Improved Initialisation and Gaussian Mixture Pairwise Terms for Dense Rnadom 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/

Many labelling problems in computer vision are often modelled as discrete optimisation problems such as object class segmentation, stereo correspondence, image de-noising etc. 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 which captures long range interactions between variables. However the complexity increases with more interactions.

Recently, Krahenbuhl and Koltun [1] proposed an efficient bilateral filtering based method for inference in dense pairwise CRFs, where pairwise weights take the form of a weighted combination of Gaussian kernels. They formulate multilabelling problems as performing approximate maximum posterior marginal (MPM) inference in a mean-field approximation to the CRF. Empirically they achieve a significant speed-up compared to graph-cuts based methods, 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. Though this simplified model leads to efficient and tractable models for learning and inference, the mean-field inference methods are generally sensitive to initialisation. 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 better initialisation. At the coarser level, we use a SIFT-flow [3] based label transfer method for better initialisation. SIFT-flow provides an algorithm for finding the correspondence between two images, where images are taken from different view points but share similar high level scene characteristics. 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. 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. These transferred labels provide good initial conditions, and act as a soft constraint on our solutions.

The second issue relates to the form of the pairwise weights in [1] 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. 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. Assuming there are M Gaussian mixture components, our energy function takes the following form:

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

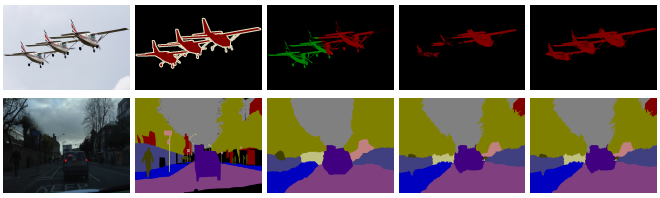

Figure 1: Qualitative results on PascalVOC-10 (first row) and CamVid (last row) datasets. From left to right: input image, ground truth, output from [2] (U+P), output from [1] (dense CRF), output from our dense CRF with better initialisation and Gaussian mixture.

where $\kappa(x_i,x_j)$ is the label compatibility function between pair of labels, λ is a weighting function, μ_m, and Σ_m are the mean, and co-variance matrix of mth mixture component, α_m^{ij} is the m^{th} mixing coefficient between the i^{th} and j^{th} labels, and I is an image derived feature. Given the number of mixture components, we learn the mixing co-efficient $\alpha^{(\cdot,\cdot)}$, the mean μ_m, and the co-variance matrix Σ_m. 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.

1 Experiments

We demonstrate the accuracy and efficiency offered by our approach on object-class segmentation problems on CamVid and PascalVOC-10 segmentation dataset. We assess the average union/intersection measure per class (defined in terms of the true/false positives/negatives for a given class as TP/(TP+FP+FN)). On CamVid dataset, we observe an overall improvement of 6.5% compared to graph-cuts based α-expansion and 5.5% compared to dense CRF method [1]. Here we use only unary and pairwise connections. Further, our model with unary and pairwise connections perform better than [4] who use unary, pairwise and higher order terms by almost 0.2%. On PascalVOC-10 segmentation dataset, with our final model which includes the mixture components and the better initialisation strategy, we observe an improvement of 3.5% and 3% over the α-expansion method. Further, in our model which 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 [2] who include higher order terms, detector potentials, and object co-occurrence terms along with unary and pairwise potentials. We also observe a qualitative improvement in the results on both CamVid and PascalVOC dataset, some of the images are shown in the Fig. 1.

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

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

[3] 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.

[4] 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.