From Patches to Natural Images via Hierarchical Dirichlet Processes

Geng Ji[1], **Michael C. Hughes**[2], and **Erik B. Sudderth**[1]

[1]Department of Computer Science, Brown University, {gji,sudderth}@cs.brown.edu
[2]School of Engineering and Applied Sciences, Harvard University, mike@michaelchughes.com

Abstract

We propose a hierarchical generative model that captures the self-similar structure of image regions as well as how this structure is shared across image collections. Our model is based on a novel, variational interpretation of the popular expected patch log-likelihood (EPLL) method as a model for randomly positioned grids of image patches. While previous EPLL methods modeled the density of image patches with finite Gaussian mixtures, we use nonparametric Dirichlet process (DP) mixtures to create models whose complexity grows as additional images are observed. An extension based on the hierarchical DP then captures the repetitive and self-similar structure of image regions via image-specific variations in cluster frequencies. We derive a structured variational inference algorithm that uses birth and delete moves to create new patch clusters and thus more accurately model novel image textures. Our denoising performance on standard benchmarks is superior to EPLL and comparable to the state-of-the-art, while providing a novel statistical interpretation for many common image processing heuristics.

1 Introduction

Models of the statistical structure of natural images play a key role in many computer vision and image processing tasks [1]. Due to the high dimensionality of the images captured by modern cameras, a rich research literature instead models the statistics of small image patches. For example, the K-SVD method [2] generalizes K-means clustering to learn a dictionary for sparse coding of image patches. The state-of-the-art *learned simultaneous sparse coding* (LSSC) [3] and *block matching and 3D filtering* (BM3D) [4] methods integrate clustering, dictionary learning, and denoising to extract information directly from a single corrupted image. Alternatively, the accurate *expected patch log-likelihood* (EPLL) [5] method models an overlapping grid of natural image patches using finite Gaussian mixtures learned from a collection of uncorrupted natural images.

We show that with minor modifications, the objective function underlying EPLL is equivalent to a variational log-likelihood bound for a novel generative model of whole images. Our model coherently captures overlapping image patches via a randomly positioned spatial grid. By deriving a rigorous variational bound, we then develop improved nonparametric models of natural image statistics using the *hierarchical Dirichlet process* (HDP) [6]. In particular, DP mixtures allow an appropriate model complexity to be inferred from data, while the hierarchical DP captures the patch self-similarities and repetitions that are ubiquitous in natural images [1]. Unlike previous whole-image generative models such as *fields of experts* (FoE) [7], which use a single set of Markov random field parameters to model all images, our HDP model allows the learning of image-specific clusters to accurately model distinctive textures. Coupled with a scalable structured variational inference algorithm, we match the state-of-the-art denoising performance of the LSSC and BM3D algorithms, while providing a Bayesian nonparametric model with a broader range of potential applications.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

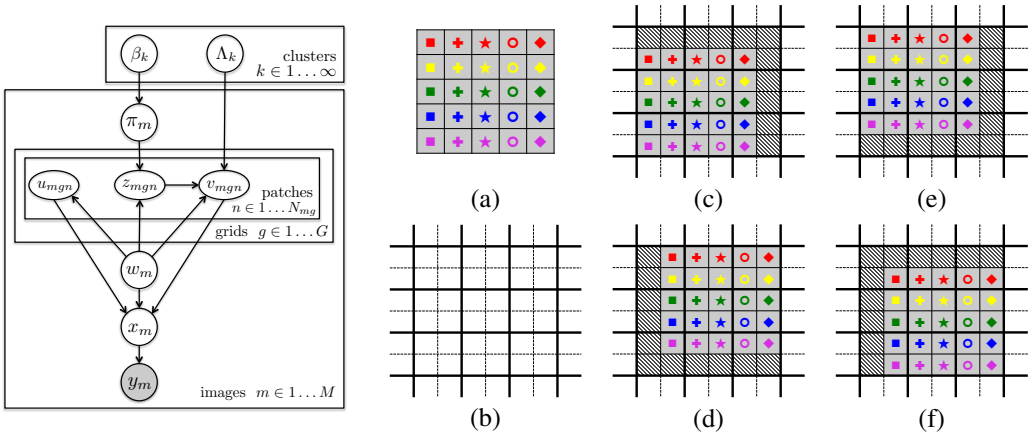

Figure 1: *Left:* Graphical representation for our HDP-Grid model. *Right:* Diagram of how our model generates whole images from a randomly selected grid of non-overlapping patches: (a) A 5×5 pixel image whose pixels are uniquely specified by the color and shape of the symbol inside; (b) An infinitely large 2-dimensional grid of pixels, divided into 2×2 patches; (c-f) The four possible ways for the image in (a) to be constructed from 2×2 patches. The shaded pixels would be clipped by the image boundary, as described in Sec. 2.

