How Well Do Filter-Based MRFs Model Natural Images?

Qi Gao and Stefan Roth

Department of Computer Science, TU Darmstadt

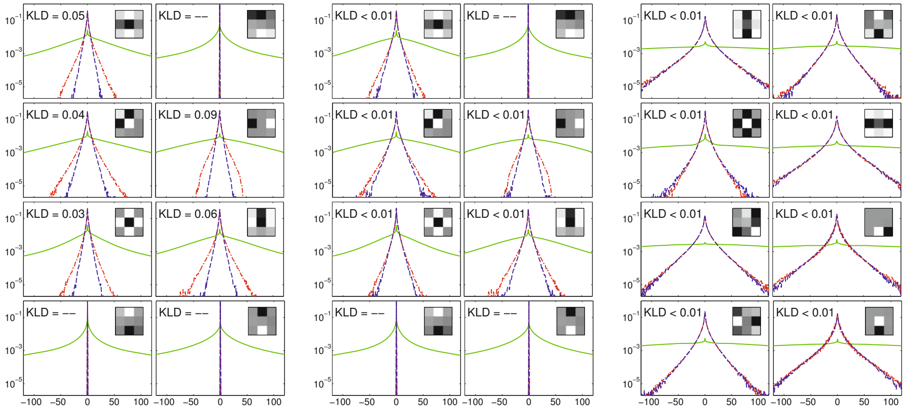
Abstract. Markov random fields (MRFs) have found widespread use as models of natural image and scene statistics. Despite progress in modeling image properties beyond gradient statistics with high-order cliques, and learning image models from example data, existing MRFs only exhibit a limited ability of actually capturing natural image statistics. In this paper we investigate this limitation of previous filter-based MRF models, which appears in contradiction to their maximum entropy interpretation. We argue that this is due to inadequacies in the leaning procedure and suggest various modifications to address them. We demonstrate that the proposed learning scheme allows training more suitable potential functions, whose shape approaches that of a Dirac-delta function, as well as models with larger and more filters. Our experiments not only indicate a substantial improvement of the models' ability to capture relevant statistical properties of natural images, but also demonstrate a significant performance increase in a denoising application to levels previously unattained by generative approaches.

1 Introduction and Related Work

Both analysis and modeling of the statistics of natural images and scenes have a long history. Statistical analyses have been carried out for natural images, image categories, range images, optical flow, *etc*. [4,13,15]. They have revealed many characteristic properties of natural images and scenes, including power-law frequency spectra, non-Gaussian highly kurtotic marginals, scale-invariant statistics, and non-Gaussian joint statistics of nearby image features [11,15]. These properties have been exploited in various statistical models, local ones that attempt to capture the statistics of one or a few features [11], as well as global models that aim to represent the properties of entire images. Latter often take the form of Markov random fields (MRFs).

MRFs based on linear filter responses, here termed *filter-based MRFs*, are perhaps the most popular form for modeling natural image priors [2,14,19,21]. The design of such models involves various choices, including the size and shape of the cliques, the selection of the image filters, and the shape of the potential functions. Pairwise MRFs are most widely used, for which the filters are simple image derivatives. The FRAME model [21] is an early instance of filter-based, *high-order* MRFs, in which discretized potentials are learned from data, and the filters are automatically chosen from a hand-designed set of candidates. The more recent Fields of Experts (FoEs) [14] use continuous potential functions and additionally learn the filters from training data to achieve better results in practice. Despite success in various applications [9,19], recent work [16] based on drawing samples from FoE priors [14,19] found that they represent

A. Pinz et al. (Eds.): DAGM/OAGM 2012, LNCS 7476, pp. 62–72, 2012.
© Springer-Verlag Berlin Heidelberg 2012



(a) CD-trained FoE; parameters from the public implementation of [16].

(b) FoE with filters from (a), but potentials re-trained with PCD.

(c) FoE trained with proposed learning procedure.

