A Learned Joint Depth and Intensity Prior using Markov Random Fields
Supplemental Material

Daniel Herrera C.[1], Juho Kannala[1], Peter Sturm[2], and Janne Heikkilä[1]

[1]Center for Machine Vision Research
University of Oulu
{dherrera, jkannala, jth}@ee.oulu.fi
[2]INRIA Grenoble, Rhône-Alpes
peter.sturm@inria.fr

1. Derivations

Schmidt [2] presents an in-depth analysis of the Field-of-Experts model for a single channel image using Gaussian Scale Mixtures as experts. Many parts of the analysis apply to our extension to intensity and depth images. Here we show the derivation of the equations that need to be extended to account for the two channels. For reference, we expand here Eq. (1) from our main paper using GSM as expert functions

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{-\epsilon||\mathbf{x}||^2/2} \prod_{k=1}^{K}\prod_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij} \cdot \mathcal{N}(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; 0, \sigma_i^2/s_j). \tag{1}$$

1.1. Learning

To train the model using contrastive divergence we need the derivatives of the model's likelihood with respect to the model parameters. As in Section 5 of [2], it is advantageous to look at the FoE model density in terms of its energy

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(-E(\mathbf{x};\boldsymbol{\theta})), \tag{2}$$

$$E(\mathbf{x};\boldsymbol{\theta}) = \frac{\epsilon}{2}||\mathbf{x}||^2 - \sum_{k=1}^{K}\sum_{i=1}^{N} \log \phi(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; \alpha_i). \tag{3}$$

To ensure that the learned weights for the GSM are always positive and sum to one, we perform the following variable replacement

$$\alpha_{ij} = \frac{\omega_{ij}}{\sum_{j=1}^{J} \omega_{ij}} \quad \text{where} \quad \omega_{ij} = \exp(\hat{\alpha}_{ij}) \tag{4}$$

and we learn $\hat{\alpha}_{ij}$ instead of the original α_{ij}. The GSM experts can be written concisely as vector products

$$\phi(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; \hat{\alpha_i}) = \frac{1}{\sum_{j=1}^{J} w_{ij}} \sum_{j=1}^{J} w_{ij} \cdot \mathcal{N}(\mathbf{h}_i^\top \mathbf{u}(k) + \mathbf{g}_i^\top \mathbf{v}_{(k)}; 0, \sigma_i^2/s_j) \tag{5}$$

$$= (\boldsymbol{\omega}_i^\top \mathbf{1})^{-1} \boldsymbol{\omega}_i^\top \boldsymbol{\varphi}(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}), \tag{6}$$

where $\boldsymbol{\omega}_i = [\omega_{i1}, \dots \omega_{iJ}]^\top$, $\mathbf{1}$ denotes the J-dimensional 1-vector, and $\boldsymbol{\varphi}(x) = \{\mathcal{N}(x; 0, \sigma_i^2/s_j) | j = 1, \dots, J\}$ is a vector-valued function where we denote vectors of element-wise derivatives with $\boldsymbol{\varphi}'(x)$,$\boldsymbol{\varphi}''(x)$, *etc.*

The log-likelihood of the model is defined over the training data $\mathbf{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(T)}\}$, where each $\mathbf{x}^{(t)}$ is an i.i.d. image with an intensity and a depth channel, *i.e.* $\mathbf{x}^\top = [\mathbf{u}^\top \mathbf{v}^\top]$. The log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \log \prod_{t=1}^{T} p(\mathbf{x}^{(t)}; \boldsymbol{\theta}) = \sum_{t=1}^{T} \log p(\mathbf{x}^{(t)}; \boldsymbol{\theta}) = \sum_{t=1}^{T} -\log Z(\boldsymbol{\theta}) - E(\mathbf{x}^{(t)}, \boldsymbol{\theta}). \tag{7}$$

As shown in [2], the derivative w.r.t the model parameters $\boldsymbol{\theta} = \{\mathbf{h}_i, \mathbf{g}_i, \hat{\boldsymbol{\alpha}}_i | i = 1, \dots, N\}$ required for contrastive divergence are

