Consensus Message Passing for Layered Graphical Models
Supplementary Material

Varun Jampani[†] **S. M. Ali Eslami**[†]**, Daniel Tarlow, Pushmeet Kohli and John Winn**
MPI for Intelligent Systems, Tübingen Microsoft Research, Cambridge

1 Random regression forests for CMP

We wish to learn a mapping f from contextual messages $\mathbf{c}$ to the consensus message m from training data $\{(\mathbf{c}_d, m_d)\}_{d=1\ldots D}$. This is challenging since the inputs and outputs of the regression problem are both messages (*i.e.* distributions), and special care needs to be taken to account for this fact. We follow closely the methodology of Eslami et al. (2014), who use random forests to predict outgoing messages from a factor given the incoming messages to it. For a review of forests see (Criminisi and Shotton, 2013).

In approximate message passing (*e.g.* EP; Minka, 2001 and VMP; Winn and Bishop, 2005), messages can be represented using only a few numbers, *e.g.* a Gaussian message can be represented by its natural parameters. We represent the contextual messages $\mathbf{c}$ collectively, in two different ways: the first is a concatenation of the parameters of its constituent messages which we call the 'regression parameterization' and denote by $\mathbf{r}_\mathrm{c}$; and the second is a vector of features computed on the set which we call the 'tree parameterization' and denote by $\mathbf{t}_\mathrm{c}$. This parametrization typically contains features of the set as a whole (*e.g.* moments of their means). We represent the outgoing message m by a vector of real valued numbers $\mathbf{r}_\mathrm{m}$.

Prediction model. Each leaf node is associated with a subset of the labelled training data. During testing, a previously unseen set of contextual messages represented by $\mathbf{t}_\mathrm{c}$ traverses the tree until it reaches a leaf which by construction is likely to contain similar training examples. We therefore use the statistics of the data gathered in that leaf to predict the consensus message with a multivariate regression model of the form: $\mathbf{r}_\mathrm{m} = \mathbf{W} \cdot \mathbf{r}_\mathrm{c} + \epsilon$ where ϵ is a vector of normal error terms. We use the learned matrix of coefficients $\mathbf{W}$ at test time to make predictions $\overline{\mathbf{r}}_\mathrm{m}$ for each $\mathbf{r}_\mathrm{c}$. To recap, $\mathbf{t}_\mathrm{c}$ is used to traverse the contextual messages down to leaves, and $\mathbf{r}_\mathrm{c}$ is used by a linear regressor to predict the parameters $\mathbf{r}_\mathrm{m}$ of the consensus message.

Training objective function. The optimization of the split functions proceeds in a greedy manner. At each node j, depending on the subset of the incoming training set $\mathcal{S}_j$ we learn the function that 'best' splits $\mathcal{S}_j$ into the training sets corresponding to each child, $\mathcal{S}_j^\mathrm{L}$ and $\mathcal{S}_j^\mathrm{R}$, *i.e.* the parameters of the split criterion $\boldsymbol{\tau}_j = \mathrm{argmax}_{\boldsymbol{\tau}\in\mathcal{T}_j} I(\mathcal{S}_j, \boldsymbol{\tau})$. This optimization is performed as a search over a discrete set $\mathcal{T}_j$ of a random sample of possible parameter settings. The objective function I is:

$$I(\mathcal{S}_j, \boldsymbol{\tau}) = -E(\mathcal{S}_j^\mathrm{L}, \mathbf{W}^\mathrm{L}) - E(\mathcal{S}_j^\mathrm{R}, \mathbf{W}^\mathrm{R}), \tag{1}$$

where $\mathbf{W}^\mathrm{L}$ and $\mathbf{W}^\mathrm{R}$ are the parameters of the regression models corresponding to the left and right training sets $\mathcal{S}_j^\mathrm{L}$ and $\mathcal{S}_j^\mathrm{R}$, and E is the 'fit residual' as defined in (Eslami et al., 2014). In simple terms, this objective function splits the training data at each node in a way that the relationship between the incoming and outgoing messages is well captured by the regression in each child.

Ensemble model. During testing, a set of contextual messages simultaneously traverses every tree in the forest from their roots until it reaches their leaves. Combining the predictions into a single forest prediction might be done by averaging the parameters $\overline{\mathbf{r}}_\mathrm{m}^t$ of the predicted messages $\overline{m}^t$ by each tree t, however this would be sensitive to the chosen parameterization. Instead we compute the moment average $\overline{m}$ of the distributions $\{\overline{m}^t\}$ by averaging the first few moments of the predictions across trees, and solving for the distribution parameters which match the averaged moments (see *e.g.* Grosse et al., 2013).

[†]The first two authors contribute equally to this work.

2 Results on the face problem

2.1 Qualitative results

Figure 1 shows inference results for reflectance maps, normal maps and lights for randomly chosen test images, and Figure 2 shows reflectance estimation results on multiple images of the same subject produced under different illumination conditions. Consensus message passing is able to produce reflectance estimates that are closer to the photometric stereo groundtruth across subjects and across different illumination conditions.

