

Normalized filter pool for prior modeling of nature images

Yangang Wang¹ · Jinli Suo² · Qionghai Dai²

Received: 10 March 2015 / Revised: 18 January 2016 / Accepted: 20 January 2016 / Published online: 27 February 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Markov random field (MRF), as one of special undirected graphs, is widely used in modeling priors of natural images. Targeting to learn better prior models from a given database, we explore the natural image statistics at different scales and build normalized filter pool, a kind of high-order MRF, for prior learning of nature images. The main contribution of the proposed model is that we construct a multi-scale MRF model through constraining the norms of filters in kernel space and integrate all the filtering responses in a unified framework. We formulate both learning and inference as constrained optimization problems and solve them using augmented Lagrange method. The experiment results demonstrate that the normalization of filters at different scales helps to achieve fast convergence in learning stage and obtain superior performance in image restoration, e.g., image denoising and image inpainting.

Keywords High-order Markov random field · Multi-scale · Normalized filter pool · Image denoising · Image inpainting

1 Introduction

Over the past decades, Markov random fields (MRFs) have become a popular framework for natural image model-

ing due to its effectiveness in describing image statistics and successful applications, e.g., super-resolution [6,32], optical flow [1,25], image restoration [7,9,12,37], reconstruction [4,14,23] etc. The main issue of MRFs is to choose appropriate cliques and potentials to describe the knowledge of neighboring interactions. The forms of cliques and potentials can be designed or learned from the contextual constraints in natural images [27].

The mainstream of MRFs image models is trying to fit the heavy-tailed marginal distribution of natural images [31]. Historically, simple neighborhood structures (i.e., pairwise MRFs) are broadly used in many low-level vision tasks [4,8] because of their simplicity and low computational cost [16]. Numerous efforts have been made on considering more complex statistical dependencies in natural images and building high-order MRFs [12,37]. Roth and Black [27] compare the performance under different number and shape settings of maximum clique and validate the superior performance of high-order cliques compared to the pairwise ones. Meanwhile, the potentials have also developed from Gaussian function, which is used in early studies of MRFs [8], to complex ones, e.g., generalized Laplacian [2,13,28], Gaussian scale mixtures (GSMs) [31,35]. Schmidt et al. [31] evaluate the quality of image priors captured by various potentials and develop a flexible one by viewing MRFs in generative aspects. In sum, *high-order, non-Gaussian* types gain momentum over *pairwise, Gaussian* ones. It is worth noting that although the above approaches make great efforts on image prior modeling, many intrinsic characteristics of natural images are not well covered; this motivates researchers to address specific image characteristics in model building.

One consideration is that natural images exhibit spatial variations in orientations, Some researchers propose to use spatially adaptive potentials in MRFs, e.g., Lyu et al. [18], Roth and Black [26] introduce predefined derivative tensors

✉ Yangang Wang
ygwang.thu@gmail.com; yangangw@microsoft.com
Jinli Suo
jlsuo@tsinghua.edu.cn
Qionghai Dai
qhdai@tsinghua.edu.cn

¹ Microsoft Research, Beijing, China

² Tsinghua University, Beijing, China

to steer the filters towards the orientations of local image structures. In this way, they could adaptively adjust the potentials according to the structure of specific images, while at the expense of largely increased complexity.

Another important property of natural images is the cross scale variation, i.e., scaling an image would display different structures [30,34] and a multi-scale scheme tends to extract more information under a given training database. So far, few previous work explores the multi-scale MRFs models of natural images except for some empirical studies on making use of multi-scale information for specific tasks [3,22]. It should be noted that although natural images ensure the similarity at different scales, which is also known as scale invariance [30], there still exist statistical divergences among filter responses (see the 3rd column of Fig. 1); straightforward integration of the nature image information at different scales into a unified framework may not be an efficient way, which is demonstrated in the experiment section of this paper. This paper targets to explore the scheme to integrate the nature image information at different scales into a unified framework for performance boosting.

Our key idea to model the scale information of natural images is to construct a multi-scale high-order MRFs with a set of pyramid filters (i.e., different sizes of filters). Through constraining the norms of filter's response at different scales in kernel space, we build a filter pool to integrate the multi-scale nature image information. Different from the traditional way which describes the multi-scale information by a set of pyramid images, the proposed strategy would dramatically reduce the training data and training time. We learn the whole parameters of the proposed filter pool simultaneously from a given database. The learned multi-scale MRFs model can be applied for various tasks and we demonstrate its effectiveness via image denoising and image inpainting. The illustration of the proposed MRFs model is shown in Fig. 1.

2 Preliminaries

2.1 Why multi-scale is beneficial in MRFs?

Zhu and Mumford [37] have described that natural images exhibit information across a large entropy range, and an ideal prior model for natural images is supposed to provide a full description. Traditional MRFs are usually with single fixed-sized filter and the model parameters are learned from a large training database, which is composed of image patches at different entropy rates (as shown in Fig. 2a). In order to improve the flexibility of MRFs, Roth and Black [27] proposed a typical MRF model, named Fields of Experts (FoE), to reflect the key characteristics of natural images with a bank of fixed-sized filters.

