Applied Informatics

RESEARCH

Open Access

Learning scene-aware image priors with high-order Markov random fields

Dong Gong, Yanning Zhang, Qingsen Yan and Haisen Li*

*Correspondence:
haisenli.nwpu@gmail.com
School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, Shaanxi, China

Abstract

Many methods have been proposed to learn image priors from natural images for the ill-posed image restoration tasks. However, many prior learning algorithms assume that a general prior distribution is suitable for over all kinds of images. Since the contents of the natural images and the corresponding low-level statistical characteristics vary from scene to scene, we argue that learning a universal generative prior for all natural images may be imperfect. Although the universal generative prior can remove artifacts and reserve a natural smoothness in image restoration, it also tends to introduce unreal flatness and clutter textures. To address this issue, in this paper, we present to learn a scene-aware image prior based on the high-order Markov random field (MRF) model (SA-MRF). With this model, we jointly learn a set of shared low-level features and different potentials for specific scene contents. In prediction, a good prior can be adapted to the given degenerated image with the scene content perception. Experimental results on the image denoising and inpainting tasks demonstrate the efficiency of the SA-MRF on both numerical evaluation and visual compression.

Keywords: Image restoration, Markov random fields, Scene-aware image prior learning

Introduction

Image restoration tasks, such as denoising (Tappen et al. 2007; Schmidt et al. 2010; Schmidt and Roth 2014), deblurring (Krishnan and Fergus 2009; Krishnan et al. 2011; Levin et al. 2009; Zhang et al. 2013; Gong et al. 2016, 2017) and super resolution (Tappen and Liu 2012) are all inherently ill-posed. Some knowledge of natural images is used as prior to boost the estimation stability and to recover information lost in non-ideal imaging processes. Recently, many image priors work on image gradients for briefness of modeling and better performance (Fergus et al. 2006; Levin et al. 2007, 2009; Krishnan et al. 2011; Krishnan and Fergus 2009; Xu et al. 2013; Zhang et al. 2013). However, the representation of image prior distribution in gradient domain is fragile for sophisticated concept of natural, as the variant of image content and/or scale makes the gradient characteristics unstable for modeling the unique clear individual images.

Motivated by the demand of capturing stable and accurate prior knowledge of natural image, many low-level modeling technologies including feature representation and related distribution are studied. Recent years have seen a trend to figure out this issue through the use of probabilistic graphical models (e.g., MRF and CRF) with non-Gaussian potential functions (Roth and Black 2005; Weiss and Freeman 2007; Samuel and Tappen 2009; Schmidt et al. 2010, 2014; Schmidt and Roth 2014; Chen et al. 2015), such as fields of experts (FoE) (Roth and Black 2005; Schmidt et al. 2010).

All of the manually designed priors and learned priors expect to model a universal distribution to represent all real-world natural images (in a specific discussed domain). Unfortunately, different images with different scene contents have varying statistics on usual low-level features like gradients or responses of learned filters in high-order MRF cliques (Fig. 1). Figure 1 shows that images with different contents (Left) have different responses on the gradient filter (Middle) or the learned high-order filters in Schmidt et al. (2010) (Right). Therefore, relying on universal generative image prior to recover every specific image is improper.

Considering the gap between the universal image prior and the special property of individual images, a series of content-related image priors are exploited in many image restoration tasks (Tappen et al. 2007; Cho et al. 2010; Sun et al. 2010; Schmidt and Roth 2014; McAuley et al. 2006). In Tappen et al. (2007) and Cho et al. (2010), local features are utilized to adapt the prior works on local areas in restoration tasks. However, as the local features like gradient filter responses (Tappen et al. 2007) and local texture (Cho et al. 2010) are usually not striking on weak edges or regions with ambiguous content, these local-specific models face inaccurate labeling problems, and the restoration results often suffer artifacts. In addition, the models in Tappen et al. (2007) and Schmidt and Roth (2014) can only be learned for specific state of the degeneration, which limits the range of application. The previous related works trying to approach the content-aware prior mainly focus on connecting the contents with some simple features such as statistics on gradients, since connecting the complex low-level features (e.g., any filter responses) with the high-level features representing the scene contents is more difficult. Additionally, recently, McAuley et al. (2006) proposed to the high-order MRF prior for