2.2 Quantitative results

Figure 3 shows quantitative results for both real images from 'Yale B' and 'Extended Yale B' datasets (Georghiades et al., 2001; Lee et al., 2005) and synthetic shadowless images. The synthetic shadowless images were created using the same light, reflectance and normal map statistics as that of images in the real dataset (however estimated using photometric stereo (Quéau et al., 2013)). Subject recognition results indicate superior performance of CMP in comparison to other baselines in both real and synthetic image settings.

Figure 4 shows the quantitative results of light inference using the different inference techniques. We use the cosine angle distance between the estimated light and the photometric stereo groundtruth (error $= \cos^{-1}(\hat{\mathbf{l}}_{\text{est}}.\hat{\mathbf{l}}_{\text{ps}})$) as an error metric. Here, $\hat{\mathbf{l}}_{\text{est}}$ is a unit vector in the same direction as the mean of the posterior light estimate of CMP and $\hat{\mathbf{l}}_{\text{ps}}$ is a unit vector in the same direction as the corresponding photometric stereo groundtruth. Again, these results indicate the superior performance of CMP in comparison to other baselines in both real and synthetic image settings.

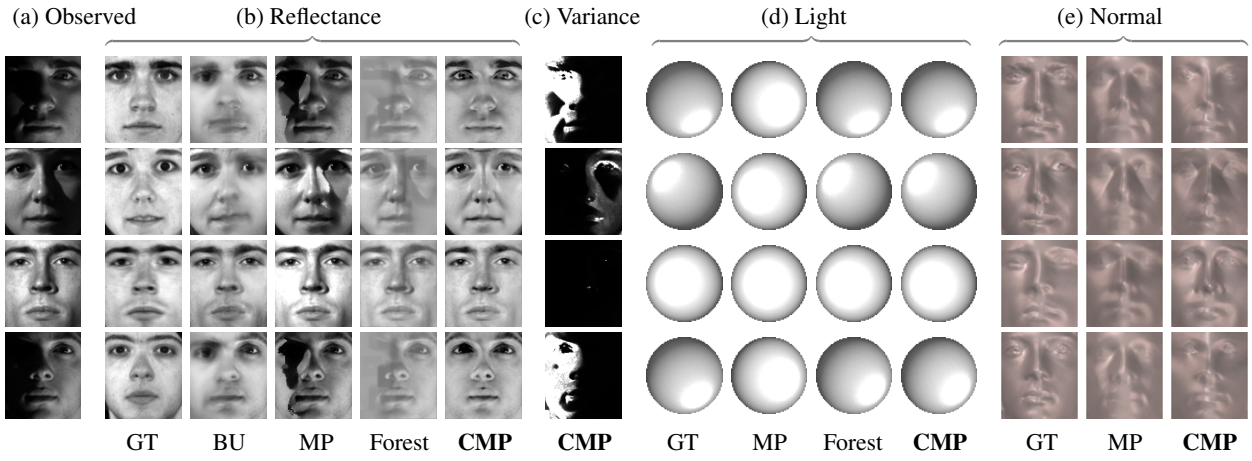


Figure 1: **A visual comparison of inference results.** For 4 randomly chosen test images, we show inference results obtained by competing methods. (a) Observed images. (b) Inferred reflectance maps. *GT* is the photometric stereo groundtruth, *BU* is the Biswas *et al.* (2009) reflectance estimate and *Forest* is the consensus prediction. (c) The variance of the inferred reflectance estimate produced by CMP (normalized across rows). High variance regions correlate strongly with cast shadows. (d) Visualization of inferred light directions. (e) Inferred normal maps.