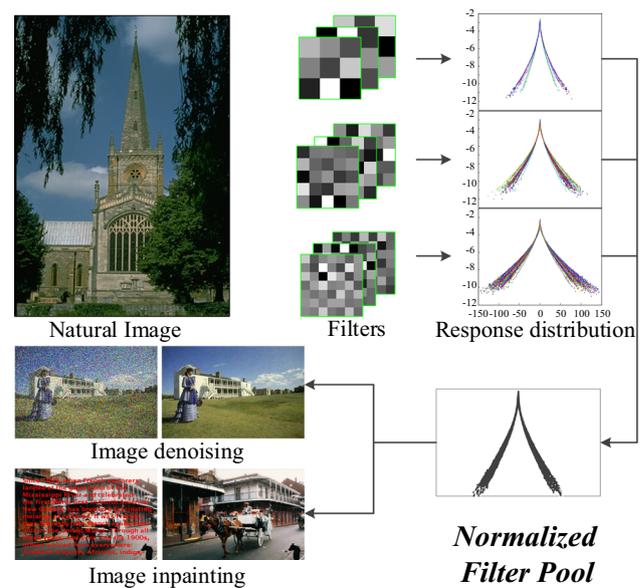


Fig. 1 Framework of normalized filter pool and its applications

However, it is impossible to arrive at the true statistics of natural images with fixed-sized filters (i.e., single scale) under a limited database; this motivates us to improve MRFs model alternatively. Following the scale space theory [34], there exist perceptual transitions in image pyramids where we can extract various information from different pyramid levels or different cropped-sized sub-images. Fig. 2 gives an intuitive example; comparing (a) and (b) one can see that the different-sized patches cropped from the same image (with a small entropy range) apparently reveal similar statistics with those cropped from multiple images at various entropy rates. Inspired from above analysis and observations, we aim at building a multi-scale MRF model to extract more information than single-scale models.

In addition, studies on several applications using MRF models (e.g., [3,17,22]) show that considering images in the scale space would improve the performance. This reveals that we may improve the MRFs by considering different scales in modeling. In this paper, we use different sizes of filters to build a high-order MRF to model the multi-scale information of nature images.

2.2 How to integrate multi-scale information?

Although multi-scale scheme in MRFs tends to extract richer prior knowledge, the integration is nontrivial since either the filter or filter responses may exhibit large cross-scale variances. In this section, we will explore the underlying problems and solutions in multi-scale integration.

Denoting x as an image, f as a zero-mean filter and $*$ as convolution, the variance of the filter response is

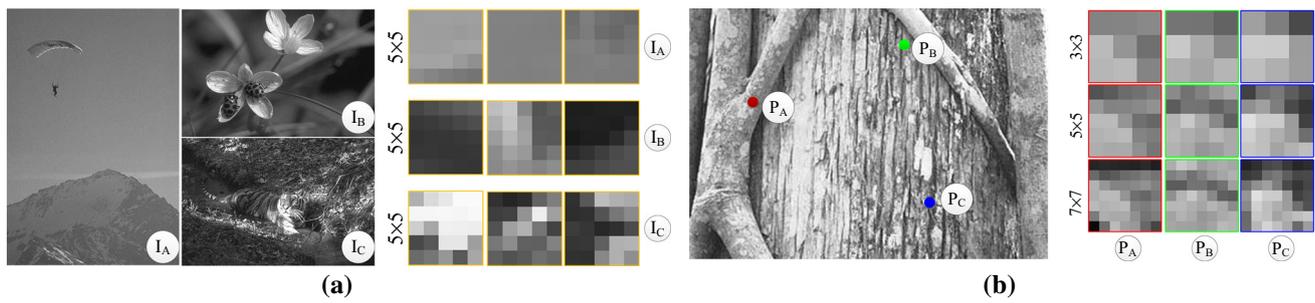


Fig. 2 Motivation of multi-scale MRFs. **a** Traditional single-scale MRFs learn prior models from images at different entropy rates. **b** Proposed multi-scale MRF extracts similar statistics as in **a** from a single image

$$\sigma_{x*f}^2 = \frac{1}{n}(x * f)^2, \tag{1}$$

where n is the number of pixels in x . Note that we assume the mean of $x * f$ is zero because of the zero-mean filter and the statistically similar color values in a small region. For convenience, we regard convolution as the multiplication of clique matrix A_x and column vector F , which are computed from x and f respectively, and yield

$$\sigma_{x*f}^2 = \frac{1}{n}(A_x F)^T(A_x F). \tag{2}$$

After performing a singular value decomposition to $\frac{1}{n}(A_x^T A_x)$ as $\frac{1}{n}(A_x^T A_x) = Q \Lambda Q^T$, we get

$$\sigma_{x*f}^2 = \|\Lambda^{\frac{1}{2}} Q^T F\|_2^2. \tag{3}$$

For simplicity, we define

$$B = \Lambda^{\frac{1}{2}} Q^T. \tag{4}$$

Here, B is just the scaled principal components of clique matrix A_x , with the scaling factor being $\Lambda^{\frac{1}{2}}$. Experiments have shown that these principal components of small patches are statistically stable across different natural images. In implementation, we pursue B from a large training database for prior learning.

Equation (3) tells us that the variance of filtering responses with x by any filter $B^{-1}H$ (i.e., $F = B^{-1}H$) is $\|H\|_2^2$. Since variance is the main factor affecting the shape of a heavy-tailed distribution, we can obtain approximately similar responses via constraining the l_2 norm of H . Note that the constraint on filter norm has been used by Köster et al. [15], but in different ways and for different purpose.

Besides, perhaps the most interesting result in Eq. (3) is that we can control the shape of filtering responses by constraining the norm of H without considering the filter size. Therefore, although there may exist large variances among cross-scale filter size and filtering responses, we can con-

trol $\|H\|_2$ to ensure similar potentials at different scales and integrate them in a unified framework.

