Improved Initialisation and Gaussian Mixture Pairwise Terms for Dense Random Fields with Mean-field Inference

Vibhav Vineet
vibhav.vineet-2010@brookes.ac.uk
Jonathan Warrell
jwarrell@brookes.ac.uk
Paul Sturgess
paul.sturgess@brookes.ac.uk
Philip H.S. Torr
philiptorr@brookes.ac.uk

Department of Computing
Oxford Brookes University
Oxford, UK
cms.brookes.ac.uk/research/visiongroup/

Abstract

Recently, Krahenbuhl and Koltun proposed an efficient inference method for densely connected pairwise random fields using the mean-field approximation for a Conditional Random Field (CRF). However, they restrict their pairwise weights to take the form of a weighted combination of Gaussian kernels where each Gaussian component is allowed to take only zero mean, and can only be rescaled by a single value for each label pair. Further, their method is sensitive to initialization. In this paper, we propose methods to alleviate these issues. First, we propose a hierarchical mean-field approach where labelling from the coarser level is propagated to the finer level for better initialisation. Further, we use SIFT-flow based label transfer to provide a good initial condition at the coarsest level. Second, we allow our approach to take general Gaussian pairwise weights, where we learn the mean, the co-variance matrix, and the mixing co-efficient for every mixture component. We propose a variation of *Expectation Maximization (EM)* for piecewise learning of the parameters of the mixture model determined by the maximum likelihood function. Finally, we demonstrate the efficiency and accuracy offered by our method for object class segmentation problems on two challenging datasets: PascalVOC-10 segmentation and CamVid datasets. We show that we are able to achieve state of the art performance on the CamVid dataset, and an almost 3% improvement on the PascalVOC-10 dataset compared to baseline graph-cut and mean-field methods, while also reducing the inference time by almost a factor of 3 compared to graph-cuts based methods.

1 Introduction

Labelling problems in computer vision are often modelled as discrete optimisation problems. Examples include object class segmentation, stereo correspondence, image de-noising etc [9, 17]. Generally, these problems are solved in a Markov Random Field (MRF) or Conditional Random Field (CRF) framework, where the basic model includes pairwise terms defined over a grid with 4 or 8 neighbours. A more expressive model is to allow dense connectivity

http://dx.doi.org/10.5244/C.26.73

which captures long range interactions between variables. However the complexity increases with more interactions.

Recently, Krahenbuhl and Koltun [8] proposed an efficient bilateral filtering based method for inference in dense pairwise CRFs [1], where pairwise weights take the form of a weighted combination of Gaussian kernels. Given this approach, they formulate multilabelling problems as performing approximate maximum posterior marginal (MPM) inference in a mean-field approximation to the CRF [7]. Within this framework, empirically they achieve a significant speed-up compared to graph-cuts based methods [9], and observe improvements in the accuracy on object-class segmentation problems. However, in its current form, their method is associated with two limitations.

The first issue is related to the fact that the mean-field approximation assumes complete factorisation over the individual variables [7]. Though this simplified model leads to efficient and tractable models for learning and inference [6], the mean-field inference methods are generally sensitive to initialisation [13]. Over the years, many different variants of mean-field inference methods have been developed to solve the issue of initialisation [5, 7, 13]. In this work, we propose a hierarchical mean-field approach to improve the quality of the solutions by providing good initial conditions. We perform mean-field inference at the coarser level, and transfer the labels from the coarser level to the finer level for initialisation. At the coarser level, we use a SIFT-flow [10] based label transfer method for better initialisation. SIFT-flow provides an elegant algorithm for finding the correspondence between two images, where images are taken from different view points but share similar high level scene characteristics. Given a query image, we use this strategy to find the nearest neighbour from the training set, and warp the corresponding labelled ground truth image to the current query image. We use this warped image to initialize the current labelling for the mean-field inference method at the coarser level. Further, assuming the hierarchical mean-field proposes a good hypothesis for initialisation, we re-weight the unary potentials based on the initialisations.