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -T \frac{\partial \log Z(\boldsymbol{\theta})}{\boldsymbol{\theta}} - \sum_{t=1}^{T} \frac{\partial E(\mathbf{x}^{(t)}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{8}$$

$$= T \left[\left\langle \frac{\partial E(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle_p - \left\langle \frac{\partial E(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle_{\mathbf{X}} \right] \tag{9}$$

where $\langle \cdot \rangle_{\mathbf{X}}$ and $\langle \cdot \rangle_p$ are the expected values over the training data set and a hallucinated data set drawn from distribution p. We rely on sampling to draw these samples and approximate the expected derivative, as in [2].

The derivative of the energy w.r.t. the GSM parameters is

$$\frac{\partial E(\mathbf{x}; \boldsymbol{\theta})}{\partial \hat{\alpha}_{ij}} = -\sum_{k=1}^{K} \frac{\partial \log \phi(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; \hat{\alpha}_{ij})}{\partial \hat{\alpha}_{ij}} \tag{10}$$

$$= \frac{\omega_{ij}}{\boldsymbol{\omega}_i^T \mathbf{1}} \left[K - \sum_{k=1}^{K} \frac{\mathcal{N}(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; 0, \sigma_i^2/s_j)}{\phi(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; \hat{\alpha}_{ij})} \right]. \tag{11}$$

The derivative w.r.t. to the intensity filter coefficients is

$$\frac{\partial E(\mathbf{x}; \boldsymbol{\theta})}{\partial h_{im}} = -\sum_{k=1}^{K} \frac{\partial \log \phi(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; \hat{\alpha}_{ij})}{\partial h_{im}} \tag{12}$$

$$= -\sum_{k=1}^{K} \frac{\boldsymbol{\omega}_i^\top \boldsymbol{\varphi}_i'(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)})}{\boldsymbol{\omega}_i^\top \boldsymbol{\varphi}_i(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)})} \left[\mathbf{u}_{(k)}\right]_m, \tag{13}$$

where $[\cdot]_m$ is the m^{th} element of the vector. And likewise for the depth coefficients

$$\frac{\partial E(\mathbf{x}; \boldsymbol{\theta})}{\partial g_{im}} = -\sum_{k=1}^{K} \frac{\boldsymbol{\omega}_i^\top \boldsymbol{\varphi}_i'(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)})}{\boldsymbol{\omega}_i^\top \boldsymbol{\varphi}_i(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)})} \left[\mathbf{v}_{(k)}\right]_m. \tag{14}$$

1.2. Sampling the Prior

As suggested by Welling *et al.* [3] we introduce a discrete hidden random vector $\mathbf{z} \in \{1, \cdots, J\}^{N \times K}$, where entry z_{ik} indicates the active scale for expert i on clique k. By definition the distribution of each variable corresponds to the GSM mistrue weight, *i.e.* $p(z_{ik}) = \alpha_{iz_{ik}}$. Because we can ignore the Gaussians that are not active, the joint probability of the image and this hidden random vector is then

$$p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{-\epsilon ||\mathbf{x}||^2/2} \prod_{k=1}^{K} \prod_{i=1}^{N} \alpha_{iz_{ik}} \cdot \mathcal{N}(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; 0, \sigma_i^2/s_{z_{ik}}). \tag{15}$$

To obtain samples from the prior we alternate between sampling the conditional distributions $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})$ and $p(\mathbf{x}|\mathbf{z}; \boldsymbol{\theta})$. In the former case, the variables z_{ik} are independent given the image, so the conditional probability simplifies to

$$p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) \propto \alpha_{iz_{ik}} \cdot \mathcal{N}(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}; 0, \sigma_i^2/s_{z_{ik}}). \tag{16}$$