3 Normalized filter pool

The purpose of image prior modeling is to dig out the statistics (joint probabilities) in images. MRFs provide a concise framework to build the joint probabilities, which describe an image with graphical representations $G = (V, E)$ with V denoting the pixels and E representing the connections between neighboring pixels. According to Hammersley–Clifford theorem, we could factorize the joint probability p for all pixels in image x as follows:

$$p(x) = \frac{1}{\mathbf{Z}} \prod_{c \in \mathcal{C}} V_c(x_c), \tag{5}$$

where x_c indicates a specific clique; V_c is the potential defined over x_c ; \mathcal{C} is the set of all cliques, and \mathbf{Z} is a normalizing factor. For simplicity, MRF is usually homogenous, i.e., V_c is the same for all cliques in graph G [26]. As a reminder, every V_c needs to map the clique x_c to a real positive number. There are lots of work that define different mechanism of mapping. The common and influential one uses two neighboring nodes or pixels and compute the potentials as $V_{i,j}(|x_i - x_j|)$. One particular work is the FRAME model proposed by Zhu and Mumford [37]. The FRAME model utilizes the filters to describe the clique information in graph, which facilitates the parameters learning from database. The filters with the same shape of cliques are designed to describe the contextual correlation among the signal elements and the joint probability of all pixels in x could be computed as

$$p(x; \Theta) = \frac{1}{\mathbf{Z}(\Theta)} \prod_{c \in \mathcal{C}} \phi((x * f)_c; \alpha). \tag{6}$$

Here, Θ is the collection of all parameters; $\mathbf{Z}(\Theta)$ is a normalizing factor. Note that we could use a heavy-tailed function $\phi(\cdot; \alpha)$ (e.g., $\phi(x) = e^{-|x|^\alpha}$) instead of delta function to penalize large variations within a local region. Inspired by

the PoE model [10], Roth and Black [27] propose the FoE model which integrates K aforementioned models and can capture the statistics of natural images more reliably. Thus, Eq. (5) can be extended as follows:

$$p(x; \Theta) = \frac{1}{\mathbf{Z}(\Theta)} \prod_{c \in \mathcal{C}} \prod_{i=1}^K \phi((x * f_i)_c; \alpha_i). \tag{7}$$

In this paper, we extend the above ideas further, i.e., learning statistics of natural images across different scales. Since the shape of filtering response distribution is decided by the l_2 norm of the filters multiplied by B^{-1} (cf. Sect. 2.2), we constrain the filter norms to ensure similar responses at different scales, and then the factoring rule in traditional MRFs can be extended across scales to construct a multi-scale integrated high-order MRF prior model.

The statistical characteristics of natural images [11,29] reveal that the filtering responses of natural images display similar responses, so we constrain the l_2 norm of filters in space B within a range. (We will discuss the parameter settings in the experimentation section.)

Considering S different scales, we denote the number of filters at scale $s \in S$ as K_s , f_i^s as the i th filter, and F_i^s as its corresponding column vector. Eq. (5) can be further refined in an integrated manner (see Fig. 3):

$$p(x; \Theta) = \frac{1}{\mathbf{Z}(\Theta)} \prod_{s \in S} \prod_{c \in \mathcal{C}_s} \prod_{i=1}^{K_s} \phi((x * f_i^s)_c; \alpha_i^s) \tag{8}$$

s.t. $F_i^s \mathbf{1} = 0, \quad d_1 \leq \|B^s F_i^s\|_2 \leq d_2.$

Here, $\phi(\cdot; \alpha_i^s)$ is a heavy-tailed function (we use student- T as in [27]); Θ is a collection of the parameters $\{f_i^s, \alpha_i^s | i = 1, 2, \dots, K_s; s \in S\}$; \mathcal{C}_s is clique set at scale s ; $\mathbf{Z}(\Theta)$ is a normalizing factor; B^s is a transform matrix at scale s ; $\mathbf{1}$ is an all-one vector to ensure the zero-mean proper-

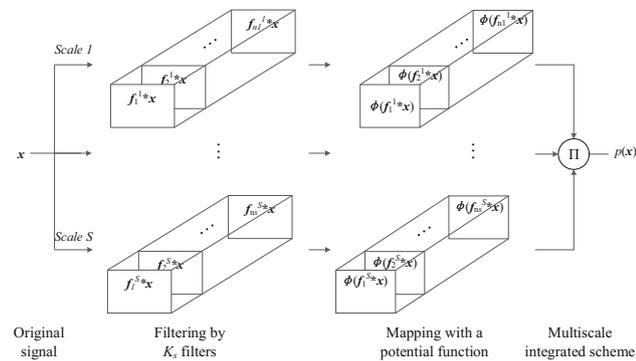


Fig. 3 The flow chart of proposed NFP. We use S different sized filters to convolve with image x , map all the filtering results with $\phi(\cdot)$, and multiply them together to construct the joint probability

ties of filters as in [11,28]; $[d_1, d_2]$ is the range for filter norms.

Since we normalize the filter norms for building a multi-scale MRF, we name the prior model defined in Eq. (8) as *normalized filter pool (NFP)*, whose energy is

$$E(x; \Theta) = - \sum_{s \in S} \sum_{c \in \mathcal{C}_s} \sum_{i=1}^{K_s} \log \phi((x * f_i^s)_c; \alpha_i^s). \tag{9}$$

4 Learning and inference

In this section, we will explain the learning and inference algorithms respectively.

4.1 Constrained parameters learning

We denote the training data set as X , which includes M images x_1, x_2, \dots, x_M . Because NFP is a model with constraints on parameter set Θ , we adopt the augmented Lagrange method and change the log-likelihood function as the following equation:

$$L(X; \Theta) = \sum_{i=1}^M \log p(x_i; \Theta) + \sum_{j=1}^J \lambda_j c_j(\Theta) + \sum_{j=1}^J \frac{1}{2\mu_j} c_j^2(\Theta). \tag{10}$$

Here, $c_j(\cdot)$ is j th constraint, which can be an equality constraint or inequality constraint [21]; J is the number of constraints ($J = 3$ in this paper); λ and μ are Lagrange multipliers and penalty parameters, respectively.