Fig. 1. Filter-based MRFs and image statistics: Image filter (3×3, top right) with corresponding learned potential function (solid, green), marginal histograms of natural images (dash-dotted, red), and model marginals obtained by sampling (dashed, blue). The proposed learning scheme leads to a better match to natural image statistics (top left – marginal KL-divergence).

relevant statistical properties of natural images only quite crudely. This appears in contradiction to the maximum entropy interpretation of filter-based MRFs [21], which suggests that the model should capture the statistics of the in-built features, at least if the potential functions are chosen and learned appropriately. [16] attributed this to previously used potentials not being heavy-tailed enough and suggested learning the shape of a more flexible potential, taken to be a Gaussian scale mixture (GSM) [11]. While this allowed to learn pairwise MRFs that capture derivative statistics correctly, the high-order case was more problematic: Marginal statistics of model samples were found not to be as heavy-tailed as those of natural images (Fig. 1a). Moreover, their study was limited to moderate clique sizes and a comparatively small number of filters.

In this paper we aim to address these issues and explore the limits of how well filter-based MRFs can capture natural image statistics. Motivated by the observation that larger support sizes lead to clearly improved performance bounds for image denoising [6], we also aim to learn models with larger cliques. We propose an improved learning procedure that *(1)* reduces training bias by replacing contrastive divergence (CD) [3] with persistent CD [17] as the learning objective; *(2)* improves robustness by imposing filter normalization and using initializations that allow the model to learn more varied filters; and *(3)* uses a new boundary handling method for sampling, which reduces sampling bias and thus increases both accuracy and efficiency. Our approach has various benefits: It makes learning more robust and consequently enables training models with larger and more filters that exhibit a more structured appearance (Fig. 5). Moreover, it enables learning improved potential functions that are extremely heavy-tailed, almost Dirac-delta like (Fig. 1c), which allow the model to capture the statistics of the model features correctly; the trained models are thus real *maximum entropy models*. More importantly and in contrast to previous approaches, the trained models also represent multi-scale derivative statistics, random filter statistics, as well as joint feature statistics of natural images quite accurately. Image denoising experiments show that this improvement in modeling accuracy also translates into sizable performance gains of approximately 0.3dB to the level of state-of-the-art denoising techniques, such as non-local sparse coding [8] or BM3D [1]. To the best of our knowledge this is the first time this has been achieved with a purely generative, global model of natural images.

Other Related Work. There is an extensive literature on learning methods for MRF models, including those of natural images. Difficulties arise from the intractability of inference and from the likelihood being generally multimodal. Besides MCMC-based approaches [21] and approximations including contrastive divergence [3] or persistent CD [17], deterministic methods including basis rotation [19] and score matching [5] have been used. These approaches have relied on particular potential functions, either fitted off-line or with limited expressiveness, which constrains their applicability. We instead learn potentials based on Gaussian scale mixtures (GSMs), which have found widespread use in local image models [11] and as potentials in global MRF models [16,19]. One of our contributions is to show that GSMs are sufficiently flexible to allow even high-order MRFs to model image statistics to a high degree of accuracy.

2 Basic Model and Learning Procedure

In this paper we explore the capabilities of filter-based, high-order MRFs (*e.g.* [2]). For ease of comparison to previous analyses, we use the particular form of [16]. The prior probability density of a natural image $\mathbf{x}$ under such a model is written as

$$p(\mathbf{x}; \boldsymbol{\Omega}) = \frac{\mathcal{N}_\epsilon(\mathbf{x})}{Z(\boldsymbol{\Omega})} \prod\nolimits_{c \in C} \prod\nolimits_{i=1}^{F} \phi\big(\mathbf{f}_i^{\mathrm{T}} \mathbf{x}_{(c)}; \boldsymbol{\omega}_i\big). \tag{1}$$

The $\mathbf{f}_i$ are the linear filters, and $\phi(\cdot; \boldsymbol{\omega}_i)$ is the respective potential function (or factor/ expert) with parameter $\boldsymbol{\omega}_i$. Further, $c \in C$ denote the model cliques, and $Z(\boldsymbol{\Omega})$ is the partition function that depends on the model parameters $\boldsymbol{\Omega} = \{\mathbf{f}_i, \boldsymbol{\omega}_i | i = 1 \dots F\}$. A broad (unnormalized) Gaussian $\mathcal{N}_\epsilon(\mathbf{x}) = e^{-\epsilon \|\mathbf{x}\|^2/2}$ ensures normalizability (*cf*. [19]).