2 A Hierarchical DP Mixture Model for Grids of Image Patches

Hierarchical Dirichlet process mixture models. The *hierarchical Dirichlet process* (HDP) [6] is a Bayesian nonparametric prior used to perform clustering across groups of related data. We use the HDP prior to model natural images, where each image is a "group" of patches. The HDP allows us to share structure, such as patches of grass or patches of bricks, by sharing a common set of clusters (also called *topics*) across images. In addition, the HDP models image-specific variability by allowing each image to use this shared set of clusters with custom frequencies, so that grass might be abundant in one image but absent in another. Finally, we can learn the proper number of hidden clusters or topics from data, and discover new clusters as we collect new images with novel visual textures.

For each of the countably infinite set of clusters indexed by k, the HDP uses a stick-breaking construction to generate a corpus-wide frequency vector $\eta = [\eta_1, \eta_2, \dots, \eta_k, \dots]$ that sums to one:

$$\beta_k \sim \text{Beta}(1, \gamma), \quad \eta_k \triangleq \beta_k \textstyle\prod_{\ell=1}^{k-1}(1 - \beta_\ell). \tag{1}$$

Just like topic models of text data [8], the HDP allows each image m to have its own topic frequencies π_m, where the vector η determines the mean of a DP prior on the frequencies of shared topics:

$$\pi_m \sim \text{DP}(\alpha\eta), \qquad E[\pi_{mk}] = \eta_k. \tag{2}$$

When the concentration parameter $\alpha < 1$, we capture the "burstiness" and self-similarity of natural image regions [1] by placing most probability mass in π_m on a sparse subset of the global topics.

Image generation via random patch grids. Unlike text data, images are not naturally divided into meaningful word tokens. We thus generate the pixels in image m via a randomly placed grid of patches, each with G total pixels. (In our experiments, $G = 8 \times 8 = 64$.) This implies there are exactly G possible grid alignments for an image of arbitrary size, as illustrated in Fig. 1. Let $w_m \in \{1, \dots, G\}$, the alignment randomly chosen for image m, be sampled from a uniform prior:

$$w_m \sim \text{Categorical}(1/G, \dots, 1/G). \tag{3}$$

Patch generation via Gaussian mixtures. Given a particular non-overlapping grid of patches, our HDP-Grid model generates each patch as an independent draw from a Gaussian mixture model. Gaussian mixtures have recently been shown to produce surprisingly good density models for natural image patches [9]. Our model assumes that each cluster is defined by a zero-mean, full-covariance Gaussian over the G pixels in each image patch. We parameterize cluster k by a precision (inverse covariance) matrix Λ_k, which is drawn from a conjugate Wishart prior. The set of image-specific frequencies $\{\pi_{mk}\}_{k=1}^{\infty}$ and precision matrices $\{\Lambda_k\}_{k=1}^{\infty}$ then define an "infinite" Gaussian mixture.

If image m is assigned to grid alignment g, we sample topic assignments to each of the N_{mg} patches in grid g. For the patch at index n, let $z_{mgn} \in \{1, 2, \dots\}$ denote the integer id of the chosen topic.

Patch n also has a vector of pixel values v_{mgn} of length G, generated as follows:

$$z_{mgn}|w_m = g \sim \text{Categorical}(\pi_m), \quad v_{mgn}|w_m = g, z_{mgn} = k \sim \mathcal{N}(0, \Lambda_k^{-1}). \tag{4}$$

Note that as the Gaussian topics all have zero mean, to match the pixel intensity distributions of real images, we add a *DC offset* u_{mgn} to each patch [5], a scalar value not dependent on z_{mgn}:

$$u_{mgn}|w_m = g \sim \mathcal{N}(r, s^2). \tag{5}$$

Finally, we assemble the generated patches v_m into a whole clean image, which we denote by x_m:

$$x_m|w_m = g, u_m, v_m \sim \mathcal{N}\Big(\sum\nolimits_{n=1}^{N_{mg}} P_{mgn}^T\big(C_{mgn}v_{mgn} + u_{mgn}\big), \delta^2 I\Big) \tag{6}$$

This model sets the mean of the image x_m by stitching together all patches in the chosen grid g, and then adds per-pixel noise with small variance δ^2. Most patches in the chosen grid will be fully observed in x_m, but some near the boundaries may have pixels clipped off, as shown in Fig. 1. The matrix C_{mgn} performs this mapping for each patch, so that vector $C_{mgn}v_{mgn} + u_{mgn}$ represents the observed pixels in patch n. This observed pixel vector is then projected to its appropriate place within the whole image vector via a binary indicator matrix P_{mgn}.