We search for the parameter Θ that maximizes the above objective function over the training data set. In initialization stage, we use linear transformation (multiplied by B^{-1}) and normalization to the samples from Gaussian distribution, as described in Step (3) and (4) of Algorithm 1.

After the initialization, we adopt the gradient descent method to optimize the objective function. The gradient of Eq. (10) with respect to Θ is computed as

$$\nabla_{\Theta} L(X, \Theta) = \sum_{i=1}^M \nabla_{\Theta} \log p(x_i; \Theta) + \sum_{j=1}^J \left(\lambda_j - \frac{c_j(\Theta)}{\mu_j} \right) \nabla_{\Theta} c_j(\Theta). \tag{11}$$

Algorithm 1 Parameters learning of NFP

Input:

- X : training data;
- S : number of scales;
- K_s : number of filter f_i^s in scale s ;
- $\phi(\cdot; \alpha)$: form of potential function;

Output:

$\langle f_i^s, \alpha_i^s \rangle: i = 1, 2, \dots, n_s; s = 1, 2, \dots, S;$

1. Compute clique matrix A^s with database X ;
2. Compute matrix $(B^s)^{-1}$ as shown in Eq. (4);
3. Generate K_s random \tilde{f}_i^s , which are i.i.d. from Gaussian distribution with zero mean and given variance;
4. Initialize the filters as $f_i^s = (B^s)^{-1} \tilde{f}_i^s$;
5. Repeat the following steps until $\|\nabla_{\Theta} L(X; \Theta)\|_2 < \epsilon$;
 - 5.1. Update λ_j and $\mu_j, j = 1, 2, 3$;
 - 5.2. Sample a new clique matrix A^s with CD;
 - 5.3. Compute $\nabla_{f_i^s} L(X; \Theta)$, and update f_i^s ;
 - 5.4. Compute $\nabla_{\alpha_i^s} L(X; \Theta)$, and update α_i^s ;

The first term of Eq. (11) is computed as [33]

$$\sum_{i=1}^M \nabla_{\Theta} \log p(x_i; \Theta) = M \cdot [-\langle \nabla_{\Theta} E(x; \Theta) \rangle_X + \langle \nabla_{\Theta} E(x; \Theta) \rangle_p]. \tag{12}$$

Here $\langle \cdot \rangle_X$ and $\langle \cdot \rangle_p$ are the expectations with respect to empirical distribution of X and model distribution p , respectively. It is very efficient to adopt Contrastive Divergence (CD) algorithm [33] to compute Eq. (12).

The second term computes the gradient of constraints $c_j(\Theta)$. Taking derivatives of Eq. (8), we have

$$\begin{aligned} \nabla_{f_i^s} c_1 &= \mathbf{1}, \\ \nabla_{f_i^s} c_{2,3} &= (B^s)^T B^s F_i^s. \end{aligned} \tag{13}$$

The Lagrange parameters λ_j and penalty parameters μ_j can be updated according to the following equalities [21]:

$$\begin{aligned} \lambda_j^{(k+1)} &= \lambda_j^{(k)} - c_j(\Theta^{(k)})/u_j^{(k)}, \\ \mu_j^{(k+1)} &= \eta \mu_j^{(k)}. \end{aligned} \tag{14}$$

where k indexes the iterations and η denotes an adaptive learning rate, $0 < \eta \leq 1$.

For clarity, we summarize the steps of parameter learning in Algorithm 1.

4.2 Constrained inference using priors

For more flexible usage, we formulate the inference as a constrained optimization instead of the widely used Maximum

a Posterior (MAP) method:

$$\begin{aligned} x^* &= \arg \max_x \log p(x) \\ \text{s.t. } d_j(x) &= 0, \quad j = 1, 2, \dots, J. \end{aligned} \tag{15}$$

Here, $d_j(\cdot)$ is the j th constraint function w.r.t. x and J is the number of constraints.

Removing the constraints via augmented Lagrange method, Eq. (15) can be rewritten as follows:

$$F(x) = \log p(x) + \sum_{j=1}^J \lambda_j d_j(x) + \sum_{j=1}^J \frac{1}{2\mu_j} d_j^2(x). \tag{16}$$

We also utilize gradient ascent method to find the optimal x . The gradient of $F(x)$ w.r.t. x is computed as

$$\nabla_x F(x) = \nabla_x \log p(x) + \sum_{j=1}^J \left(\lambda_j - \frac{d_j(x)}{\mu_j} \right) \nabla_x d_j(x). \tag{17}$$

Similar to [37], we compute the first term of Eq. (17) as

$$\nabla_x \log p(x) = \sum_{s=1}^S \sum_{i=1}^n \tilde{f}_i^s * \psi(f_i^s * x; \alpha_i^s), \tag{18}$$

where $\psi(\cdot)$ is the gradient of $\log \phi(\cdot)$ w.r.t. its parameter; \tilde{f}_i^s is mirrored by f_i^s .

The second term in Eq. (17) depends on specific inference tasks. For example, in image denoising with noise being zero-mean and variance being σ^2 , $d(x)$ takes the following form:

$$d_1(x) : \overline{x - x^*} = 0, \tag{19}$$

$$d_2(x) : \frac{1}{n} \sqrt{\text{tr}((x - x^*)^T (x - x^*))} = \sigma^2. \tag{20}$$

In inpainting, $d(x)$ is null, we only compute the pixels in unknown regions via optimization.

The Lagrange multipliers and penalty parameters are updated similar as Eq. (14).

For clarity, we summarize the algorithm of inference in Algorithm 2.

5 Experiments

5.1 Model analysis

The central issue of the proposed NFP is the normalization to multi-scale filters and the corresponding constrained optimization algorithm. In this section, we design experiments to analyze the subsequent advantages quantitatively.