The second issue relates to the form of the pairwise weights in [8] which are a linear combination of Gaussian kernels. Although they learn the standard deviation and weighting co-efficient of each component, they allow each Gaussian component to take only zero mean. Further, they use the same combination of Gaussian kernels for each label pair, though they scale this by a learnt label compatibility function. In this paper, we propose an approach that extends this model. Specifically, we allow our model to take a more general Gaussian mixture model for every pair of labels, where we learn the mean, the co-variance matrix and the mixing co-efficient of each Gaussian component. We propose a piecewise learning framework where given a set of data points, we fit Gaussian mixture model to those points. We use a variation of the *Expectation Maximization (EM)* algorithm for estimating the parameters determined by maximum likelihood. It should be noted though that we do not learn the label compatibility function as in [8].

In summary, our main contributions are:

- Proposing a hierarchical mean-field approach for transferring labels from the coarser level to the finer level for better initialisation,
- An adaptation of the SIFT-flow based label transfer method to further improve initialisation by finding a semantically close nearest neighbours,
- An *EM* based algorithm for piecewise learning of a general Gaussian mixture model for the pairwise terms, increasing the expressivity of the filter-based mean-field inference.

We evaluate the accuracy and efficiency proposed by our method on object class segmentation problem using the PascalVOC-10 segmentation [4] and CamVid datasets [3]. We compare our results with the filter-based dense pairwise CRF method [8], and graph cuts based α-expansion [9, 15] which does not incorporate dense pairwise connections, and validate the significance of our methods. We achieve state of the art results on the CamVid dataset, and observe an improvement of almost 3% on PascalVOC compared to baseline graph-cut and mean-field methods, while also reducing the inference time by almost a factor of 3 compared to graph-cuts based methods.

The rest of the paper is structured as follows. In section 2, we outline the bilater filter-based efficient inference for dense pairwise CRF by Krahenbuhl and Koltun [8]. Section 3.1 provides the details of our hierarchical mean-field approach, and Sec. 3.2 gives the details of learning and inference with general Gaussian pairwise terms. The experimental evaluations are presented in Sec. 4.

2 Mean-field Inference in Dense Random Fields

We formulate the multilabel problem in a conditional random field (CRF) framework where each random variable corresponds to a pixel in the image. Let $\mathbf{X} = \{X_1, X_2, ..., X_N\}$ denote the set of random variables corresponding to the image pixels $i \in \{1, 2, ..., N\}$. Each random variable takes a label from the label set $L = \{l_1, l_2, ..., l_k\}$. A labelling $\mathbf{x}$ refers to any possible assignment of labels to the random variables and takes values from the set $\mathbf{L} = L^N$.

We define a fully connected pairwise CRF where each variable is connected to all other variables. Given this framework, the probability distribution $\mathrm{P}(\mathbf{x}|\mathbf{I})$ over the labellings of the CRF can be written as:

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

where $E(\mathbf{x}|\mathbf{I})$ is the energy function corresponding to the current configuration, $\mathrm{Z}(\mathbf{I})$ is the normalizing constant, and $\mathbf{I}$ is the given image data. For fully connected pairwise CRFs, the energy function takes the form $E(\mathbf{x}|\mathbf{I}) = \sum_i \psi_i(x_i) + \sum_{i<j} \psi_{ij}(x_i, x_j)$. The unary potentials $\psi_i(x_i)$ are based on local feature responses and can take arbitrary form, but [8] restrict the pairwise potentials to take the form of a linear combination of Gaussian kernels:

$$\psi_{ij}(x_i, x_j) = \kappa(x_i, x_j) \sum_{v=1}^{V} w^{(v)} k^{(v)}(\mathbf{f}_i, \mathbf{f}_j) \tag{1}$$

where $\kappa(.,.)$ is an arbitrary *label compatibility function*, while $k^{(v)}(.,.)$, $v = 1...V$ are Gaussian kernels defined on feature vectors $\mathbf{f}_i$, $\mathbf{f}_j$ derived from the image data at locations i and j (where [8] 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 $w^{(v)}$, $v = 1...V$ are used to weight the kernels.

