | 43.8 | 43.4 | 30.9 |
| Dense+Potts+Cooc | 16.1 | 16.8 | 23.4 | 43.8 | 38.4 | 21.1 | 30.9 | 15.5 | 44.0 | 36.8 | 32.35 |

Shown are the intersection/union scores per class as a %-age defined as $\frac{TP}{TP+FN+FP}$, for Ladický et al. (2010) and our full model

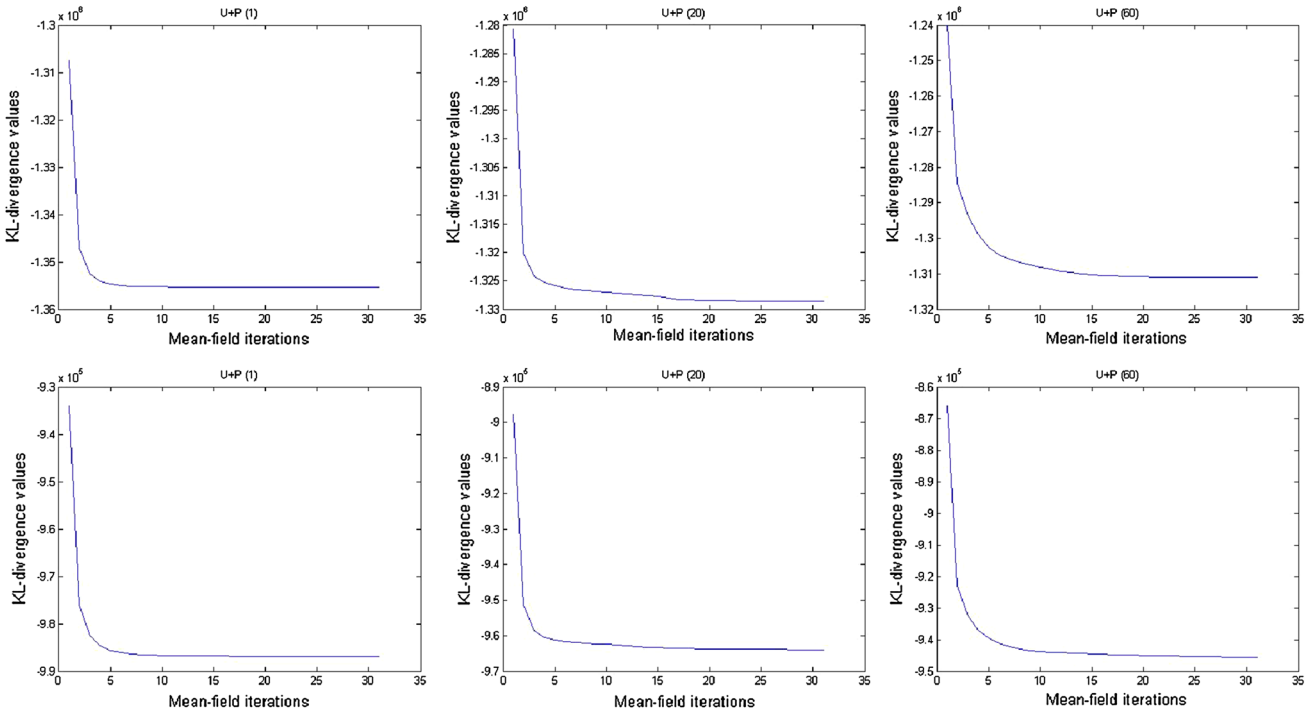


**Fig. 4** Convergence analysis on PascalVOC-10 dataset: these figures show the KL-divergence values of the mean-field approximation after each iteration when CRF consists of only unary and pairwise terms under different decision choices, i.e. varying neighbour-hood sizes (each column) and using noisy unary terms (2nd row). Each column shows the affect of using different neighbour-hood sizes, i.e. when standard spatial deviation is varied from 1 to 20 to 60 pixels. While first row consists of noiseless unary terms, we added some noise to the unary terms in the second row to show the convergence of the mean-field when the unary terms are noisy. We observe that in practice the KL-divergence values always decrease even though we are using parallel updates

improvement in the average score across classes. We do well on some of the difficult classes such as cycle, dinning-table and motor-bike where the relative improvement is almost 6–10 % against Ladickỳ et al. (2010). We also improve on many classes which had high scores like sheep, train, aeroplane, and see a slight dip in certain classes, e.g. boat, person, TV. Since Ladickỳ et al. (2010) includes similar higher-order potentials to ours, the improved performance of our model can be attributed to its dense connectivity and/or our use of mean-field filter-based inference as opposed to graph-cuts (see below Sect. 7).