Algorithm 2 Inference with the learned priors

Input:

x_0 : input data;
 f_i^s : learned filters;
 $\phi(\cdot; \alpha_i^s)$: potential function;

Output:

x : Result;

1. Initialize the result: $x = x_0$;
2. Initialize the Lagrange multipliers and penalty factors: $\lambda = 1, \mu = 1$;
3. Repeat the following steps until $\|\nabla_x F(x)\|_2 < \epsilon$;
 - 3.1. Update λ_j and μ_j ;
 - 3.2. Compute $\nabla_x F(x)$ according to Eq. (17);
 - 3.3. Update x ;

Taking image denoising as an example, we use Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) as performance criteria, and compare the performance of our algorithm under different constraint settings (with and without constraint) to measure the contribution from the introduced normalization. For a closer look, we also compare the performance between initial parameters and learned parameters.

Single scale experiment. We first test the denoising performance of initial filters with and without constraints at a single scale 3×3 . In non-constraint setting, we get the denoising performance with the filters initialized by the samples drawn from Gaussian distribution with different variances(cf. Algorithm 1), as shown in horizontal axis of Fig. 4. In constrained setting, we crop 20,000 patches (15×15 sub-images) randomly selected from the Berkeley image segmentation database [19] to compute matrix B^{-1} as shown in Algorithm 1 and normalize all the filters to a fixed norm in space B^{-1} (i.e., \tilde{f}_i^s). For objective comparison, we initialize all α_i^s to be 0.05 and adopt the same sampler Hybrid Monte Carol (HMC) in both models. In implementation, we run the sampling code downloaded from website [20] under the following parameter settings: each HMC step consists of 30 leaps, the leap-frog step size is initialized with 0.01, and adjusted adaptively to keep the acceptance range between 90 and 98 %.

Then we test the denoising performance on 68 test images from Berkeley database (transformed to gray scale) with manually added zero-mean noise ($\sigma = 25$), and the result is shown in Fig. 4. One can see that the performance of filters drawn from a unit Gaussian distribution and without normalization (the method in [27]) is much inferior to that of the normalized ones, while *normalizing the filters or constraining the norm of the filters in space B^{-1} would help provide good initialization.*

Further, we compare the denoising performance of non-constraint algorithm [27] and constrained algorithm under

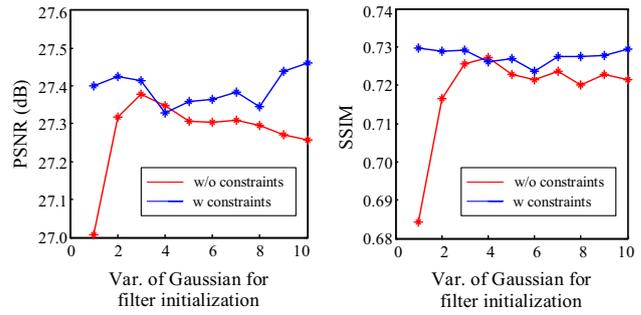


Fig. 4 Performance comparison between normalized and un-normalized filter initialization

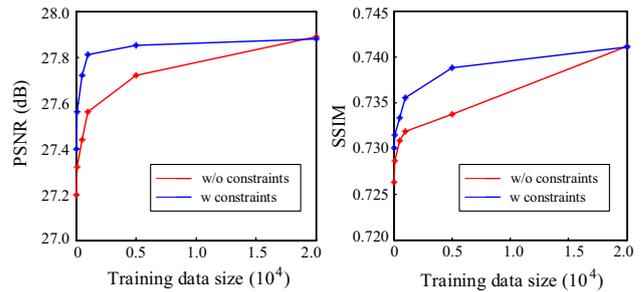


Fig. 5 Performance comparison between normalized and un-normalized filters learned from the same training data

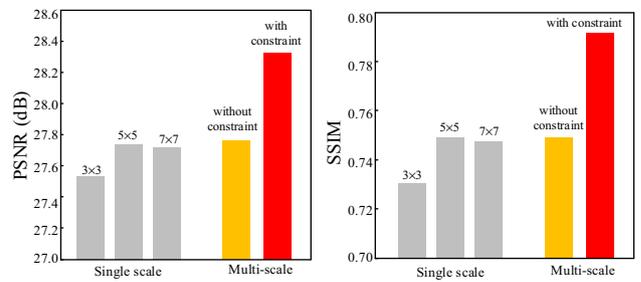


Fig. 6 Performance comparison of learned normalized and un-normalized filters under multi-scale settings

single-scale settings. The result in Fig. 5 shows that adding constraints brings little improvement on a large training database including 20,000 patches (the maximum number in previous work [27,36] within unacceptable training time). However, when we reduce the number of training data and find that our constrained algorithm gives much superior performance to the non-constraint one: we obtain the peak performance with 25 % of the training data necessary for non-constraint method, corresponding running time also decreases linearly. *This indicates that adding constraints within a single-scale MRF would largely reduce the necessary training data and the training time.*

Multi-scale experiment. In this experiment, we only use 10 % of the patches used in single-scale experiment (i.e., 2,000) to learn our multi-scale MRF model, and drop the constraint to get non-constraint result. We use three different cliques sizes

Table 1 Denoising results for 4 test images [24]

σ /PSNR	Barbara	Boats	Barco	House
05/34.15	37.20	36.35	38.43	38.46
10/28.13	32.92	33.24	34.49	35.07
15/24.61	30.39	31.41	32.36	33.46
20/22.11	28.68	30.14	30.94	32.16
25/20.17	27.40	29.05	29.98	31.26

Table 2 Performance comparison of image denoising results in 68 test images with $\sigma = 25$ [31]. Note that our result outperforms the state-of-the-art

Model	Inference	Avg. PSNR (dB)
Buades et al. [5]	MMSE	27.50
Roth et al. [27]	MAP (w/λ)	27.74
Schmidt et al. [31]	MMSE	27.95
Portilla et al. [24]	MMSE	28.02
Ours	Constrained	28.32

(3×3 , 5×5 , and 7×7) and remove one filter indicating the mean value at every scale to learn 80 filters (i.e., 8 filters at scale 3×3 , 24 ones at scale 5×5 and 48 ones at scale 7×7). All the filters are initialized by the samples drawn from a Gaussian distribution with fixed variance (variance is 4 in implementation and result is insensitive to the parameter setting) and then normalized in space B^{-1} . Initialization of α_i^s s and the sampling are the same with that in single scale experiment.