Given this form of energy function, Krahenbuhl and Koltun [8] proposed a filter-based method for performing fast inference in the mean-field approximation to the CRF. The mean-field is a very simple distribution which assumes complete factorisation of the probability distribution as: $Q(X) = \prod_i Q_i(x_i)$. The mean-field inference algorithm tries to minimize the KL-divergence $\mathbf{D}(\mathrm{Q}||\mathrm{P})$ between the approximate distribution Q, and the true distribution P. By considering the conditions that must be satisfied at the minima, 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 [7] for a derivation):

$$Q_i(x_i = l) = \frac{1}{Z_i} \exp\{-\psi_i(x_i) - \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j(x_j = l') \psi_{ij}(x_i, x_j)\} \tag{2}$$

where $Z_i = \sum_{x_i=l\in\mathcal{L}} \exp\{-\psi_i(x_i) - \sum_{l'\in\mathcal{L}} \sum_{j\neq i} Q_j(x_j = l')\psi_{ij}(x_i,x_j)\}$ is a constant which normalizes the marginal at pixel i. [8] show the expensive update equation in the mean-field is approximated by a convolution with a bilateral filter in a high dimensional space as follows:

$$\tilde{Q}_i^{(v)}(l) \quad = \quad \sum_{j\neq i} k^{(v)}(\mathbf{f}_i,\mathbf{f}_j)Q_j(l) = [\mathcal{G}_v \otimes Q(l)](\mathbf{f}_i) - Q_i(l) \tag{3}$$

where $\mathcal{G}_v$ is a Gaussian kernel, and $\tilde{Q}_i^{(v)}(l)$ can be used to approximate the pairwise terms in Eq. 2. Given this Gaussian convolution, they use a permutohedral lattice based bilateral filtering method [1] for performing efficient inference. They run the update equation for a fixed number of iterations, where each iteration leads to decrease in the KL-divergence value. To extract a solution, they evaluate the approximate maximum posterior marginal as $x_i = \max_{x_i} Q_i(x_i)$.

3 Methods

In this section, we describe our proposed algorithms for improving the robustness of the filter-based mean-field inference method [8], and thus improving the quality of the solutions offered by the method. Specifically, we focus on two aspects: providing a good initial condition, and improving the quality of the pairwise weights, details of which are provided in the Sec. 3.1, and Sec. 3.2 respectively.

3.1 Initialisation with hierarchical mean-field approximation

The mean-field approximation assumes complete factorisation of the probability distribution, and is thus far from properly approximating the true marginal distribution. Consequently, one problem with mean-field inference is that it is too easy to get stuck in local minima resulting in sensitivity to initialisation [13]. To validate this, we conduct experiments on object class segmentation on the PascalVOC-10 segmentation dataset, and we observe such behaviour with mean-field inference. If we initialize our starting labelling with the maximum unary potential responses, the mean-field approximation [8] achieves 28.52%, where the accuracy is intersection/union (I/U) score measured per class (defined in terms of the true/false positives/negatives for a given class as TP/(TP+FP+FN)). If we initialize the solution with the ground truth labelling, the mean-field results improved by almost 13% compared to the previous results. Thus, estimating a good starting point is critical to the mean-field inference methods. We outline our approach based on the hierarchical mean-field and SIFT-flow approaches. The benefits are two fold: we use it to initialize the mean-field method as well using the labels to re-rank the unary potentials.

Many problems in computer vision have been modelled by a coarse to fine hierarchy [6]. In this work, we investigate such a coarse to fine hierarchy for the mean-field approaximation. Here we restrict ourselves to using only two layers. But, the approach can be generalized to any number of layers. We define a variable at a coarser level to correspond to four variables at the next finer level. We apply mean-field inference at the coarse level, and use the solution to initialize the mean-field inference at the finer level by assigning the same label to all four variables corresponding to one at the coarse level. Further, we use a SIFT-flow [11] based label transfer method to initialize the coarse level.

Ce Liu et.al. [10, 11] propose the SIFT-flow method for higher level image alignment, where images are taken from different view points but share similar higher-level scene characteristics. Correspondences between images are established based on SIFT features [12]