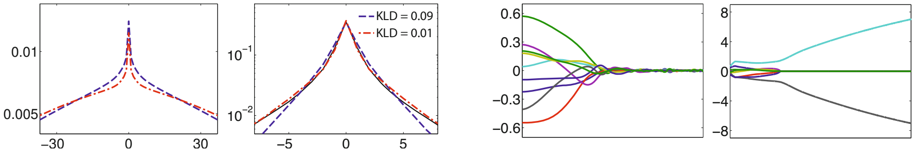
Due to their flexibility for representing a wide variety of heavy-tailed, highly kurtotic distributions, we follow previous work [7,16,19] and use Gaussian scale mixtures (GSMs) [11] to represent the potentials. In their finite form they can be written as

$$\phi\big(\mathbf{f}_i^{\mathrm{T}} \mathbf{x}_{(c)}; \boldsymbol{\omega}_i\big) = \sum\nolimits_{k=1}^{K} \omega_{ik} \cdot \mathcal{N}\big(\mathbf{f}_i^{\mathrm{T}} \mathbf{x}_{(c)}; 0, z_{ik} \cdot \sigma_i^2\big), \tag{2}$$

where $\omega_{ik} \geq 0, \sum_k \omega_{ik} = 1$ are the mixture weights of the scales z_{ik}. Note that here we use fixed variances $z_{ik} \cdot \sigma_i^2$ of the Gaussian components.

Basic Learning Strategy. Learning the model parameters $\boldsymbol{\Omega}$ from data involves estimating the weights ω_{ik} of the GSM, and in case of Fields of Experts also the filters $\mathbf{f}_i$. The classical learning objective for training models of natural images is maximum likelihood (ML, see Sec. 1 for alternatives). A gradient ascent on the log-likelihood for a parameter Ω_i leads to the update

$$\Omega_i^{(t+1)} = \Omega_i^{(t)} + \eta \big[\big\langle \tfrac{\partial E}{\partial \Omega_i} \big\rangle_p - \big\langle \tfrac{\partial E}{\partial \Omega_i} \big\rangle_{\mathbf{X}^0} \big], \tag{3}$$



(a) (*left*) Potentials; (*right*) marginal statistics.

(b) Evolution of filter coefficients during learning.

Fig. 2. *(a)* Difference between potentials (for one filter from Fig. 1(a), 2nd row & 2nd column) trained with 1-step CD (dashed, blue) and PCD (dash-dotted, red), as well as resulting model marginals (magnified for display). The marginal KL-divergence is given w.r.t. natural images (solid, black). *(b)* Filter coefficients may decay or disperse without filter normalization.

where E is the unnormalized Gibbs energy according to $p(\mathbf{x}; \boldsymbol{\Omega}) = e^{-E(\mathbf{x};\boldsymbol{\Omega})}/Z(\boldsymbol{\Omega})$, η is the learning rate, $\langle\cdot\rangle_{\mathbf{X^0}}$ denotes the average over the training data $\mathbf{X^0}$, and $\langle\cdot\rangle_p$ denotes the expectation value w.r.t. the model distribution $p(\mathbf{x}; \boldsymbol{\Omega}^{(t)})$.

One conceptual advantage is that this minimizes the Kullback-Leibler (KL) divergence between the model and the data distribution and, in principle, makes the model statistics as close to those of natural images as possible. Various difficulties, however, arise in practice. First, there is no closed form expression for the model expectation, and an exact computation is intractable. Approximate inference, *e.g*. using sampling, must thus be used. Markov chain Monte Carlo (MCMC) approximations are historically most common (*e.g*. [21]), but very inefficient. Consequently, ML estimation itself was frequently approximated by contrastive divergence (CD) [3], which avoids costly equilibrium samples: Samplers are initialized at the training data $\mathbf{X^0}$ and only run for n (usually a small number) MCMC iterations to yield the sample set $\mathbf{X}^n$. Then $\langle \partial E/\partial \Omega_i\rangle_{\mathbf{X}^n}$ is used to replace $\langle \partial E/\partial \Omega_i\rangle_p$ in Eq. (3). We here use CD as the basis. A second challenge is the speed of mixing, which is usually addressed with efficient sampling methods, such as hybrid Monte Carlo [20] or auxiliary-variable Gibbs samplers [16]. We employ the latter and use the publicly available implementation of [16].