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 Krahenbuhl and Koltun (2011). 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 Everingham et al. (2011)). We also note that Krahenbuhl and Koltun (2011) 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.

In addition, we also evaluate the convergence of our mean-field algorithm after inclusion of Potts and co-occurrence based higher order terms. We show the KL-divergence values between $Q$ and $P$ distributions after each iteration of our mean-field update under different decision choices, i.e. varying the neighbour-hood sizes and use of noisy unary terms. We first briefly provide analysis for the CRF with only uanry and pairwise terms before going into the CRFs with the higher order terms. In practice, we consistently observe that the KL-divergence values always decrease when the energy functions consist of only unary and pairwise terms even though we are using parallel updates shown in Fig. 4 under all the different decision choices mentioned earlier. However, they can oscillate for some iterations when we include the higher order terms, although we observe a convergence to a local minima overall as shown in Fig. 5. Further, Fig. 6 visually shows the convergence of our mean-field method with higher order terms across iterations, and how the confidence of car pixels increases after inclusion of higher order terms. In all these cases, we observe that the mean-field method reaches very close to the local optima in few iterations.

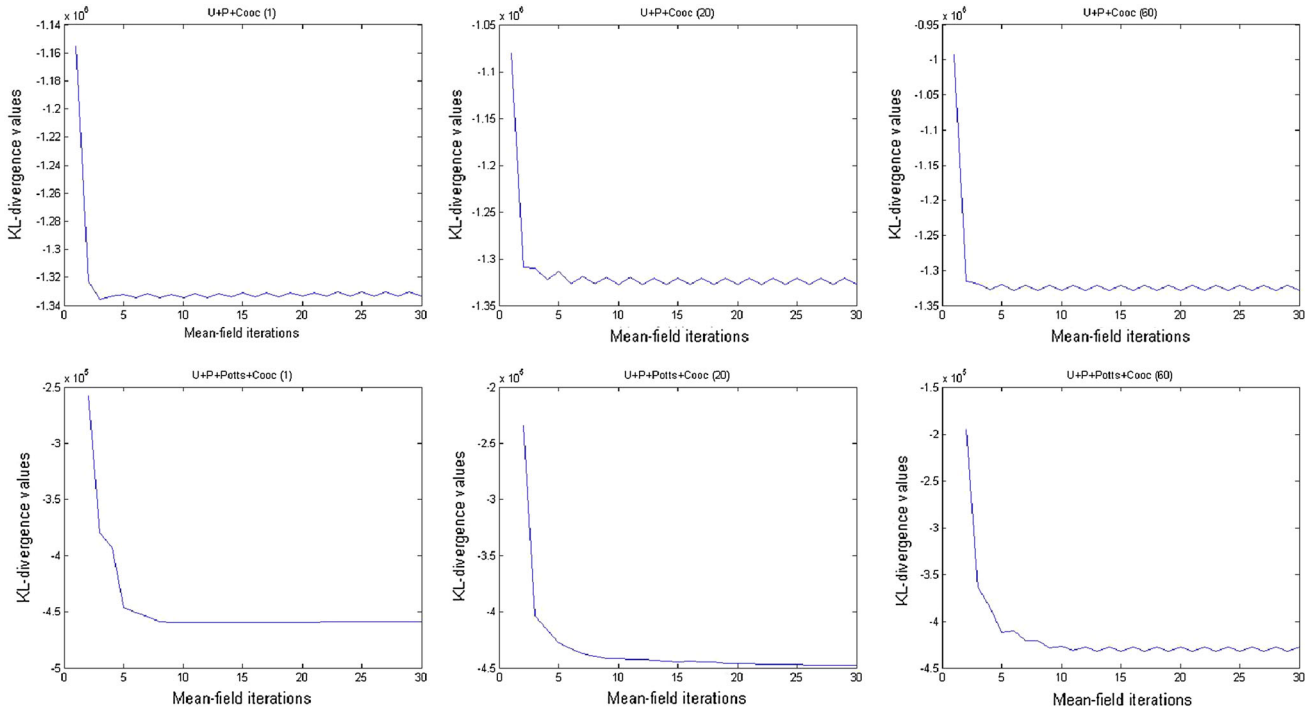