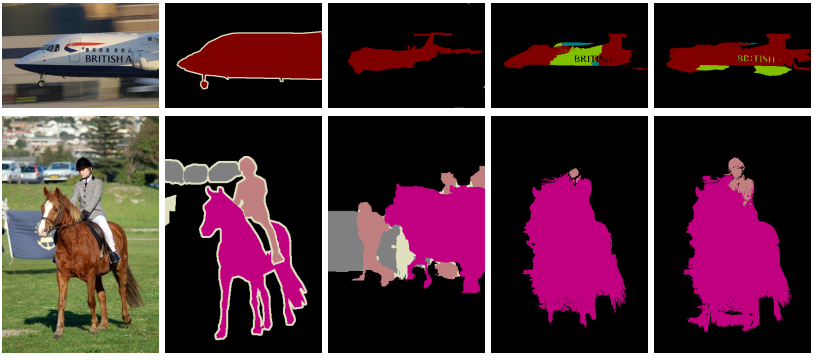

Figure 1: Qualitative results of SIFT-flow method on PascalVOC-10 dataset. From left to right: input test image, test image ground truth, ground truth of nearest training image, output of mean-field [8] without and with SIFT-flow initialisation respectively.

in an energy minimization framework. They apply it in many problem domains including video retrieval, scene alignment, face recognition, and object recognition tasks.

In this work, we use a similar strategy for label transfer. These transferred labels provide good initial conditions, and act as a soft constraint on our solutions. Suppose we have a large training set of annotated ground truth images with per pixel class labels. Now, given a test image, we first find the K-nearest neighbour images using GIST features [13]. 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. Figure 1 shows some of query images, nearest neighbours, and the reranked images based on the SIFT-flow values.

3.2 General Gaussian mixture pairwise terms

Krahenbuhl and Koltun [8] use the following pairwise energy function:

$$E(\mathrm{X}|\mathrm{I}) = \sum_i \psi(x_i) + \sum_{i<j} \kappa(x_i, x_j) \sum_{v=1}^{V} w^{(v)} k^{(v)}(\mathbf{f}_i, \mathbf{f}_j) \tag{4}$$

where $\kappa(x_i, x_j)$ is the label compatibility function between pairs of labels, and $k^{(v)}(\mathbf{f}_i, \mathbf{f}_j)$ is the v^{th} Gaussian kernel with zero mean and an arbitrary standard deviation. In this work, we propose a method to alleviate this restrictive assumption by incorporating a more general class of Gaussian mixture function to Eq. 4. Let our mixture function $\mathcal{G}^{ij}_{(\mathrm{mix})}(\mathrm{I})$ for the i^{th} and j^{th} pair of labels take the following form:

$$\mathcal{G}^{ij}_{(\mathrm{mix})}(\mathrm{I}) = \sum_{\mathrm{m}=1}^{\mathrm{M}} \alpha^{ij}_m \mathcal{G}_{\mathrm{m}}(\mathrm{I}, \mu_{\mathrm{m}}, \Sigma_{\mathrm{m}}) \tag{5}$$

where α^{ij}_m, μ_{m}, and Σ_{m} are the mixing co-efficients, mean, and co-variance matrix of m^{th} Gaussian mixture component $\mathcal{G}_{\mathrm{m}}$ corresponding to the $(ij)^{th}$ label pair, and I is an image derived feature. Further, we assume the mixing co-efficients α^{ij} to come from a probability

distribution. On incorporating this Gaussian mixture function into Eq. 4, our final more general pairwise energy function takes the following form:

$$E(\mathrm{X}|\mathrm{I}) = \sum_i \psi(x_i) + \sum_{i<j} \kappa(x_i, x_j) \sum_{v=1}^{V} w^{(v)} k^{(v)}(\mathbf{f}_i, \mathbf{f}_j) - \lambda \sum_{i<j} \sum_{m=1}^{M} \alpha_m^{(x_i, x_j)} \mathcal{G}_m(\mathrm{I}, \mu_m, \Sigma_m) \quad (6)$$

Here λ is a weight which combines the contributions from two different sets of Gaussian kernels, zero-mean kernels $K^{(v)}(.,.)$ and our general learnt kernels $\mathcal{G}^{ij}_{(\mathrm{mix})}$. In principal, we do not need a separate set of zero-mean Gaussian kernels, since they can be absorbed into our general mixture model. However, we found it useful to treat these separately for parameter setting. We now explain our learning method for the mixing co-efficients $\alpha^{(\cdots)}$, the mean μ_m, and the co-variance matrix Σ_m.

Learning mixture models: Given this CRF model, we follow the piecewise strategy of [14] for learning the parameters of the CRF. They show how the piecewise method provides an efficient and accurate alternative to joint learning of the parameters. Thus, first we set the parameters of unary $\psi(x_i)$, and first pairwise weights $\sum_{i<j} \kappa(x_i, x_j) \sum_{v=1}^{V} w^{(v)} k^{(v)}(\mathbf{f}_i, \mathbf{f}_j)$ following the works of [14] and [8] respectively. We then set the parameters α, μ, Σ using the method described below, and the value of λ is set through cross validation.