The conditional probability $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$ can also be expressed as a multivariate Gaussian. We first express the per-clique filtering operations as filters on the entire image

$$\begin{aligned}
p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}) &\propto e^{-\epsilon||\mathbf{x}||^2/2} \prod_{k=1}^{K}\prod_{i=1}^{N} \exp\left(-\frac{s_{z_{ik}}}{2\sigma_i^2}(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)})^2\right) \\
&\propto \exp\left(-\frac{\epsilon}{2}\mathbf{x}^\top\mathbf{x} - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\frac{s_{z_{ik}}}{\sigma_i^2}\left(\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{g}_i^\top \mathbf{v}_{(k)}\right)^2\right) \\
&\propto \exp\left(-\frac{\epsilon}{2}\mathbf{x}^\top\mathbf{x} - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N}\frac{s_{z_{ik}}}{\sigma_i^2}\left(\mathbf{u}_{(k)}^\top\mathbf{h}_i\mathbf{h}_i^\top \mathbf{u}_{(k)} + \mathbf{v}_{(k)}^\top\mathbf{g}_i\mathbf{g}_i^\top \mathbf{v}_{(k)} + \mathbf{u}_{(k)}^\top\mathbf{h}_i\mathbf{g}_i^\top \mathbf{v}_{(k)} + \mathbf{v}_{(k)}^\top\mathbf{g}_i\mathbf{h}_i^\top \mathbf{u}_{(k)}\right)\right) \\
&\propto \exp\left(-\frac{\epsilon}{2}\mathbf{x}^\top\mathbf{x} - \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{u}^\top\sum_{k=1}^{K}\frac{s_{z_{ik}}}{\sigma_i^2}\mathbf{h}'_{ik}\mathbf{h}'^\top_{ik}\mathbf{u} + \mathbf{v}^\top\sum_{k=1}^{K}\frac{s_{z_{ik}}}{\sigma_i^2}\mathbf{g}'_{ik}\mathbf{g}'^\top_{ik}\mathbf{v} + \mathbf{u}^\top\sum_{k=1}^{K}\frac{s_{z_{ik}}}{\sigma_i^2}\mathbf{h}'_{ik}\mathbf{g}'^\top_{ik}\mathbf{v} + \mathbf{v}^\top\sum_{k=1}^{K}\frac{s_{z_{ik}}}{\sigma_i^2}\mathbf{g}'_{ik}\mathbf{h}'^\top_{ik}\mathbf{u}\right)\right) \\
&\propto \exp\left(-\frac{\epsilon}{2}\mathbf{x}^\top\mathbf{x} - \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{u}^\top\mathbf{H}_i\mathbf{Z}_i\mathbf{H}_i^\top\mathbf{u} + \mathbf{v}^\top\mathbf{G}_i\mathbf{Z}_i\mathbf{G}_i^\top\mathbf{v} + \mathbf{u}^\top\mathbf{H}_i\mathbf{Z}_i\mathbf{G}_i^\top\mathbf{v} + \mathbf{v}^\top\mathbf{G}_i\mathbf{Z}_i\mathbf{H}_i^\top\mathbf{u}\right)\right)
\end{aligned} \tag{17}$$

where $\mathbf{h}'_{ik}$ is defined so that $\mathbf{h}'^\top_{ik}\mathbf{u}$ is the result of applying filter $\mathbf{h}_i$ to clique k of the intensity channel $\mathbf{u}$, and $\mathbf{g}'_{ik}$ is defined likewise for the depth channel. $\mathbf{Z}_i = \operatorname{diag}\{s_{z_{ik}}/\sigma_i^2\}$ are diagonal matrices with entries for each clique. $\mathbf{H}_i$ and $\mathbf{G}_i$ are a filter matrices corresponding to a convolution with filter i of the intensity and depth channels respectively, *i.e.* $\mathbf{H}_i^\top\mathbf{u} = [\mathbf{h}'^\top_{i1}\mathbf{u}, \ldots, \mathbf{h}'^\top_{iK}\mathbf{u}]^\top = [\mathbf{h}_i^\top\mathbf{u}_{(1)}, \ldots, \mathbf{h}_i^\top\mathbf{u}_{(K)}]^\top$. Not only is this more efficient to implement through image convolutions, but it also allows the filter matrices to be combined