For image restoration tasks, the observed image y_m is a corrupted version of some clean image x_m that we would like to estimate. Specifically, we consider denoising tasks where the model is:

$$y_m|x_m \sim \mathcal{N}(x_m, \sigma^2 I). \tag{7}$$

The variance $\sigma^2 \gg \delta^2$ indicates the noise level.

3 Structured Variational Inference

Posterior inference for each image m must consider multiple possible grid alignments that could have generated the image. This reasoning about multiple alignments is what makes our model a powerful generative model for whole images; grid overlap produces posterior dependencies between patches.

In this section, we provide a high-level summary of our structured mean-field variational posterior q:

$$q \triangleq \prod\nolimits_{k=1}^{K} q(\Lambda_k)q(\beta_k) \cdot \prod\nolimits_{m=1}^{M} q(\pi_m)q(w_m)q(x_m) \prod\nolimits_{n=1}^{N_{gn}} q(u_{mgn}|w_m)q(z_{mgn}, v_{mgn}|w_m) \tag{8}$$

We train this approximate posterior by optimizing a lower bound on the marginal log-likelihood of the observed images. This objective is also known as the *evidence lower bound* (ELBO), and could be rewritten (up to a constant) as the negative of the KL divergence between the joint posterior and the variational approximation q [10]. Below, we describe each factor of q and its associated free parameters (which are denoted as letters with hats). Our algorithm performs coordinate ascent on the factors of q, iteratively updating each one while holding others fixed. Many updates are standard due to the model's conjugate structure and are omitted for brevity.

Patch-level. We assume the variables specific to patch n in grid g have a structured posterior, which conditions on the chosen grid alignment g and chosen topic assignment k. We have the following posteriors for assignments z, pixel vectors v, and scalar offsets u:

$$q(z_{mgn}|w_m = g) = \text{Categorical}(\hat{r}_{mgn1}, ..., \hat{r}_{mgnK}), \tag{9}$$

$$q(v_{mgn}|w_m = g, z_{mgn} = k) = \mathcal{N}(\hat{v}_{mgnk}, \hat{\phi}^v_{mgnk}),\ q(u_{mgn}|w_m = g) = \mathcal{N}(\hat{u}_{mgn}, \hat{\phi}^u_{mgn}). \tag{10}$$

This construction provides a rigorous derivation for the heuristically motivated EPLL algorithm [5].

Image-level. In the real world, we expect no single grid alignment to do a noticeably better job of explaining some natural image. Thus, we directly enforce in our approximate posterior that all G possible grid locations have the same uniform probability: $q(w_m) = \text{Categorical}(\frac{1}{G}, \ldots, \frac{1}{G})$.

Other image-level variables have standard forms for the approximate posterior. We set the whole-image posterior $q(x_m)$ to be a normal distribution, whose optimal covariance is diagonal. Given a current hypothesis K for the number of topics that have been observed at least once, we truncate our posterior on image-specific topic frequencies $q(\pi_m)$ to a finite Dirichlet distribution as in [11].

Corpus-level. To model the structure shared across a collection of images, we learn the parameters of our hierarchical DP Gaussian mixture via variational posteriors $q(\Lambda_k) = \text{Wishart}(\hat{\nu}_k, \hat{W}_k)$ and $q(\beta_k) = \text{Beta}\big(\hat{\rho}_k\hat{\omega}_k, (1 - \hat{\rho}_k)\hat{\omega}_k\big)$. We only explicitly compute posterior statistics for the K topics that have been assigned to at least one patch. The infinite set of not-yet-observed image topics is then tractably approximated by setting their variational parameters to match the prior [11].