**Fig. 5** Convergence analysis on PascalVOC-10 dataset: these figures show the KL-divergence values after each iteration of the mean-field approximation for two cases. First row shows the affect of varying the density of the CRF when the CRF consists of co-occurrence terms, and the second row shows the affect when the CRF also includes the Potts potentials. Each column shows the affect of using different neighbourhood sizes, i.e. when standard spatial deviation is varied from 1 to 20 to 60 pixels. We observe that in practice the KL-divergence oscillates after inclusion of Potts and co-occurrence potentials

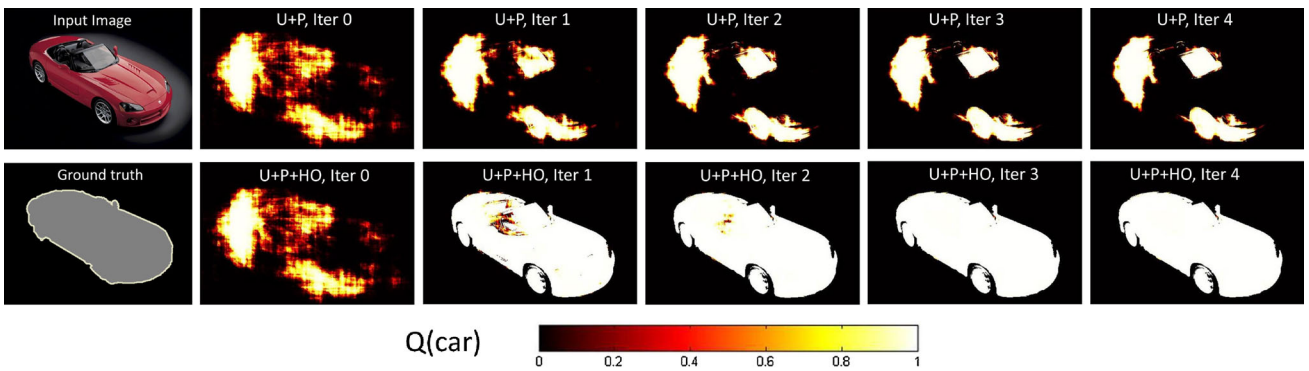


**Fig. 6** It shows the Q distribution values across different iterations of the mean-field method for car class on PascalVOC-10 dataset before (1st row) and after (2nd row) inclusion of higher order terms. We can observe how the confidence of car pixels increases after inclusion of higher order terms

## 7 Mean-Field Analysis

### 7.1 Mean-Field Versus Graph-Cuts Inference

The results shows that the mean-field methods perform equally well or outperform graph-cut methods on all problems we consider. Since the mean-field methods allow us to perform inference in densely connected CRF models, while we restrict attention to models with 8-connected pairwise terms for graph-cuts (with/without higher-order terms in both cases), the question arises as to whether the performance gains are due to the models used or the optimization technique (or both). To investigate this, we rerun our object-class segmentation experiments on PascalVOC-10 using mean-field and graph-cuts ($\alpha$-expansion (Boykov et al. 2001)) inference in CRF models with matching forms of pairwise potential based on Gaussian kernels as in Krahenbuhl and Koltun (2011) (see Sect. 2 of the main paper), using as default stan-