$$\begin{aligned}
p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}) &\propto \exp\left(-\frac{\epsilon}{2}\mathbf{x}^\top\mathbf{x} - \frac{1}{2}\begin{bmatrix}\mathbf{u}^\top & \mathbf{v}^\top\end{bmatrix}\sum_{i=1}^{N}\begin{bmatrix}\mathbf{H}_i\mathbf{Z}_i\mathbf{H}_i^\top & \mathbf{H}_i\mathbf{Z}_i\mathbf{G}_i^\top \\ \mathbf{G}_i\mathbf{Z}_i\mathbf{H}_i^\top & \mathbf{G}_i\mathbf{Z}_i\mathbf{G}_i^\top\end{bmatrix}\begin{bmatrix}\mathbf{u} \\ \mathbf{v}\end{bmatrix}\right) \\
&\propto \exp\left(-\frac{\epsilon}{2}\mathbf{x}^\top\mathbf{x} - \frac{1}{2}\begin{bmatrix}\mathbf{u}^\top & \mathbf{v}^\top\end{bmatrix}\sum_{i=1}^{N}\begin{bmatrix}\mathbf{H}_i \\ \mathbf{G}_i\end{bmatrix}\mathbf{Z}_i\begin{bmatrix}\mathbf{H}_i^\top & \mathbf{G}_i^\top\end{bmatrix}\begin{bmatrix}\mathbf{u} \\ \mathbf{v}\end{bmatrix}\right) \\
&\propto \exp\left(-\frac{1}{2}\mathbf{x}^\top\left(\epsilon\mathbf{I} + \sum_{i=1}^{N}\mathbf{W}_i\mathbf{Z}_i\mathbf{W}_i^\top\right)\mathbf{x}\right) \\
&\propto \mathcal{N}\left(\mathbf{x}; \mathbf{0}, \left(\epsilon\mathbf{I} + \sum_{i=1}^{N}\mathbf{W}_i\mathbf{Z}_i\mathbf{W}_i^\top\right)^{-1}\right),
\end{aligned} \tag{18}$$

where we use $\mathbf{W}_i^\top = [\mathbf{H}_i^\top\mathbf{G}_i^\top]^\top$ to reach an equation that has the same form as that used by Schmidt [2, 1] for sampling. As in their case, we can further rewrite the covariance matrix as the matrix product

$$\Sigma = \left(\epsilon\mathbf{I} + \sum_{i=1}^{N}\mathbf{W}_i\mathbf{Z}_i\mathbf{W}_i^\top\right)^{-1} = \left(\begin{bmatrix}\mathbf{W}_1, \ldots, \mathbf{W}_N, \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{Z}_1 & & \ldots & 0 \\ & \ddots & & \vdots \\ \vdots & & \mathbf{Z}_N & \\ 0 & \ldots & & \epsilon\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{W}_1^\top \\ \vdots \\ \mathbf{W}_N^\top \\ \mathbf{I}\end{bmatrix}\right)^{-1} = \left(\mathbf{W}\mathbf{Z}\mathbf{W}^\top\right)^{-1} \tag{19}$$

and sample $\mathbf{y} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ to obtain a sample $\mathbf{x}$ from $p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$ by solving the least-squares problem

$$\mathbf{W}\mathbf{Z}\mathbf{W}^\top\mathbf{x} = \mathbf{W}\sqrt{\mathbf{Z}}\mathbf{y} \tag{20}$$

as was shown in [1].

1.3. Conditional sampling

The formulas obtained on the previous section provide the theoretical framework for sampling. However, we often have information for some of the elements of $\mathbf{x}$ and we only want to sample the rest. In this case we need to sample the conditional probability $p(\mathbf{x}_A|\mathbf{x}_B, \mathbf{z}; \boldsymbol{\theta})$, where $\mathbf{x}_B$ and $\mathbf{x}_A$ are the known and unknown elements of $\mathbf{x}$ respectively. This is useful both for learning, where extreme values are less constrained in boundary pixels, and during inference, where, for example, only certain areas of the image need to be inpainted.