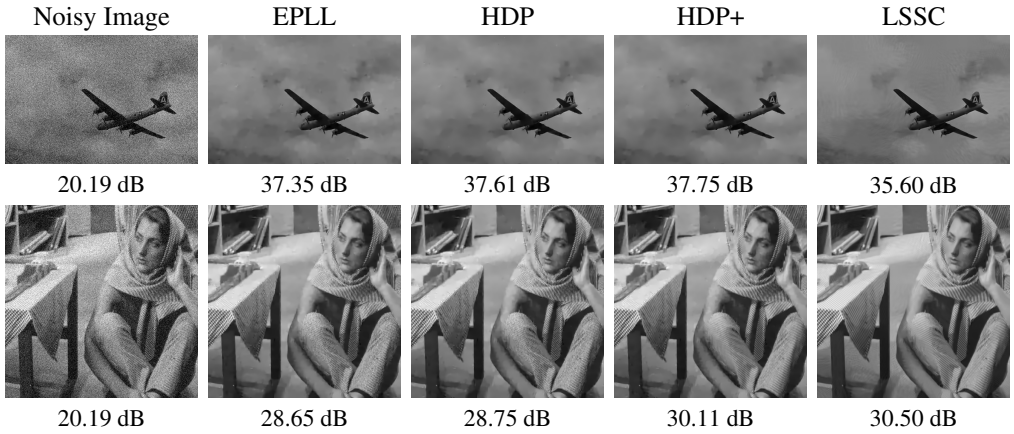

Figure 3: *Top:* An image from BSDS-68. Algorithms like LLSC, which learn solely from the noisy image, may not accurately model the range of natural image textures. *Bottom:* An image from classic-12 that shows our ability to create new, image-specific topics. In areas like the pants, scarf, and tablecloth, HDP+ is clearly superior to the HDP and EPLL methods in restoring stripes of varying width and orientation. **Best viewed online.**

Image-specific topics. Most images contain unique textural patterns. The Bayesian nonparametric nature of our model allows us to create novel, image-specific topics during inference. For each test image, we augment the existing K topics with $K' = 100$ new topics, initialized using a generalization of `k-means++` [12] to the Bregman divergence associated with our zero-mean, full-covariance Gaussian likelihood [13]. This procedure samples K' diverse patches from the test image. We use these patches to initialize K' new clusters, and refine them via the corpus-level variational updates. The resulting posterior approximation to the test image, and all training images, has $K + K'$ total topics. While we initialize K' to a large number to avoid local optima, this may lead to extraneous topics, so we delete any new topics that are not assigned to any patch. In the bottom image of Fig. 3, this leaves 9 image-specific topics. This deletion improves model interpretability and algorithm speed, because costs scale linearly with the number of instantiated topics.

4 Experimental Validation: Image Denoising

We train our HDP-Grid model using the 400 training and validation images from the Berkeley Segmentation Dataset (BSDS) [14]. We consider three variants of our method: (i) a DP mixture constrains images to have the same distribution $\pi_m = \eta$, and learns $K = 90$ topics; (ii) an HDP mixture allows image-specific variation in frequencies π_m, and learns $K = 146$ topics; (iii) HDP+, the HDP mixture augmented with up to $K' = 100$ novel image-specific topics. Fig. 2 compares clean image ELBO and PSNR values on 12 "classic" images used in many previous denoising papers [3; 5]. We see clear increases in both ELBO and PSNR as we transition from DP to HDP to HDP+, suggesting that learning image-specific frequencies and textures is useful. Examples in Fig. 3 illustrate this trend.

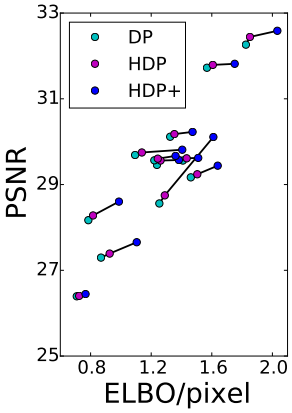

Figure 2: Clean image ELBO vs. PSNR for 12 "classic" images. Better statistical models have improved denoising accuracy.

Table 1 compares average denoising performance to many other methods; BSDS-68 denotes the 68 images from [14] used by [7; 5]. Our denoising performance is superior to the state-of-the-art on this dataset, illustrating the value of Bayesian nonparametric learning from large image collections. Our performance is competitive with top methods tuned to perform well on classic-12, where repeated textures make the HDP+ variant of our model particularly effective.

Table 1: Average PSNR values on two benchmark datasets, for Gaussian noise with standard deviation 25.

Dataset	DP	HDP	HDP+	EPLL	FoE	KSVDG	KSVD	BM3D	LSSC
classic-12	29.33	29.42	29.63	29.39	28.28	28.88	29.10	29.74	**29.75**
BSDS-68	28.67	28.72	**28.78**	28.72	27.71	28.28	28.27	28.56	28.70

References

[1] Anuj Srivastava, Ann B Lee, Eero P Simoncelli, and Song Chun Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003.

[2] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[3] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *International Conference on Computer Vision*, 2009.

[4] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image restoration by sparse 3d transform-domain collaborative filtering. In *Electronic Imaging*, 2008.

[5] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *International Conference on Computer Vision*, 2011.

[6] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[7] Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *Computer Vision and Pattern Recognition*, 2005.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[9] Daniel Zoran and Yair Weiss. Natural images, Gaussian mixtures and dead leaves. In *Neural Information Processing Systems*, 2012.

[10] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[11] Michael C Hughes, Dae Il Kim, and Erik B Sudderth. Reliable and scalable variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2015.

[12] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[13] Marcel R. Ackermann and Johannes Blömer. Coresets and approximate clustering for Bregman divergences. In *ACM-SIAM Symposium on Discrete Algorithms*, 2009.

[14] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *International Conference on Computer Vision*, 2001.