3 Improved Learning Scheme

The basic learning procedure from Sec. 2 involves a series of approximations. Moreover, the data likelihood is generally multimodal, leading to locally optimal parameters. Since previous filter-based, high-order MRFs failed to capture image statistics accurately (*cf*. Fig. 1a), the shortcomings in learning are a possible cause. We here investigate this issue, show that such a standard learning procedure is insufficient to learn accurate models of natural images, and propose an improved learning scheme.

3.1 PCD *vs*. CD

Although contrastive divergence is a reasonably good and formally justified approximation of maximum likelihood [3], it may still incur a training bias. While using n-step CD (with large n) may reduce the bias, learning becomes much less efficient. Thus previous work typically relied on 1-step CD [20], particularly for high-order models [16]. We instead use *persistent contrastive divergence* (PCD) [17], in which the samplers are not reinitialized each time the parameters are updated. Instead, the samples from the previous iteration are retained and used for initializing the next iteration. Combined with a small learning rate, the samplers are thus held close to the stationary distribution:

$$\left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_{\mathbf{X}^{\mathrm{PCD}}} \approx \left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_{\mathbf{X}^{\infty}} \approx \left\langle \frac{\partial E}{\partial \Omega_i} \right\rangle_{p}. \tag{4}$$

Thus each parameter update closely approximates a true ML update step as in Eq. (3). Note that even with a small learning rate, PCD has an efficiency comparable to that of 1-step CD, but substantially reduces bias as the experiments below show. Note that while PCD has been used to train Restricted Boltzmann Machines [17] and filter-based MRFs with Student-t potentials [12], this is the first time it has been investigated in conjunction with more flexible GSM potentials.

Replacing CD with PCD not only reduces training bias, but more importantly improves the models' properties significantly. To demonstrate this, we use the 3×3 FoE from the public implementation of [16] as a basis and retrain the potentials with PCD, while keeping the filters fixed. The resulting marginal statistics of the in-built model features (Fig. 1b) match those of natural images well; all marginal KL-divergences are below 0.01. Fig. 2(a) shows in detail the parts where PCD affects the potential shape the most and most improves the resulting model marginal. Another notable benefit of using PCD is that it enables the following improved boundary handling scheme.

3.2 Boundary Handling for Sampling

Boundary pixels are a common source of problems in MRFs, since they are overlapped by fewer cliques, making them less constrained than those in the image interior. When sampling the model, boundary pixels of the samples tend to take extreme values, which affects both learning and analysis of the model through sampling. Norouzi *et al.* [10] proposed to use conditional sampling, *i.e.* keeping a small number of pixels around the boundary fixed and conditionally sampling the interior. The drawback of this scheme is that the boundary pixels will significantly diffuse into the interior during sampling, which can be seen from the example in Fig. 3(a,b). To reduce bias in learning and evaluation of the model, a thick boundary from the samples thus has to be discarded. The disadvantage is that this lowers the accuracy and the efficiency of learning.

To address this, we instead use *toroidal sampling*, in which the cliques are extended to also cover the boundary pixels in a wrap-around fashion. The toroidal topology used

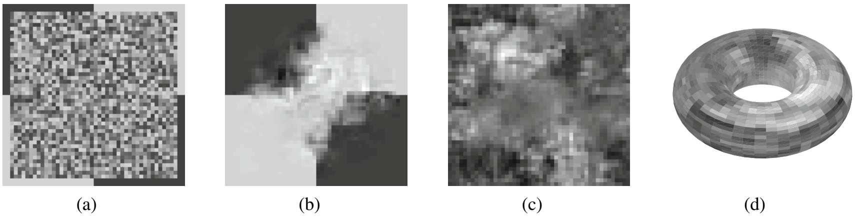


Fig. 3. Effect of boundary handling on samples: *(a)* Initialization of the sampler; *(b)* typical sample generated by conditional sampling. Note how the boundaries affect the interior of the sample; *(c, d)* typical sample and its topology generated by the proposed toroidal sampling.

during sampling is shown in Fig. 3(d). The obvious benefit of using this topology is the absence of any boundary pixels; all pixels are overlapped by the same number of cliques, and there are as many cliques as pixels. Since all pixels are constrained equally, boundary artifacts are avoided, and bias from the boundaries during learning is avoided.