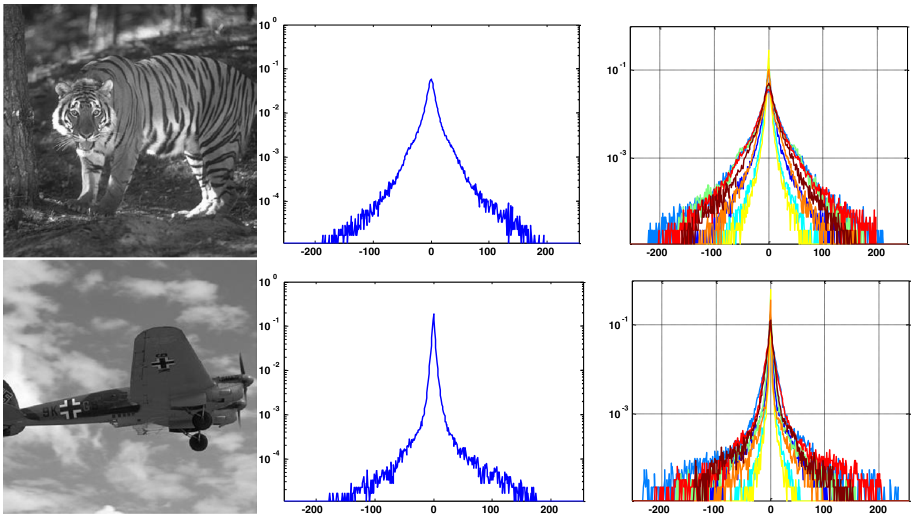

Fig. 1 Statistical characteristics change with the variations of image scene/content. Left: two typical real-world scenes. Middle: the empirical distributions of gradients in the images on left. Right: filter statistics of the images with the eight filters learned in Schmidt et al. (2010)

color images. In Feng et al. (2016), a high-order natural image prior model was proposed for reducing the Poisson noise. Ren et.al. (2013) introduced the "context-aware" concept into the sparse representation for image denoising and super-resolution. Considering the limitation of expression ability of the classical MRF, Wu et.al. proposed to compact the MRFs with deep neural networks (Wu et al. 2016).

In a natural image, low-level statistical characteristics are usually generated by the contents in the captured scene (Torralba and Oliva 2003). And the scene perception for an image is usually more robust than the pixel-level (low-level) characteristics. Based on this observation, we focus on developing a scene-aware prior model that can adapt the manifolds of the scene-related content in an image globally instead of taking the local structures. In this paper, we propose a scene-aware Markov random field (SA-MRF) model to capture the scene-discriminating statistical prior of any whole natural image; the SA-MRF model owes high-order non-Gaussian potential conditioned on a *scene coefficient* extracted from high-level concepts of observations. This is based on an assumption that the high-level contents are preserved fairly even in degenerated observations. Then related efficient algorithms for learning and inference are proposed. Experiments on image restoration tasks, denoising and inpainting, illustrate that the SA-MRF-based scene-aware image prior captures the image statistic characteristics accurately and improves the quality of images effectively.

Scene-aware image prior based on MRFs

The purpose of this paper is to build a system, in which (1) a high-order MRF model depending on scene content of the image is proposed to model the low-level statistical distribution and (2) the observed image can be adapted to a specific proper prior in restoration procedure. Overview of the system is illustrated in Fig. 2.

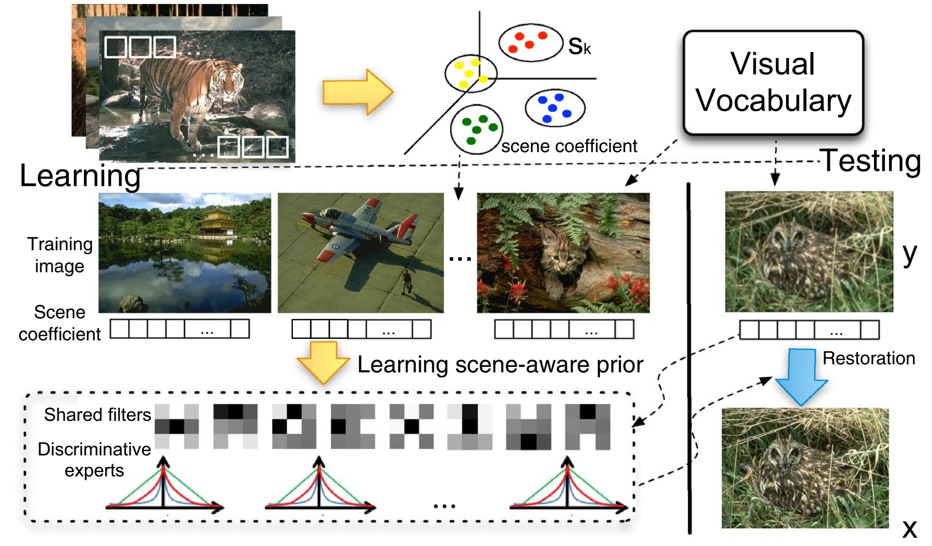

Fig. 2 Scene-aware prior learning and testing diagram. Scene coefficient is represented by bag-of-words, distribution of the bag-of-words descriptors provide scene discrimination in learning process and give scene-aware guides in testing

Scene-aware MRF model

The distribution of natural image $\mathbf{x}$ is formulated as a high-order MRF (Schmidt et al. 2010). To let the scene content information guide the modeling, we introduce an explicit *scene coefficient* as a parameter of the distribution.