Suppose we have a set of data points $\mathbf{f'_1}, \mathbf{f'_2}, ..., \mathbf{f'_n}$ where each feature vector $\mathbf{f'_i}$ is derived from the image data at two locations $\mathbf{f'_i} = \mathbf{f}_i - \mathbf{f}_j$, and their ground truth labels are l_i and l_j. Let us represent our data by an $N \times D$ matrix $\mathbf{F'}$, where the rows correspond to D dimensional feature points $\mathbf{f'_i}$. Assuming the data points are drawn in an i.i.d. fashion, we define a log-likelihood function as follows:

$$\log P(\mathbf{F'}|\alpha, \mu, \Sigma) = \sum_{\mathbf{i}=1}^{N} \log\{ \sum_{m=1}^{M} \alpha_m^{(l_i l_j)} \mathcal{G}_m(\mathbf{f'_i}|\alpha_m, \Sigma_m)\} \quad (7)$$

Given this setting, we propose an *Expectation Maximization (EM)* method for learning parameters of the mixture models in maximum likelihood framework. During the M step, to satisfy the conditions at the maximum of the function, we first take partial derivatives of the function with respect to each parameter, and we set them to zero. First, we derive the conditions for μ_m by setting the derivative of $\log P(\mathbf{F'}|\alpha, \mu, \Sigma)$ w.r.t. μ_m to zero as follows:

$$\frac{\partial \log P(\mathbf{F'}|\alpha, \mu, \Sigma)}{\partial \mu_m} = -\sum_{\mathbf{i}=1}^{N} \frac{\alpha_m^{l_i l_j} \mathcal{G}(\mathbf{f'_i}|\mu_m, \Sigma_m)}{\sum_{m'} \alpha_{m'}^{l_i l_j} \mathcal{G}(\mathbf{f'_i}|\mu_{m'}, \Sigma_{m'})} \sum_m (\mathbf{f'_i} - \mu_m) = -\sum_{\mathbf{i}=1}^{N} \gamma_{\mathbf{i}m} (\mathbf{f'_i} - \mu_m) = 0 \quad (8)$$

On rearranging this, we get the update equation for μ_m as $\mu_m = \frac{1}{N_m} \sum_{\mathbf{i}=1}^{N} \gamma_{\mathbf{i}m} \mathbf{f'_i}$. Following similar strategy, we get the following update equations for Σ_m, and $\alpha^{l_i l_j}$:

$$\Sigma_m = \frac{1}{N_m} \sum_{\mathbf{i}=1}^{N} \gamma_{\mathbf{i}m} (\mathbf{f'_i} - \mu_m)(\mathbf{f'_i} - \mu_m)^{\mathrm{T}}; \alpha_m^{l_i l_j} = \frac{N_m^{l_i l_j}}{N^{l_i l_j}} \quad (9)$$

where $N_m^{l_1 l_2} = \sum_{\mathbf{i}} \gamma_{\mathbf{i}m}^{l_i l_j} [l_i = l_1 \wedge l_j = l_2]$, $N^{l_1 l_2} = \sum_{\mathbf{i}} [l_i = l_1 \wedge l_j = l_2]$, and $N_m = \sum_{\mathbf{i}} \gamma_{\mathbf{i}m}$. During the M step we assume that the value of $\gamma_{\mathbf{i}m}$ is constant. Then, during E step we evaluate the value of $\gamma_{\mathbf{i}m} = \frac{\alpha_m^{l_i l_j} \mathcal{G}(\mathbf{f'_i}|\mu_m, \Sigma_m)}{\sum_{m'} \alpha_{m'}^{l_i l_j} \mathcal{G}(\mathbf{f'_i}|\mu_{m'}, \Sigma_{m'})}$ assuming the Gaussian parameters μ_m, Σ_m and α_m are constant. Details of whole iterative procedure are given in the Algorithm 1.