Fig. 3(c) shows how toroidal sampling is less affected by its initialization and can quickly explore a large space. The generated samples will in turn make learning more accurate, while not requiring boundary pixels to be discarded. This increases the learning efficiency, because fewer parallel samplers suffice to estimate the likelihood gradient accurately. It is important to note that while PCD allows using toroidal sampling, the more common CD does not. This is because CD repeatedly initializes the samplers from the training images, which usually do not satisfy periodic boundary conditions.

3.3 Filter Normalization

Some researchers (*e.g.* [5,20]) have suggested to impose constraints on the norms of filters, because filters may otherwise become "inactive", *i.e.* decay to zero during training. As zero filters and the corresponding potentials do not contribute to the model at all, this is an issue, especially when a large number of filters are trained [5,20]. But even with fewer filters as are used here (the flexibility of the GSM potentials imposes limits on the attainable number of filters), we observe that filter coefficients may decay or disperse during learning (Fig. 2b). To address this, we normalize the coefficients of each filter to unit ℓ^2 norm after each parameter update. This incurs no loss of generality due to the redundancy between the GSM scales and the filter coefficient norm: GSM potentials with an infinitely large range of scales can in principle adapt to filters with arbitrary coefficient norm. The necessarily limited range of GSM scales in practice, however, does not allow to properly model the potentials if the filters take extreme values. Moreover, removing the parameter redundancy increases robustness, and in turn enables learning more filters. Fig. 1(c) shows an example in which all 8 filters are "active" and contribute to the model. This is in contrast to the learning approach of [16], for which 3 out of 8 learned filters are effectively inactive (Fig. 1a). Unlike previous uses [5,20], combining filter normalization with more flexible potentials enables learning different, heavy tailed potentials. These notably improve the ability to capture the marginal statistics of the in-built features (Fig. 1c). This also suggests that the learning procedure rather than the representational capabilities of GSMs had been the limiting factor in previous work.

3.4 Initialization of Parameters

The final aspect we address is that of initialization, which is crucial due to the non-convexity of the data likelihood. Specifically, we found that the initialization of the potential shape (GSM weights) can significantly affect learning, including the filters. A uniform initialization of the GSM weights (Fig. 4, red curve) as used, *e.g.*, by [16] is problematic. This overly constrains the pixel values and makes model samples spatially flat. The filter responses on the samples thus fall into a much smaller range than those on training images. The learning algorithm aims to reduce this difference by changing the filters toward patterns that reduce the filter-response range on natural images. The effect is that filters, particularly a Laplacian (Fig. 4, middle), are redundantly learned.

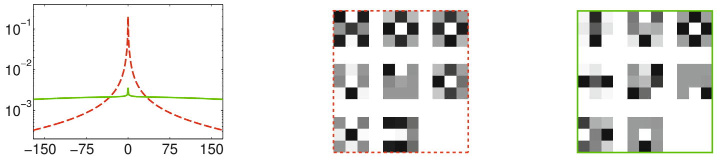


Fig. 4. Typical uniform initialization of GSM weights (dashed, red) leads to filters with fewer patterns (*middle*). Broad (δ-like) initialization (solid, green) leads to more varied filters (*right*).

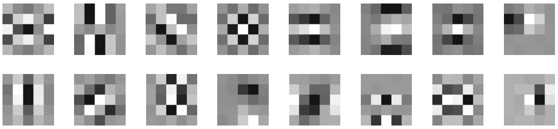


Fig. 5. 16 learned filters of size 5×5. Note their more structured appearance compared to [14]

We alleviate this by initializing the potentials such that the pixels are initially less constrained than they should be, rather than more. To that end, we initialize the potentials with a broad δ-like shape (Fig. 4, green curve). We found that this improves the robustness of learning and enables training a more varied set of filters that captures different kinds of spatial structures (Fig. 4, right). Our findings indicate that the filters, on the other hand, are best initialized randomly. Initializing them to interpretable filters, such as Gabors, is counterproductive as these are usually not optimal for filter-based MRFs / FoEs, and lead to training becoming stuck in poor local optima.