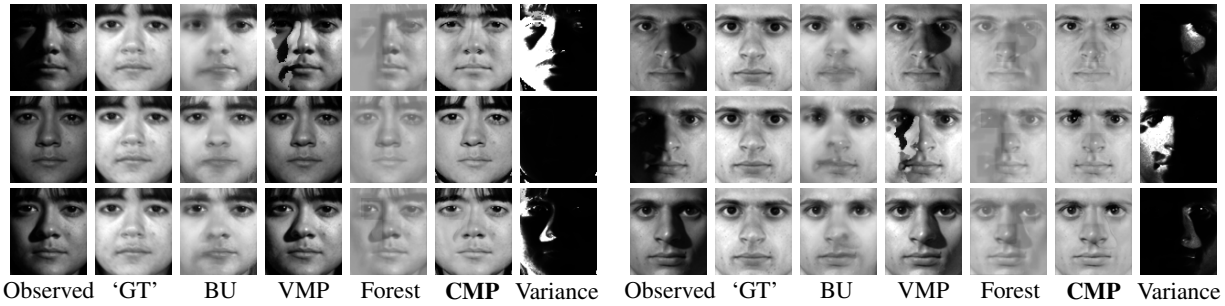
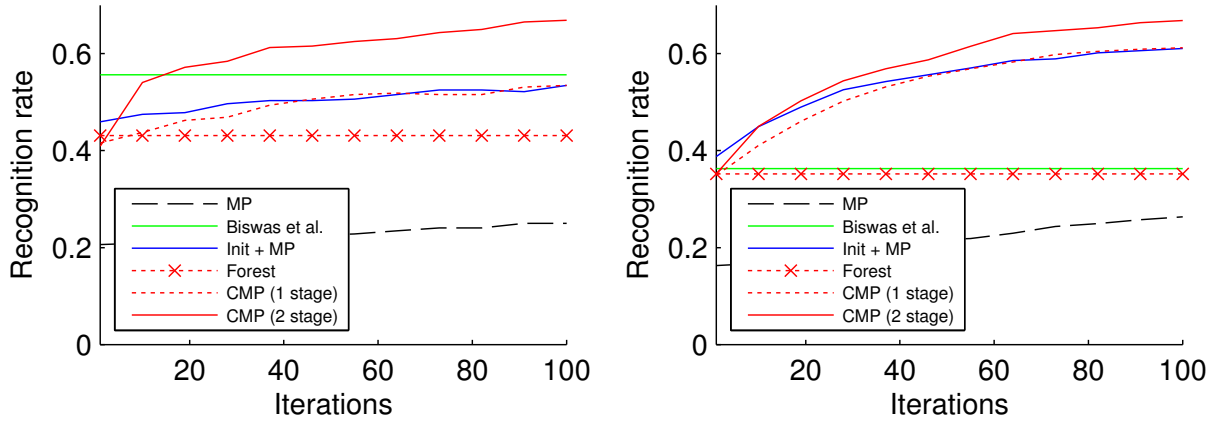
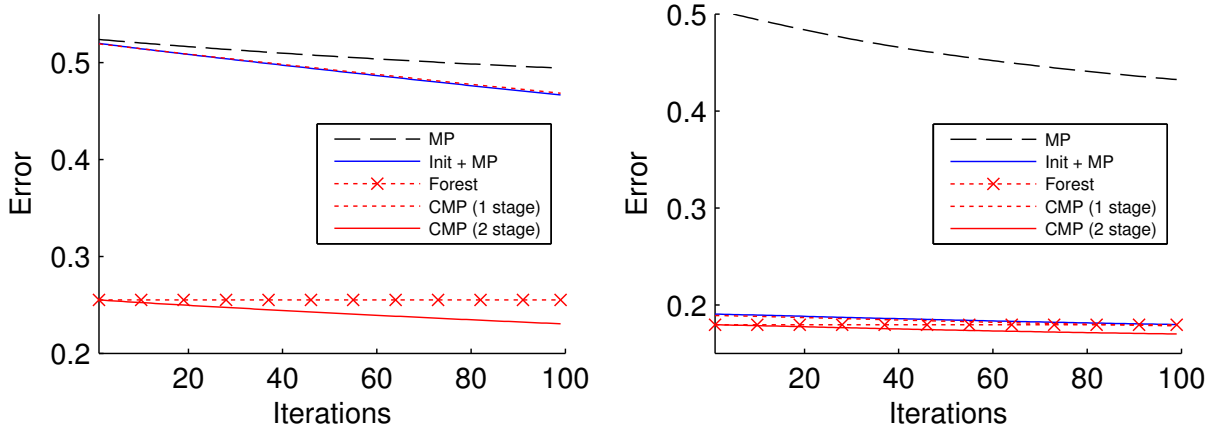


Figure 2: **Robustness to varying illumination.** Reflectance estimation on two subject images with varying illumination. Left to right: observed image, photometric stereo estimate which is used as a proxy for groundtruth, bottom-up estimate of Biswas et al. (2009), VMP result, consensus forest estimate, CMP mean, and CMP variance.



(a) Synthetic (b) Real

Figure 3: **Reflectance inference accuracy.** Results have been averaged over all images of test subjects. (a) Synthetic, shadowless images. (b) Real images.



(a) Synthetic (b) Real

Figure 4: **Light inference accuracy.** Results have been averaged over all images of test subjects. (a) Synthetic, shadowless images. (b) Real images.

References

Biswas, S., Aggarwal, G., and Chellappa, R. (2009). Robust estimation of albedo for illumination-invariant matching and shape recovery. *Pattern Analysis and Machine Intelligence*, 31(5):884–899.

Criminisi, A. and Shotton, J. (2013). *Decision Forests for Computer Vision and Medical Image Analysis*. Springer Publishing Company, Incorporated.

Eslami, S. M. A., Tarlow, D., Kohli, P., and Winn, J. (2014). Just-In-Time Learning for Fast and Flexible Inference. In *Neural Information Processing Systems (NIPS) 27*, pages 154–162.

Georghiades, A., Belhumeur, P., and Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence*, 23(6).

Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. (2013). Annealing between distributions by averaging moments. In *Neural Information Processing Systems (NIPS) 26*, pages 2769–2777.

Lee, K.-C., Ho, J., and Kriegman, D. (2005). Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence*, 27(5):684–698.

Minka, T. (2001). *Expectation Propagation for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology.

Quéau, Y., Lauze, F., and Durou, J.-D. (2013). *Solving the uncalibrated photometric stereo problem using total variation*. Springer.

Winn, J. and Bishop, C. M. (2005). Variational Message Passing. *Journal of Machine Learning Research*, 6:661–694.