Algorithm 1: *EM* based learning Gaussian mixture model

input : Initialize $\alpha^{l_i l_j}$, μ_m, Σ_m

converged := 0, v := 1;

while *converged* $= 0$ **do**

E Step: evaluate $\gamma_{\mathbf{i}m} = \frac{\alpha_m^{l_i l_j} \mathcal{G}(\mathbf{f}'_\mathbf{i}|\mu_m,\Sigma_m)}{\sum_{m'} \alpha_{m'}^{l_i l_j} \mathcal{G}(\mathbf{f}'_\mathbf{i}|\mu_{m'},\Sigma_{m'})}$;

M Step: re-estimate parameters: $\alpha^{l_i l_j}$, μ_m, Σ_m as follows: ;

$\mu_m = \frac{1}{N_m}\sum_{\mathbf{i}=1}^{N} \gamma_{\mathbf{i}m}\mathbf{f}'_\mathbf{i}$, $\Sigma_m = \frac{1}{N_m}\sum_{\mathbf{i}=1}^{N} \gamma_{\mathbf{i}m}(\mathbf{f}'_\mathbf{i}-\mu_m)(\mathbf{f}'_\mathbf{i}-\mu_m)^{\mathrm{T}}$;

$\alpha_m^{l_i l_j} = \frac{N_m^{l_i l_j}}{N^{l_i l_j}}$, $N_m^{l_1 l_2} = \sum_\mathbf{i} \gamma_{\mathbf{i}m}^{l_i l_j}[l_i = l_1 \wedge l_j = l_2]$;

$N^{l_1 l_2} = N^{l_1 l_2} = \sum_\mathbf{i}[l_i = l_1 \wedge l_j = l_2], N_m = \sum_\mathbf{i} \gamma_{\mathbf{i}m}$;

Evaluate the log likelihood $\log P(\mathbf{F}'|\alpha,\mu,\Sigma)$;

end

Return $\alpha^{l_i l_j}, \mu_m, \Sigma_m$;

Our piecewise learning strategy does not guarantee any bound on the solution achieved even though [16] bound the solution achieved in their piecewise learning framework. The first reason is that the parameter λ is not learnt jointly in the CRF. Further, we use a generative model to learn the parameters of the mixture components which given a pair of labels models the distribution of feature vectors, and use the negative likelihood within the energy. This is in contrast to [14] who maximize the conditional likelihood of the labels given the training data and use the negative conditional log-likelihood as an energy term.

Inference with mixture model: Now, we explain our approach for efficient inference using the mixture model. Each mixture component involves evaluating an extra expensive term: $\sum_{i<j}\sum_{m=1}^{M} \alpha_m^{(x_i x_j)} \mathcal{G}_m(\mathbf{I},\mu_m,\Sigma_m)$. We formulate this expensive step as an efficient Gaussian filtering operation in high dimensional space following the work of Krahenbuhl and Koltun [8]. Thus, our filtering step under a non-zero mean is given by:

$$\tilde{Q}_i^m(l) = \sum_{j\neq i} \mathcal{G}_m(\mathbf{f}_i - \mathbf{f}_j|\mu_m,\Sigma_m) = [\mathcal{G}_m \otimes Q(l)](\mathbf{f}_i - \mu_m) - \mathcal{G}_m(0)Q_{(i)}(l) \tag{10}$$

We use the permutohedral lattice based filtering method [1] for fast filtering. We first embed the feature points in the high dimensional space translating the points by the means μ_m and project them onto the lattice points. We apply blurring on the mean-shifted feature points.

4 Experiments

We demonstrate the accuracy and efficiency offered by our approach on object-class segmentation problems on two challenging datasets: Cambridge-driving Labelled Video Database (CamVid) [3], and PascalVOC-10 segmentation dataset [2]. 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. We compare our method with two baselines, the dense CRF [8], which uses filter-based inference, and graph cuts based alpha-expansion which is not densely connected. We use the permutohedral lattice [1] for filtering in all models. In all our models, we have only unary and pairwise connections. We evaluate the efficacy of learning the Gaussian mixture components for the pairwise terms in the potts setting for the object-class segmentation problem on PascalVOC dataset. We learn a model with $m = L$ Gaussian components, using data from label pairs $l_i = l_j$ only. This generates $L \times L$ mixing

Algorithm	Time (s)	Overall (%-corr)	Av. Recall	Av. U/I
α-exp (U+P) [2]	0.96	78.84	58.64	43.89
APST (U+P+H)[15]	1.6	85.18	60.06	50.62
dense CRF (U + dense P) [8]	0.2	79.96	59.29	45.18
Ours (U + dense P + hierar)	0.35	85.31	59.75	50.56

Table 1: Quantitative results on CamVid. The table compares the timing and performance of our approach (last line) against three baselines. The importance of better initialisation is confirmed by the fact that we achieve state of the art results. Futher, our model with just unary and pairwise terms is also able to slightly improve results compared to the model of [15] which along with unary and pairwise terms uses segment based higher order terms.