From the results in Fig. 6 one can clearly see that the added constraints prominently improve the integrated performance than non-constraint fusion. This indicates that there indeed exist large differences among filter responses at different scales (their large difference can also be seen from Fig. 1) and validates the normalization operation.

It is worth noting that the proposed multi-scale MRF model (NFP) can be learned from a much smaller training data under a given performance; the learning time does not increase largely compared with learning a single-scale FoE. The running time for multi-scale denoising is about twice of that in single scale.

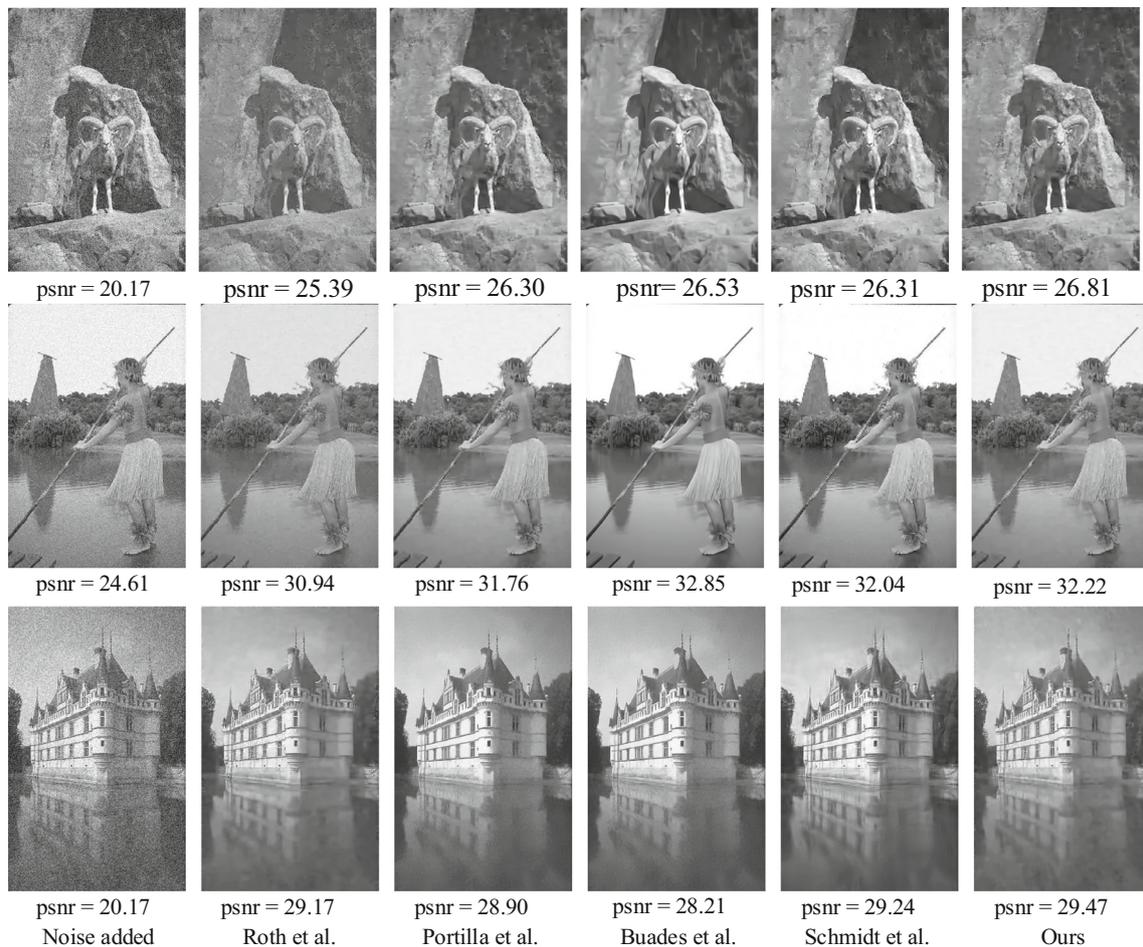


Fig. 7 Some image denoising results and the comparison with state-of-the-art approaches

5.2 Applications in image restoration

In order to validate the effectiveness of the proposed multi-scale MRF model further, we conduct a series of experiments in two typical applications: image denoising and image inpainting.

Image denoising experiment. In this experiment, we add noise manually under given standard deviation (either Gaussian or non-Gaussian) and use the learned model for denoising using Algorithm 2.

A. Denoising and performance evaluation

We apply the learned NFP for image denoising on two databases: Table 1 shows the denoising results on four widely used test images for denoising [24] and Table 2 shows the average performance on 68 test images from Berkeley image segmentation database [19]. Figure 7 shows an exemplar denoising result in parallel with some previously published work.

We compare the performance with that of state-of-the-art algorithms, as shown in Fig. 7 and Table 2, from which we can see that the proposed model outperforms the previous work in both visual results and quantitative evaluation. The promising performance may be due to the fact that we normalize the filters across image scales before integrating them together and help to achieve a more reasonable prior.

B. Effects of model parameters

In this experiment, we test the effects from two key parameters on denoising performance: clique size and l_2 norm of filters.