We note that $\mathbf{x}$ is logically divided into two parts, $\mathbf{u}$ for intensity and $\mathbf{v}$ for depth, however, the grouping together of the variables for each channel is arbitrary. We reorder $\mathbf{x}$ to group the known and unknown variables instead, *i.e.* $\mathbf{x}' = [\mathbf{x}_A^\top, \mathbf{x}_B^\top]^\top$. For this we use a permutation matrix $\mathbf{P}$ so that $\mathbf{x}' = \mathbf{P}\mathbf{x}$ and $\mathbf{P}\mathbf{P}^\top = \mathbf{P}^\top\mathbf{P} = \mathbf{I}$. This reordering does not affect the shape of the probability distribution, *i.e.* $p(\mathbf{x}'; \boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{x}'; \mathbf{0}, \Sigma')$. By definition, this new covariance matrix is

$$\Sigma' = \langle \mathbf{x}'\mathbf{x}'^\top \rangle = \langle \mathbf{P}\mathbf{x}\mathbf{x}^\top\mathbf{P}^\top \rangle = \mathbf{P}\langle \mathbf{x}\mathbf{x}^\top \rangle \mathbf{P}^\top = \mathbf{P}\Sigma\mathbf{P}^\top \tag{21}$$

Using the result from Eq. (19) we can expand this and define Σ' in terms of submatrices

$$\Sigma' = \mathbf{P}\left(\mathbf{W}\mathbf{Z}\mathbf{W}^\top\right)^{-1}\mathbf{P}^\top = \left(\mathbf{P}\mathbf{W}\mathbf{Z}\mathbf{W}^\top\mathbf{P}^\top\right)^{-1} = \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{bmatrix}^{-1} \tag{22}$$

where $\mathbf{A}$ and $\mathbf{B}$ are a square matrices with the same number of rows as $\mathbf{x}_A$ and $\mathbf{x}_B$ respectively. We can now obtain the conditional distribution of the unknown variables given those that are known

$$p(\mathbf{x}_A|\mathbf{x}_B, \mathbf{z}; \boldsymbol{\theta}) \propto \mathcal{N}\left(\mathbf{x}'; \mathbf{0}, \Sigma'\right) \tag{23}$$

$$\propto \exp\left(-\frac{1}{2}\begin{bmatrix}\mathbf{x}_A \\ \mathbf{x}_B\end{bmatrix}^\top \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{bmatrix} \begin{bmatrix}\mathbf{x}_A \\ \mathbf{x}_B\end{bmatrix}\right) \tag{24}$$

$$\propto \exp\left(-\frac{1}{2}\left(\mathbf{x}_A + \mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B\right)^\top \mathbf{A}\left(\mathbf{x}_A + \mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B\right)\right) \tag{25}$$

$$\propto \mathcal{N}\left(\mathbf{x}_A; -\mathbf{A}^{-1}\mathbf{C}\mathbf{x}_B, \mathbf{A}^{-1}\right) \tag{26}$$

This allows for the same efficient sampling scheme. The sampling of the scale vector $\mathbf{z}$ remains as before, *i.e.* $p(\mathbf{z}|\mathbf{x}_A, \mathbf{x}_B; \boldsymbol{\theta}) = p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})$.

References

[1] U. Schmidt, Q. Gao, and S. Roth. A generative perspective on MRFs in low-level vision. In *CVPR*, 2010.

[2] Uwe Schmidt. Learning and Evaluating Markov Random Fields for Natural Images. Master's thesis, 2010.

[3] M. Welling, G. Hinton, and S. Osindero. Learning sparse topographic representations with products of student-t distributions. *Advances in neural information processing systems*, pages 1383–1390, 2003.