Algorithm	Time (s)	Overall (%-corr)	Av. Recall	Av. U/I
α-exp (U+P) [2]	3.0	79.52	36.08	27.88
AHCRF (U+P+H) + Cooc [9]	36	81.43	38.01	30.9
dense CRF (U + dense P) [8]	0.67	71.63	34.53	28.4
Ours1 (U + dense P+GM)	26.7	80.23	36.41	28.73
Ours2 (U+ dense P+hierar)	0.90	79.65	41.84	30.95
Ours3 (U+ dense P+hierar+GM)	26.7	78.96	44.05	31.48

Table 2: Quantitative results on PascalVOC-10. The table compares the timing and performance of our approach (last three lines) against three baselines. The importance of better initialisation and Gaussian mixtures is confirmed by the significant improvement achieved compared to the other methods, which use only unary and pairwise connections, and slight improvement in the results compared to the model of [9] which uses segment based higher order terms, detector potentials, and co-occurrence terms as well.

co-efficients $\alpha^{ij} \in [0\ 1]$, thus allowing each $l_i = l_j$ label pair to reweight the L Gaussian components. Further, we note our overall timings do not include the timings for SIFT-flow, and for embedding the feature points in the permutohedral lattice. We assess the overall percentage of pixels correctly labelled, the average recall per class, and the intersection/union (I/U) measure per class.

CamVid dataset: We test our model on the CamVid training and test set. We use the same split as used by [15] who randomly partition 600 images into 367 training images, and 233 test images, and the same number of 11 object classes. Table 1 quantitatively compares our methods with other methods. We observe an overall improvement of 6.5% compared to graph-cuts based α-expansion and 5.5% compared to the dense CRF method [8]. In all these cases, we used only unary and pairwise connections. Further, our model with unary and pairwise connections performs better than [15] who use unary, pairwise and higher order terms by almost 0.2%. We observe a qualitative improvement, importantly we are able to recover sign-poles, and building parts missing from the output of other methods, as shown in the Fig. 2. Further, we also note our approach is able to reduce the inference time by a factor of 3 to 5 compared to graph cuts based α-expansion and the work of [15].

PascalVOC dataset: We also test our model on the PascalVOC-10 training and validation set. We use the same split as used in [8], 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 [8], and an Ising label compatibility function $\mu(l_1, l_2) = [l_1 \neq l_2]$.

Qualitative and quantitative results are shown in Fig. 2 and Tab. 2 respectively. Our approaches are able to outperform both of the baseline methods in terms of union-intersection