Figure 8 shows the denoising results with different clique sizes and the one using integration scheme. Noticeable differ-

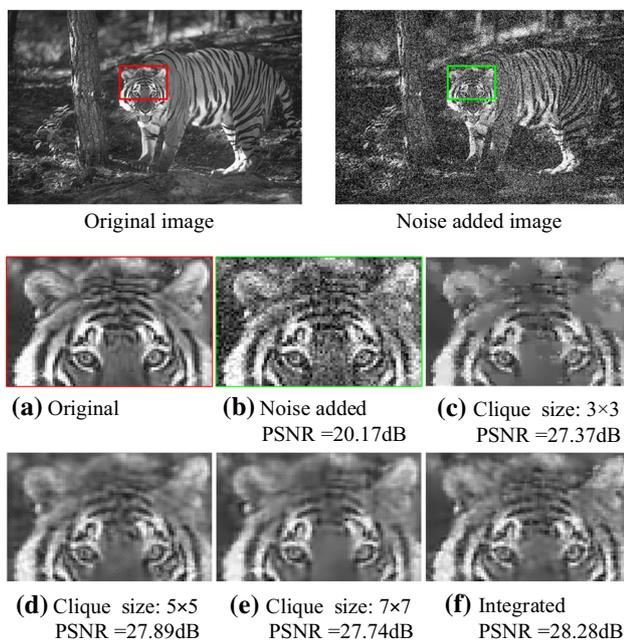


Fig. 8 Image denoising result with respect to the clique size

ences exist in the denoising results at different clique settings, among which performance of clique 5×5 is superior to that at adjacent scales. However, the integration scheme gains another 0.39 dB in PSNR, which validates the effectiveness of the proposed filter pool.

We also analyze the effects of the parameter $\|f\|_2$ on denoising performance. From the curve obtained on 3×3 cliques in Fig. 9, we can see that the performance stays relative stable within [6, 30], while degenerating largely once out of this range. Experiments on other clique settings give similar trends, which effectively validate the added constraints in our prior model.

Image inpainting experiment. We also apply the learned image priors to image inpainting. Figure 10 shows our inpainted results of the cropped ‘three children photo’ together with the ones implemented by Roth et al. [27] and Schmidt et al. [31]. For clarity, a detailed region is shown on the right bottom of each result. Although there is no quantitative measurement for inpainting performance, the visual

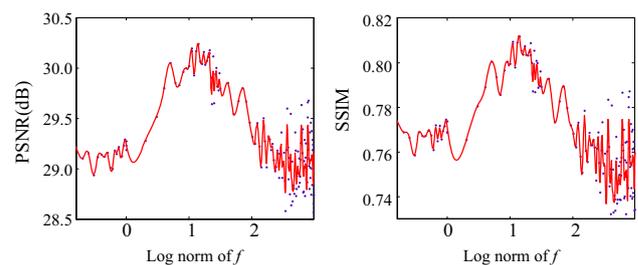


Fig. 9 Average PSNR of image denoising with respect to $\log(\|f\|_2)$

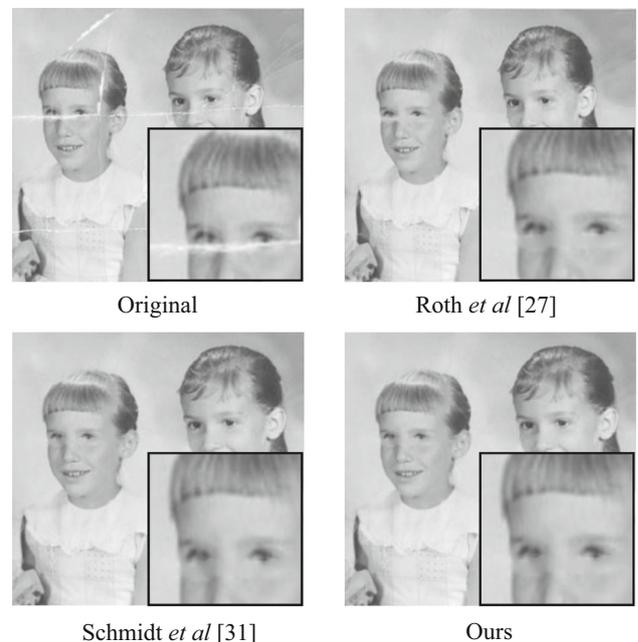


Fig. 10 Results of image inpainting

results reveal that our model is able to obtain inpainting results comparable or superior to those of state-of-the-art.

6 Conclusions and future work

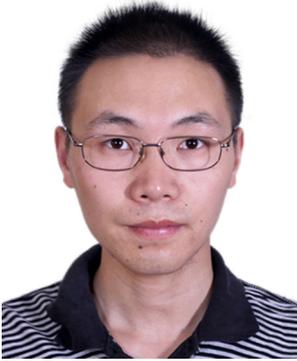
In this paper, we normalize the image statistics at different scales and build a normalized filter pool for modeling image priors; the normalization enables to integrate multi-scale information in a unified manner to learn a high-order MRF model. The proposed model obtains superior performance in both image denoising and inpainting tasks.

In the future, we would like to extend the proposed model in more flexible ways, e.g., design clique shape according to segmentation results instead of uniform grids. Applying the approach to other tasks (e.g., video, voice) is also a worth considering direction.

Acknowledgments We would like thank all the reviewers. This work was supported in part by NSFC (No. 61305026).