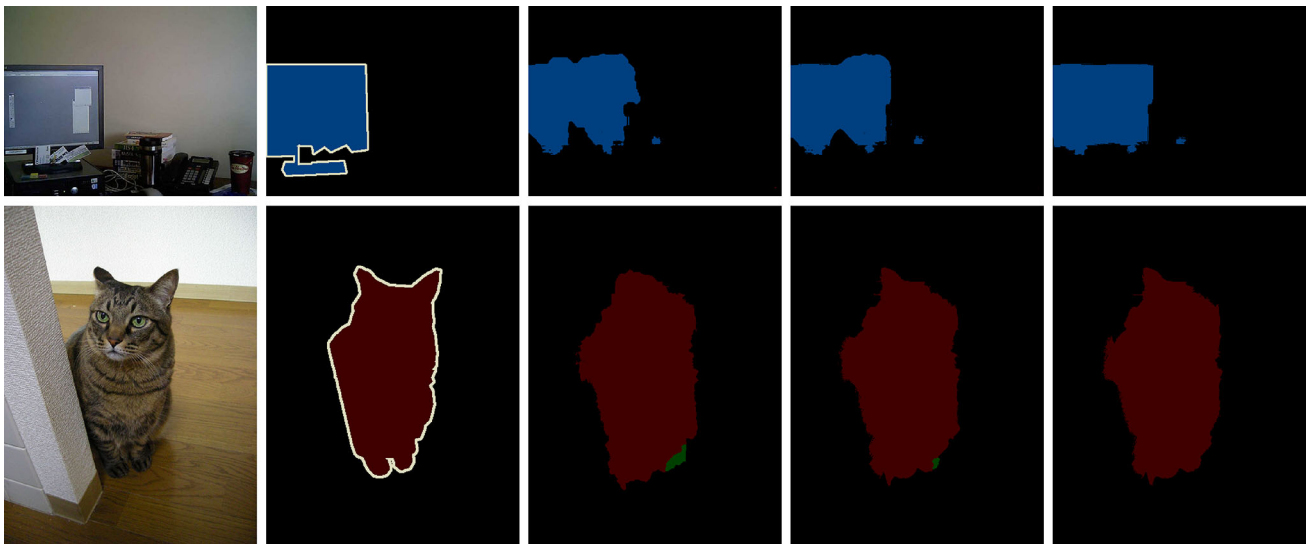

**Fig. 7** Qualitative improvement in $\alpha$-expansion output Boykov et al. (2001) on gradually increasing neighbourhood sizes for each pixel. From *left* to *right*: input image, ground truth, $\alpha$-expansion output with 8, 24 and 48 neighbours respectively

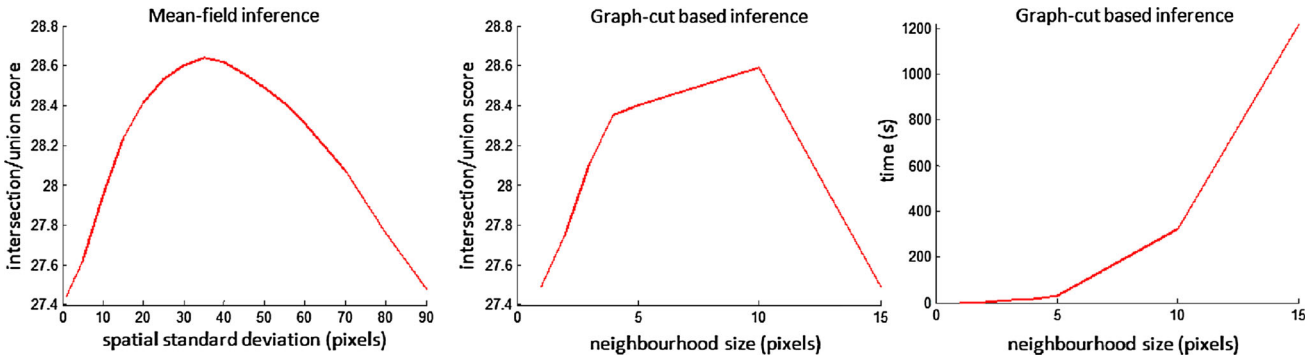


**Fig. 8** Comparison of inference algorithms on PascalVOC-10 using matched energies with pairwise terms only. The *left* plot shows the performance of mean-field inference as the spatial standard deviation of the Gaussian pairwise term is varied. The *centre* plot shows the performance of graph-cut inference ($\alpha$-expansion) as the pairwise neighbourhood size is varied (maintaining a constant spatial standard deviation of 40 pixels). On the *right* are shown the inference times per image associated with the *centre* plot. The inference time for all mean-field settings is $\sim 0.7$ s