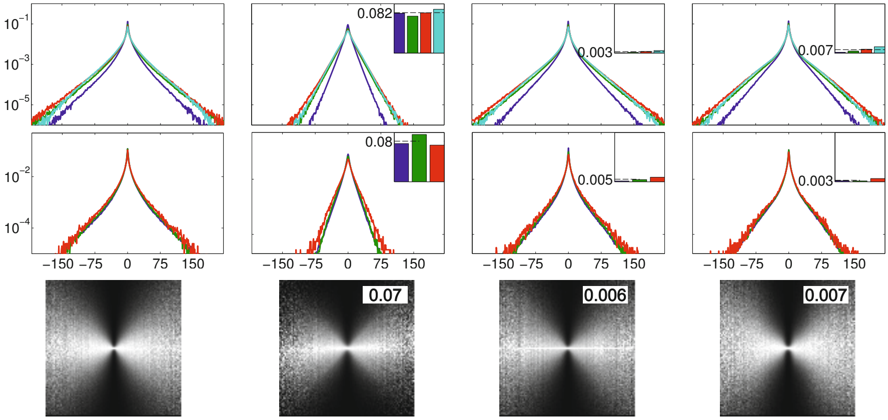
4 Experiments

Due to the intractability of the partition function, it is not possible to compare models through the likelihood of a test set. We follow [16,21] to evaluate whether other well-known properties of natural image statistics are captured by the learned models. We use a validation set of 1000 non-overlapping images of size 48×48, randomly cropped from a set of natural images. Since computing statistical properties such as marginals exactly is intractable as well, we use model samples from Gibbs sampling.

Evaluated Models. Since the 3×3 FoE of [16] represents image statistics more accurately than pairwise MRFs and, as far as we are aware, also other filter-based, high-order MRFs from the literature, we use it as performance baseline. We train the basic model from Sec. 2 using the improved learning procedure described in Sec. 3 on 1000 randomly cropped 48×48 natural image patches. To facilitate comparison, we trained a model with 8 filters of size 3×3 (Fig. 1c). To showcase the benefits of the improved learning scheme, we also trained 5×5 models with 8 and 16 filters. All models exhibit fully "active" filters and potentials with very broad shoulders and tight peaks. Due to limited space, we only show the learned 16 filters of the 5×5 model in Fig. 5.

4.1 Generative Properties

Model Features. Due to the maximum entropy interpretation of filter-based MRF priors [21], the learned model should perfectly capture its feature statistics if the potential

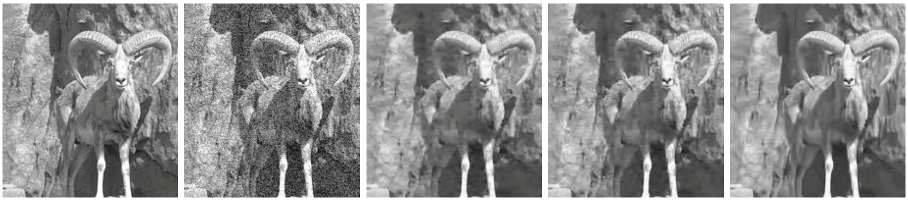


(a) Natural images (b) 3 × 3 FoE (8 filters) from [16] (c) 3 × 3 model (8 filters) w/ improved learning (d) 5 × 5 model (16 filters) w/ improved learning

Fig. 6. Random filter, multiscale derivative and conditional statistics: *(top)* Average marginal histograms of 8 random filters (mean 0, norm 1) of various sizes (3 × 3 – blue, 5 × 5 – green, 7 × 7 – red, 9 × 9 – cyan). *(middle)* Derivative statistics at three spatial scales (0 – blue, 1 – green, 2 – red; scales are powers of 2 with 0 being the original scale). *(bottom)* Conditional histograms of neighboring derivatives. Brightness corresponds to probability. (top right – KL-divergence.)

functions were chosen and learned appropriately. As can be seen from Fig. 1(a), the learning scheme in [16] does not lead to a particularly close match between model marginals and natural image statistics. In contrast, Fig. 1(c) shows that our improved learning procedure allows the identical model design to capture the marginal statistics of the in-built features very well. GSM potentials thus prove to be sufficiently flexible for modeling the potentials in such a filter-based MRF. This is also true for our 5 × 5 model (not shown); all marginal KL-divergences between model and image statistics are < 0.002. The resulting priors are thus real *maximum entropy models*.