References

- Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. *Int. J. Comput. Vis.* **92**(1), 1–31 (2011)
- Black, M.J., Rangarajan, A.: On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Int. J. Comput. Vis.* **19**(1), 57–92 (1996)
- Blunsden, S., Atallah, L.: Investigating the effects of scale in MRF texture classification. In: *VIE* (2005)
- Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(11), 1222–1239 (2001)
- Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2005)
- Chen, J., Nunez-Yanez, J., Achim, A.: Video super-resolution using generalized gaussian Markov random fields. *IEEE Signal Process. Lett.* **19**(2), 63–66 (2012)
- Cho, T.S., Zitnick, C.L., Joshi, N., Kang, S.B., Szeliski, R., Freeman, W.T.: Image restoration by matching gradient distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(4), 683–694 (2012)
- Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. *Int. J. Comput. Vis.* **40**(1), 25–47 (2000)
- Geman, S., Geman, D.: Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEEE Trans Pattern Anal. Mach. Intell.* **9**(9), 721–741 (1984)
- Hinton, G.: Product of experts. In: *ICANN* (1999)
- Huang, J.: Statistics of Nature Images and Models. PhD thesis, Brown University (2000)
- Ishikawa, H.: Higher-order clique reduction in binary graph cut. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2009)
- Kohli, P., Kumar, M.P.: Energy minimization for linear envelope MRFs. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010)
- Komodakis, N., Paragios, N.: (2009) Beyond pairwise energies: efficient optimization for higher-order MRFs. In: *IEEE Conference on Computer Vision and Pattern Recognition*
- Köster, U., Lindgren, J., Hyvärinen, A.: Estimating Markov random field potentials for natural images. In: *ICA* (2009)
- Li, S.Z.: *Markov Random Field Modeling in Image Analysis*, 3rd edn. Springer, New York (2009)
- Liu, C., Pizer, S.M., Joshi, S.: A Markov random field approach to multi-scale shape analysis. In: *SSVM* (2003)
- Lyu, S., Simoncelli, E.P., Hughes, H.: Statistical modeling of images with fields of Gaussian scale mixtures. In: *NIPS* (2006)
- Martin, D., Fowlkes, C., Tal, D., Malik, J.: 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)
- Murphy, K.: <https://github.com/bayesnet/bnt> (2014). Accessed 17 Apr 2014
- Nocedal, J., Wright, S.J.: *Numerical Optimization*, 2nd edn. Springer, New York (2006)
- Paget, R., Longstaff, I.D.: Texture synthesis via a noncausal non-parametric multiscale Markov random field. *IEEE Trans. Image Process.* **7**(6), 925–931 (1998)
- Paulsen, R.R., Brentzen, J.A., Larsen, R.: Markov random field surface reconstruction. *IEEE Trans. Vis. Comput. Graph.* **16**, 636–646 (2010). doi:10.1109/TVCG.2009.208
- Portilla, J., Strela, V., Wainwright, M.J., Simoncelli, E.P.: Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Process.* **12**(11), 1338–1351 (2003)
- Roth, S., Black, M.J.: On the spatial statistics of optical flow. In: *International Conference on Computer Vision* (2005)
- Roth, S., Black, M.J.: Steerable random fields. In: *International Conference on Computer Vision* (2007)
- Roth, S., Black, M.J.: Fields of experts. *Int. J. Comput. Vis.* **82**(2), 205–229 (2009)
- Rother, C., Kohli, P., Feng, W., Jia, J.: (2009) Minimizing sparse higher order energy functions of discrete variables. In: *IEEE Conference on Computer Vision and Pattern Recognition*
- Ruderman, D.L.: The statistics of natural images. *Netw. Comput. Neural Syst.* **5**, 517–548 (1994)
- Ruderman, D.L.: Origins of scaling in natural images. *Vis. Res.* **37**(23), 3385–3398 (1997)
- Schmidt, U., Gao, Q., Roth, S.: A generative perspective on MRFs in low-level vision. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010)
- Tappen, M.F., Russell, B.C., Freeman, W.T.: Exploiting the sparse derivative prior for super-resolution and image demosaicing. In: *International Workshop on Statistical and Computational Theories of Vision at ICCV* (2003)
- Teh, Y.W., Welling, M., Osindero, S., Hinton, G.E.: Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.* **4**, 1235–1260 (2003)
- Wang, Y., Zhu, S.: Perceptual scale-space and its applications. *Int. J. Comput. Vis.* **80**(1), 143–165 (2008)
- Weiss, Y., Freeman, W.T.: What makes a good model of natural images? In: *IEEE Conference on Computer Vision and Pattern Recognition* (2007)
- Woodford, O.J., Reid, I.D., Torr, P.H.S., Fitzgibbon, A.W.: Fields of experts for image-based rendering. In: *British Conference on Machine and Computer Vision* (2006)
- Zhu, S.C., Mumford, D.: Prior learning and Gibbs reaction-diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(11), 1236–1250 (1997)



Yangang Wang received his B.E. degree from Southeast University, Nanjing, China, in 2009 and his Ph.D. degree in control theory and technology from Tsinghua University, Beijing, China, in 2014. He is currently an associate researcher at Microsoft Research Asia. His research interests include image processing, computer vision, computer graphics, motion capture and animation.



Jinli Suo is currently a lecturer in the Department of Automation at Tsinghua University, Beijing, China. She received her B.E. degree from Shandong University, Ji'nan, China, in 2004 and her Ph. D. degree from University of Chinese Academy of Sciences, Beijing, China, in 2010. Her research interests include computational photography, computer vision, pattern recognition and statistical learning.



Qionghai Dai is currently a professor of Department of Automation at Tsinghua University, Beijing, China. He is the director of the Broadband Network and Digital Media Laboratory. He also serves as the associated head of the Department of Automation. (Please consider rephrasing the following sentence: He has authored or co-authored over 250 high-quality papers, and published 4 books/book chapters). He has published over 250 papers, and 4 books. He has filed

70 patents and some of them are of great economic value. He obtained his Ph.D. degree in 1996 from Northeastern University, Shenyang, China.