dard deviations of 40 and 6 for the spatial and range kernels respectively. Since it is infeasible in terms of time to run $\alpha$-expansion on a fully connected model, we run it on graphs with gradually increased connectivity, where for a neighbourhood size $n$, we have that each pixel is connected to all others whose $x$ and $y$ positions differ from it by no more than $n$ (for $n = 1$ this is 8-connectivity). Some qualitative results on increasing the neighbourhood size for $\alpha$-expansion are shown in Fig. 7. For mean-field inference, we use full connectivity throughout. We compare models with pairwise terms only, pairwise with $P^n$-Potts higher-order potentials, and pairwise with $P^n$-Potts and co-occurrence terms.[5] For graph-cuts inference, we begin with $n = 1$, and test $n = 1, 2, 3, 4, 5, 10, 15$, stopping when the intersection/union score ceases to increase (/does not increase). We also test mean-field inference on pairwise only models with varying kernel standard deviations, for the spatial kernel setting $\sigma_s = 1, 5, 10, 15...70, 80, 90$ pixels, and the range kernel $\sigma_r = 1, 2, 3, 4...15, 18, 20$.

In Fig. 8 we compare performance of the inference methods on the model with pairwise terms only. From the left plot, we see that the best results achieved on the dense model by mean-field occur when the spatial standard deviation is

[5] In fact we use slightly different co-occurrence potentials with graph-cuts and mean-field, since for graph-cuts we use $\psi^{\text{cooc}}$ while for mean-field we use $\psi^{\text{cooc-2}}$, although we set the costs $C(\Lambda)$ identically. We view the latter as an approximation of the former, and thus view this as a slight handicap for mean-field inference; however, further experiments would be needed to determine if the different forms of this potential lead to better/worse models.

**Table 5** Comparison of inference algorithms on PascalVOC-10 using matched energies with pairwise, pairwise and $P^n$-Potts higher-order potentials, and pairwise, $P^n$-Potts and co-occurrence potentials

| Algorithm | Model | Time (s) | Av. I/U |
|---|---|---|---|
| $\alpha$-exp ($n$=10) | Pairwise | 326.17 | 28.59 |
| Mean-field | Pairwise | 0.67 | 28.64 |
| $\alpha$-exp ($n$=3) | Pairwise+Potts | 56.8 | 29.6 |
| Mean-field | Pairwise+Potts | 4.35 | 30.11 |
| $\alpha$-exp ($n$=1) | Pairwise+Potts+Cooc | 103.94 | 30.45 |
| Mean-field | Pairwise+Potts+Cooc | 4.4 | 32.17 |

For the $\alpha$-expansion results, we fix the standard deviation of the Gaussian kernels to the same values as for mean-field (spatial deviation $\sigma_s$=40 pixels, range deviation $\sigma_r$=6), and optimize over the pairwise neighbourhood size n, where n denotes that each pixel is connected to all others with horizontal/vertial offsets of up to $n$ pixels. Shown are intersection versus union (I/U) measure defined as $\frac{TP}{TP+FN+FP}$ averaged across all the classes

around $\sim$ 40 pixels (and corresponding range standard deviation $\sim$ 6). These are the kernel parameters we use with graph-cuts in all models. The central plot shows that graph-cuts is able to achieve approximately the same performance with a neighbourhood connectivity $n = 10$. This seems to indicate that for pairwise only models, increased connectivity leads to improved performance up to a point, and both graph-cuts and mean-field inference are able to achieve similar results in such models in terms of accuracy. However, as shown on the right plot, substantially longer inference times are needed for graph-cuts at the required connectivity to equal the accuracy of mean-field methods (where the inference time remains around $\sim$ 0.7 s for all settings). We highlight here how our approach fits with some of the previous studies (Turner and Sahani 2011; Weiss 2001) which suggest that the inference based on the mean-field approximation provide relatively poor marginal posteriors. Our experimental results suggest that the dense pairwise connection is important to achieve good accuracy with the mean-field approach.