Other Important Statistics. Since natural images exhibit heavy-tailed statistics even for the marginals of *random linear filters* [4], we evaluate our models in this regard with random filters of 4 different sizes (8 of each size). Fig. 6 (top) shows the average responses to these random filters for natural images, as well as all models. Moreover, natural images have been found to exhibit *scale invariant derivative statistics* [15]. Hence, we check the marginal statistics of derivatives at 3 image scales (powers of 2), which are shown in Fig. 6 (middle). Natural images have also been found to have characteristic *conditional distributions* of two image features, with a particular bow-tie shape [11]. Fig. 6 (bottom) shows the conditional histograms of neighboring image derivatives. While previous learning approaches come reasonably close regarding all three properties (Fig. 6b), our improved learning procedure reduces the mismatch between model and image statistics by a significant factor of 10, as measured in terms of the marginal KL-divergence (shown at the top-right corner of each plot). The 5 × 5 model improves particularly in terms of (multi-scale) derivative statistics (Fig. 6d). To the best of our knowledge, this is the first time that such close matches between model and natural image statistics have been reported for filter-based MRFs, or any MRF



(a) Original image (b) Noisy image (σ=25) (c) FoE from [16], $\mathcal{P}$ 26.27dB, $\mathcal{S}$ 0.689 (d) Our 5 × 5 model, $\mathcal{P}$ 26.46dB, $\mathcal{S}$ 0.700 (e) BM3D [1], $\mathcal{P}$ 26.23dB, $\mathcal{S}$ 0.674

Fig. 7. Image denoising example. $\mathcal{P}$: PSNR; $\mathcal{S}$: SSIM.

image prior. Importantly, this also demonstrates that filter-based MRFs are indeed capable of capturing a large number of key statistical properties of natural images.

4.2 Denoising Application

To further assess the impact of the proposed learning scheme, we evaluate the learned models using image denoising. Following [16], we estimate the posterior mean (MMSE estimate) using Gibbs sampling and evaluate on the same 68 test images with additive Gaussian noise of known variance. The runtime of our Matlab implementation is on par with [16] for the 3 × 3 model (8 filters), and approximately four times slower for the 5 × 5 one (16 filters). Tab. 1 shows a consistent, substantial boost in denoising performance from our improved learning procedure. Retraining the GSM potentials using PCD yields a gain of 0.15dB; the full learning procedure improves a 3 × 3 model with 8 filters by another 0.08dB. More importantly, the proposed learning scheme also allows training models with larger (*e.g.* 5 × 5) or more (*e.g.* 16) filters, both of which lead to further improvements in terms of denoising performance. In total, we obtain an improvement of 0.3dB over [16], which uses an identical model design but an inferior learning procedure, and even 0.8dB over the 5 × 5 FoE of [14]. This is not only a significant gain in the realm of denoising, but also makes our approach competitive with the latest state of the art in denoising. In particular, it can compete with BM3D (particularly in SSIM [18]) as well as NLSC [8]. As far as we are aware, this is the first time such competitive denoising performance has been achieved with any generative, global model of natural images. An example of denoising is shown in Fig. 7.