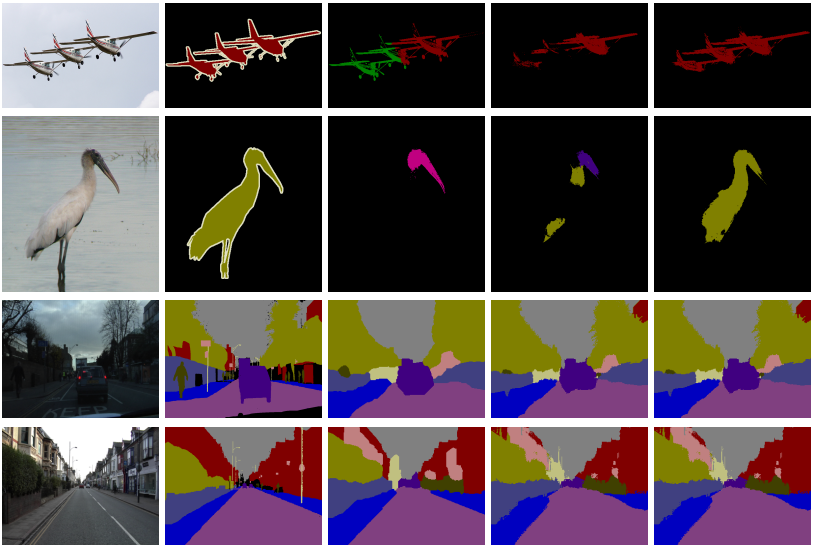

Figure 2: Qualitative results on PascalVOC-10 (first 2 rows) and CamVid (last 2 rows) datasets. From left to right: input image, ground truth, output from [9] (U+P), output from [8] (dense CRF), output from our dense CRF with better initialisation and Gaussian mixture.

(U/I) metrics, demonstrating the importance of the Gaussian mixture components, and hierarchical SIFT-flow based initialisations. As shown, we observe an improvement of almost 1% in union-intersection (U/I) score compared to graph-cuts based α-expansion and 0.3% compared to the dense CRF of [8] on inclusion of the Gaussian mixture components. Further, using our second model with better initialisations, we are able to improve the U/I accuracy by almost 3% and 2.5% compared to the baseline methods. In our final model which includes the mixture components and the better initialisation strategy, we observe an improvement of 3.5% and 3% over the baseline methods. Further, although our final model only includes unary and pairwise terms, we observe 0.5% improvement in U/I score and almost 6% improvement in the average recall scores over the work of [9] who include higher order terms, detector potentials, and object co-occurrence terms along with unary and pairwise potentials. Apart from the improved accuracy offered by our approaches, we also observe an improvement in the inference timing compared to the graph-cuts based baseline methods. Our model with better initialisation achieves a speed up of 3 times compared to the α-expansion method, and a speed up of 40 times compared to the work of [9], although our method with the Gaussian mixture components is slower as we have to evaluate the filtering step separately for each of the mixture components in the model. Finally, 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 union/intersection compared to the highest performing method on the 2011 challenge, see [4]). We also note that [8] are able to further improve their average union/intersection score to 30.2% by learning the pairwise label compatibility function, which remains a possibility for our model also.

5 Conclusion

In this paper, we propose two methods to improve the accuracy and efficiency offered by the filter-based mean-field inference method [8]. Specifically, we focus on two aspects of

the mean-field inference: providing a good initial condition, and improving the mean-field inference by using more general pairwise weights. We propose a hierarchical mean-field inference method where we perform inference at the coarser level, and propagate the solution from the coarser level to the finer level. We use SIFT-flow to initialise the mean-field inference at the coarser level. Further, we propose a piecewise framework for learning the parameters of the mixture of Gaussian pairwise weights using *Expectation Maximization (EM)* algorithm where parameters are determined by maximum likelihood functions. We extensively evaluate our method on object class segmentation on two challenging datasets, PascalVOC-10 segmentation and CamVid dataset, and observe a significant improvement in accuracy and efficiency over the other existing approaches.

Acknowledgement: The work was supported by the EPSRC and the IST programme of the European Community, under the PASCAL2 Network of Excellence. Prof. Philip H.S. Torr is in receipt of Royal Society Wolfson Research Merit Award.

References

[1] Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. *Comput. Graph. Forum*, 2010.

[2] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[3] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2): 88–97, 2009.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

[5] Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233, 1999.

[6] Taesup Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang D. Yoo. Variable grouping for energy minimization. In *CVPR*, pages 1913–1920, 2011.

[7] Daphne Koller and Nir Friedman. Probabilistic graphical models. In *MIT Press*, 2009.

[8] Philipp Krahenbuhl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[9] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. *ECCV*, 2010.

[10] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T. Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV (3)*, pages 28–42, 2008.

[11] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, pages 1972–1979, 2009.

[12] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3): 145–175, 2001.

[14] Jamie Shotton, John M. Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81 (1):2–23, 2009.

[15] Paul Sturgess, Karteek Alahari, Lubor Ladicky, and Philip H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009.

[16] Charles A. Sutton and Andrew McCallum. Piecewise training for undirected models. In *UAI*, pages 568–575, 2005.

[17] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *PAMI*, 30(6), 2008.

[18] Yair Weiss. Comparing the mean field method and belief propagation for approximate inference in mrfs. *Advanced Mean Field Methods: Theory and Practices*, 2001.