Results in Table 5 compare the performance of both algorithms on models with higher-order terms, and dense connectivity of various neighbourhood sizes for graph-cut inference, where we quote only the setting at which the optimal accuracy is achieved using the protocol described above. The intersection/union scores quoted here are similar to the one in the Table 3 for some settings, but with slight differences caused by the fact that we are ensuring that the potentials in all models take matching forms so that the contributions of model and inference method can be separated. As shown, although both mean-field and graph-cuts inference are able to achieve similar accuracies with dense models using pairwise terms only, when higher-order terms are added the $\alpha$-expansion accuracies are consistently lower than mean-field, even when we allow the former to use models with larger neighbourhood sizes (in fact, for the full model with $P^n$-Potts and co-occurrence terms, nothing is gained by running graph-cuts with neighbourhood sizes of $n > 1$ as shown). These results imply that, unlike the pairwise only case, when such higher-order terms are included not only is mean-field inference faster than graph-cuts, but it is able to optimize these energies substantially better in terms of accuracy than graph-cuts. We thus claim that the performance gains we observe in the experiments of the main paper are due to both the densely connectivity of the models we use, and the mean-field techniques we use to optimize these models.

### 7.2 Sensitivity to Initialization

It is also worth noting that the mean-field inference methods are sensitive to initialization and can thus get stuck in local minima (Weiss 2001). Thus, estimating a good starting point is critical to the mean-field methods. Here, we show how SIFT-flow based label transfer method can be used in providing a good starting point based on the work of Ce Liu et.al. (Liu et al. 2009, 2008). Suppose we have a large training set of annotated ground truth images with per pixel class labels. Given a test image, we first find the K-nearest neighbour images from the training set using GIST features (Oliva and Torralba 2001). In general, we restrict our set to 30 nearest neighbours. We then compute a dense correspondence using the SIFT-flow method from the test image to each of 30 nearest neighbours. We re-rank those nearest neighbours based on the flow values, and pick the best nearest neighbour. Once we have recovered our best candidate, we warp the corresponding ground truth of the candidate image to the current test image. We use these warped labels to initialize the mean-field inference method which acts as a soft constraint on our solutions. We re-weight the unary potential of each pixel based on the label transferred as $\tilde{\psi}_u(x_i) = \lambda * \psi_u(x_i)$, where $\lambda$ is set through cross-validation. We perform experiments with this initialization method on the PascalVOC dataset, and observe both quantitative and qualitative improvement in the accuracy. Figure 9 shows some of query images, their nearest neighbours, and qualitative results before and after SIFT-flow based initialization. Quantitatively, with the better initializations we observe an improvement of almost 2.5 % over the baseline methods with unary and pairwise terms, and almost 0.6 % over the model with unary, pairwise and higher order terms (see Table 6).

## 8 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

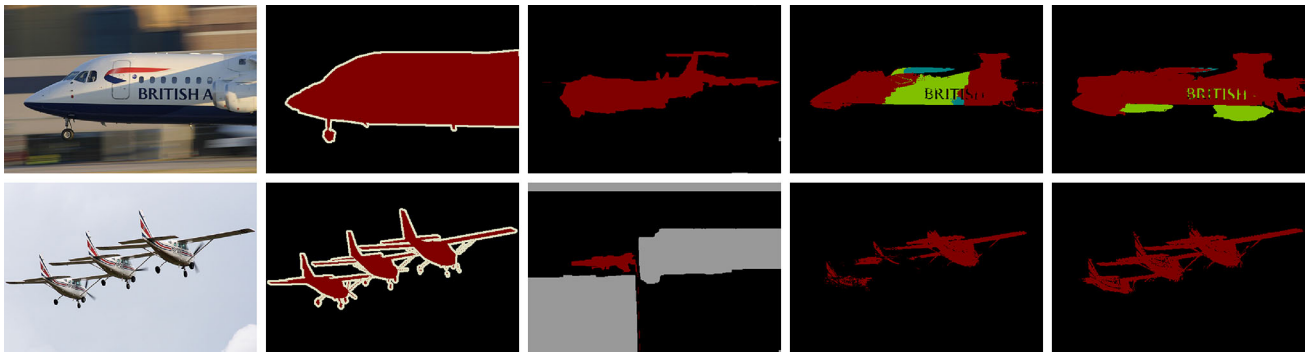

**Fig. 9** Qualitative results on PascalVOC-10 before and after better initialization. From *left* to *right*: input image, ground truth, warped ground truth of the nearest neighbour, output from our dense CRF without better initialisation, and with better initialization