Table 1. Denoising results for 68 test images [14] ($\sigma = 25$)

Model/Method	∅ PSNR (dB)	∅ SSIM
5 × 5 FoE [14]	27.44	0.746
3 × 3 FoE [16]	27.95	0.788
8 fixed 3 × 3 filters from [16], learned potentials (proposed proc.)	28.10	0.793
8 learned 3 × 3 filters & learned potentials (proposed procedure)	28.18	0.796
8 learned 5 × 5 filters & learned potentials (proposed procedure)	28.22	0.797
16 learned 5 × 5 filters & learned potentials (proposed procedure)	28.26	0.799
BLS-GSM [11]	28.02	0.789
non-local sparse coding (NLSC) [8]	28.28	0.799
BM3D [1]	28.35	0.797

5 Conclusions

In this paper, we explored the limits of filter-based MRFs for natural image statistics. We identified various shortcomings in previous learning approaches, and proposed several improvements that increase robustness and enable learning larger, more, and more varied image filters. The learned potentials were found having an almost Dirac-delta like shape. Moreover, the proposed learning procedure strongly improves the models' ability of capturing the in-built feature statistics, making them real maximum-entropy models. They also show clear improvements in capturing other important statistical properties of natural images, outlining the capabilities of filter-based MRFs. Denoising experiments demonstrate significant performance gains, bringing the results very close to the state of the art.

Although our procedure allows learning more and larger filters, pushing this even further is currently not practical. Many filters lead to slower mixing, larger ones to less-sparse linear equation systems in sampling. Future work should aim to address this. Nonetheless, the trained models already capture natural image statistics very well, suggesting that further gains are likely challenging and may require new model designs.

References

1. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE T. Image Process. 16(8), 2080–2095 (2007)
2. Geman, D., Reynolds, G.: Constrained restoration and the recovery of discontinuities. IEEE T. Pattern Anal. Mach. Intell. 14(3), 367–383 (1992)
3. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. 14(8), 1771–1800 (2002)
4. Huang, J.: Statistics of Natural Images and Models. Ph.D. thesis, Brown University (2000)
5. Köster, U., Lindgren, J.T., Hyvärinen, A.: Estimating Markov Random Field Potentials for Natural Images. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A.K. (eds.) ICA 2009. LNCS, vol. 5441, pp. 515–522. Springer, Heidelberg (2009)
6. Levin, A., Nadler, B.: Natural image denoising: Optimality and inherent bounds. In: CVPR 2011 (2011)
7. Lyu, S., Simoncelli, E.P.: Modeling multiscale subbands of photographic images with fields of Gaussian scale mixtures. IEEE T. Pattern Anal. Mach. Intell. 31(4), 693–706 (2009)
8. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: ICCV 2009 (2009)
9. McAuley, J.J., Caetano, T., Smola, A.J., Franz, M.O.: Learning high-order MRF priors of color images. In: ICML 2006, pp. 617–624 (2006)
10. Norouzi, M., Ranjbar, M., Mori, G.: Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning. In: CVPR 2009 (2009)
11. Portilla, J., Strela, V., Wainwright, M.J., Simoncelli, E.P.: Image denoising using scale mixtures of Gaussians in the wavelet domain. IEEE T. Image Process. 12(11), 1338–1351 (2003)
12. Ranzato, M., Mnih, V., Hinton, G.E.: Generating more realistic images using gated MRF's. In: NIPS 2010 (2010)
13. Roth, S., Black, M.J.: On the spatial statistics of optical flow. Int. J. Comput. Vision 74(1), 33–50 (2007)
14. Roth, S., Black, M.J.: Fields of experts. Int. J. Comput. Vision 82(2), 205–229 (2009)

15. Ruderman, D.L.: The statistics of natural images. Network: Comp. Neural 5(4), 517–548 (1994)
16. Schmidt, U., Gao, Q., Roth, S.: A generative perspective on MRFs in low-level vision. In: CVPR 2010 (2010)
17. Tieleman, T.: Training restricted boltzmann machines using approximations to the likelihood gradient. In: ICML 2008 (2008)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE T. Image Process. 13(4), 600–612 (2004)
19. Weiss, Y., Freeman, W.T.: What makes a good model of natural images? In: CVPR 2007 (2007)
20. Welling, M., Hinton, G.E., Osindero, S.: Learning sparse topographic representations with products of Student-t distributions. In: NIPS 2002, pp. 1359–1366 (2002)
21. Zhu, S.C., Mumford, D.: Prior learning and Gibbs reaction-diffusion. IEEE T. Pattern Anal. Mach. Intell. 19(11), 1236–1250 (1997)