**Table 6** Quantitative results on PascalVOC-10 before and after better initialization

| Algorithm | Time (s) | Overall (%-corr) | Av. Recall | Av. U/I |
|---|---|---|---|---|
| Ours (U+ dense P) | 0.67 | 80.39 | 35.47 | 28.44 |
| Ours (U+ dense P+Init) | 0.9 | 79.65 | 41.84 | 30.95 |
| Ours (U+ dense P+HO) | 4.4 | 80.52 | 44.46 | 33.19 |
| Ours (U+ dense P+HO+Init) | 4.7 | 80.65 | 44.8 | 33.9 |

Though the improvement is significant with unary and pairwise terms, we observe slight improvement in accuracy after inclusion of higher order terms and better initialization compared to the model with higher order terms. Here '% corr' measure corresponds to the total proportional of correctly labelled pixels, per class recall measure is defined as $\frac{TP}{TP+FN}$ and intersection versus union (I/U) measure is defined as $\frac{TP}{TP+FN+FP}$

filter-based methods remain possible for inference in such models, effectively retaining the mean-field update complexity $O(MNL^2)$ as in Krahenbuhl and Koltun (2011) when higher-order $P^n$-Potts and co-occurrence 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 labelling 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. 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/.

**Acknowledgments** We thank Paul Sturgess for his discussion on SIFT-flow based initialization. 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.

## References

Adams, A., Baek, J., & Davis, M. A. (2010). Fast high-dimensional filtering using the permutohedral lattice. *Computer Graphics Forum*, *29*(2), 753–762.

Bai, X. and Sapiro, G. (2007). A geodesic framework for fast interactive image and video segmentation and matting. In *ICCV*.

Bleyer, M., Rhemann , C. and Rother, C. (2012). Extracting 3D scene-consistent object proposals and depth from stereo images. In *ECCV*, (pp. 467–481).

Bleyer, M., Rother, C., Kohli, P., Scharstein, D. and Sinha, S. (2011). Object stereo - joint stereo matching and object segmentation. In *CVPR*, (pp. 3081–3088).

Borestein, E. and Malik, J. (2006). Shape guided object segmentation. In *CVPR*, (pp. 969–976).

Boykov, Y. and Jolly, M. (2001) Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, (pp. 105–112).

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

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach towards feature space analysis. *TPAMI*, *24*, 603–619.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE PAMI*, *24*(5), 603–619.

Criminisi, A. Sharp, T. and Blake, A. (2008). GeoS: Geodesic image segmentation. In *ECCV*, (pp. 99–112).

Everingham, M. Van Gool, L., Williams, C.K.I., Winn, J. and Zisserman, A. (2011). The PASCAL visual object classes, challenge (VOC2011).

Galleguillos, C. Rabinovich, A. and Belongiem, S. (2008). Object categorization using co-occurrence, location and appearance. In *CVPR*.

Gastla, E. S. S. L., & Oliveira, M. M. (2011). Domain transform for edge-aware image and video processing. *ACM Transactions on Graphics*, *30*(4), 69.

Goldlucke, B. and Cremers, D. (2010). Convex relaxation for multilabel problems with product label spaces. In *ECCV*, (pp. 225–238).

Gonfaus, J. M., Boix, X., Van De Weijer, J., Bagdanov, A. D., Serrat, J. and J. (2010). Gonzalez. Harmony potentials for joint classification and segmentation. In *IEEE CVPR*.

Grady, L. (2006). Random walks for image segmentation. *TPAMI*, *28*, 1768–1783.

Kohli, P., Kumar, M.P. and Torr, P.H.S. (2007). P3 & beyond: Solving energies with higher order cliques. In *IEEE CVPR*.
Koller, D., & Friedman, N. (2009). *Probabilistic graphical models*. London: MIT Press.
Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *IEEE PAMI*, *28*(10), 1568–1583.
Komodakis, N. and Paragios, N. (2009). Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *IEEE CVPR*, (pp. 2985–2992).
Komodakis, N., Paragios, N., & Tziritas, G. (2011). MRF energy minimization and beyond via dual decomposition. *IEEE PAMI*, *33*(3), 531–552.
Kornprobst, P., Tumblin, J., & Durand, F. (2009). Bilateral filtering: Theory and applications. *Foundations and Trends in Computer Graphics and Vision*, *4*(1), 1–74.
Krahenbuhl . P. and Koltun, V. (2011). Efficient inference in fully connected CRFs with gaussian edge potentials. In *NIPS*, (pp. 109–117).
Kumar, M., Torr, P. and Zisserman, A. (2005). Obj cut. In *CVPR*, (pp. 18–25).
Kumar, M. P., Veksler, O., & Torr, P. H. S. (2011). Improved moves for truncated convex models. *JMLR*, *12*, 31–67.
Ladický, L., Russell, C., Kohli, P. and Torr, P.H.S. (2009). Assiciative hierarchical CRFs for object class image segmentation. In *ICCV*, (pp. 739–746).
Ladický, L., Russell, C., Kohli, P. and Torr, P.H.S. (2010). Graph cut based inference with co-occurrence statistics. In *ECCV*, (pp. 239–253).
Ladický, L., Sturgess, P., Alahari, K., Russell, C. and Torr, P.H.S. (2010). What, where and how many? combining object detectors and crfs. In *ECCV*.
Ladický, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W.F. and Torr, P.H.S. (2010). Joint optimisation for object class segmentation and dense stereo reconstruction. In *BMVC*, (pp. 1–11).
Lan, X., Roth, S., Huttenlocker, D. and Black, M. (2009). Efficient belief propagation with learnerd higher-order markov random fields. In *ECCV*, (pp. 269–283).
Liu, C., Yuen, J., Torralba, A., Sivic, J. and Freeman, W.T. (2008). SIFT flow: Dense correspondence across different scenes. In *ECCV*.
Liu, C., Yuen, J. and Torralba, A. (2009). Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*.
Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, *42*, 145–175.
Pawan Kumar, M. and Torr, Philip H.S. (2008). Improved moves for truncated convex models. In *NIPS*, (pp. 889–896).
Payet, N. and Todorovic, S. (2010). $(RF)^2$-random forest random field. In *NIPS*.
Potetz, B., & Lee, T. S. (2008). Efficient belief propagation for higher-order cliques using linear constraint nodes. *CVIU*, *112*, 39–54.
Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E. and Belongie, S. (2007). Objects in context. In *ICCV*.
Rhemann, C., Hosni, A., Bleyer, M., Rother, C. and Gelautz. M. (2011). Fast cost-volume filtering for visual correspondence and beyond. In *CVPR*, (pp. 3017–3024).
Rother, C., Kohli, P., Feng, W. and Jia, J. (2009). Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, (pp. 1382–1389).
Rother, C., Kohli, P., Feng, W. and Jia, J. (2009). Minimizing sparse higher order energy functions of discrete variables. In *CVPR* .
Rother, C., Kolmogorov, V., & Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM TOG*, *23*, 309–314.
Shotton, J., Winn, J. M., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, *81*(1), 2–23.
Singaraju, D., Grady, L. and Vidal R. (2008). P-Brush: Continuous valued MRFs with normed pairwise distributions for image segmentation. In *CVPR*.
Torralba, A., Murphy, K. P., & Freeman, W. T. (2007). Sharing visual features for multiclass and multiview object detection. *IEEE PAMI*, *29*, 854–869.
Toyoda, T., & Hasegawa, O. (2008). Random field model for integration of local information and global information. *TPAMI*, *30*, 1483–1489.
Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In *Bayesian time series models*, (pp. 109–130).
Veksler, O. (2007). Graph cut based optimization for MRFs with truncated convex priors. In *CVPR* .
Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in MRFs. *Advanced mean field methods: Theory and practices*. Cambridge, MA: MIT Press.
Woodford, O., Torr, P. H. S., Reid, I., & Fitzgibbon, A. (2009). Global stereo reconstruction under second-order smoothness priors. *IEEE PAMI*, *31*(12), 2115–2128.