Let $\{\mathbf{F}_i\}_{i=1}^N$ denote a set of filter-based features to capture the low-level characteristics of natural images, and $f(\mathbf{x})$ denote the scene content perceiving feature, which is defined as the *scene coefficient.* With the scene coefficient, the probability distribution of the corresponding clear image is defined as

$$p(\mathbf{x}; f(\mathbf{x}), \{\mathbf{F}_i\}, \mathbf{\Theta}) = \frac{1}{Z(\mathbf{\Theta})} \prod_{c \in \mathcal{C}} \prod_{i=1}^{N} \phi(\mathbf{F}_i^T \mathbf{x}_c; \mathbf{w}_i(f(\mathbf{x}), \boldsymbol{\theta})), \tag{1}$$

where $\mathcal{C}$ is the set of maximal cliques (Koller and Friedman 2009) of the MRFs; $\mathbf{x}_c$ thus denotes a subvector of $\mathbf{x}$ corresponding to the clique[1] c; $\phi(\cdot)$ represents the potential function; $\mathbf{w}_i$ is the parameter of the potential function $\phi(\cdot)$ which depends on $f(\mathbf{x})$ and the parameter $\boldsymbol{\theta}$ through a function $\mathbf{w}(\cdot)$[2]; $\mathbf{\Theta}$ is the collection of parameters $\mathbf{F}_i$'s and $\mathbf{w}_i$'s; and $Z(\cdot)$ denotes partition function normalizing the product of the potential functions (Koller and Friedman 2009). Because our MRF model (1) explicitly considers the scene content, we call it as *scene-aware* MRF. Model (1) is called scene-aware prior since the parameters of the prior distribution in (1), i.e., $\mathbf{w}_i$, depend on he scene coefficient $f(x)$, which captures the scene content of a specific image $\mathbf{x}$ in practice.

Potential function conditioned on scene coefficient

In (1), the formulation of the potential function is still not given. In this section, we will focus on the modeling of the potential function depending on the scene coefficient.

On account of the heavy-tailed filter statistics of natural image, we formulate the non-Gaussian potential function based on Gaussian scale mixtures (GSMs) models:

$$\phi(\mathbf{F}_i^T \mathbf{x}_c; \mathbf{w}_i(f(\mathbf{x}), \boldsymbol{\theta})) = \sum_{j=1}^{J} w_{ij}(f(\mathbf{x}), \boldsymbol{\theta}) \cdot \mathcal{N}(\mathbf{F}_i^T \mathbf{x}_c; 0, \sigma_i^2 / s_j), \tag{2}$$

where J is the number of mixture components and is set as 15 in this paper; w_{ij} is the j th component of the parameter vector $\mathbf{w}_i$; σ_i^2 and s_j denote the base variance and scale of Gaussian components, respectively. Following Schmidt et al. (2010), we set the scales as $s = \exp(-9, -7, -5, -4, ..., -1, 0, 1, ..., 4, 5, 7, 9)$. Benefiting from the mixture of Gaussian formulation in (2), the inference of the model can be simplified due to the conjugacy. From (2), the scene coefficient $f(\mathbf{x})$ influences the potential function through the weights of Gaussian components w_{ij}, and links the low-level characteristics and high-level properties associated with the contents in the scene. Then we will introduce how to build the linkage, and give the definitions of $\mathbf{w}(\cdot)$ and $f(\cdot)$.

[1] In model (1), if $\mathbf{F}_i$'s are $l \times l$ filters, each $\mathbf{x}_c$ is a $l \times l$ subvector in $\mathbf{x}$.

[2] We slightly abuse the notation $\{\mathbf{w}_i\}_{i=1}^N$ as both the parameters of the potentials and the functions $\mathbf{w}_i(f(\mathbf{x}), \boldsymbol{\theta})$.

Link the image x *and the scene perception through $f(\mathbf{x})$*

Given a $\mathbf{x}$, an easy way to represent its scene is to assign the discrete labels associated with the content (e.g., objects or scene) in $\mathbf{x}$ as many scene understanding works (Li et al. 2009). However, because there is a bias between the high-level perception of the content and the low-level feature [e.g., SIFT (Lowe 2004) and GIST (Oliva and Torralba 2001)] (Li et al. 2010), even images with same content labels may have dissimilar low-level statistical distributions. Instead of tackling this issue directly, we try to *take advantage of it*. Because our task roots in the low-level tasks, we do not need to assign exact labels to the contents in the scene. We directly use the Bag-of-words (BoW) histogram of SIFT descriptors to toward scene perception. Given an image $\mathbf{x}$, we extract dense SIFT (DSIFT) from it and generate DSIFT-BoW histogram with 200 vocabularies as $\mathbf{b}_{\mathbf{x}}$. $f(\mathbf{x})$ is defined as $f(\mathbf{x}) = \mathbf{b}_{\mathbf{x}}$. To extract dense SIFT, we run the SIFT feature extractor on a dense grid of location covering all locations on an image at a fixed scale and orientation. Specifically, in prediction task, given a $\mathbf{y}$, we first roughly recover a clear image $\hat{\mathbf{x}}(\mathbf{y})$. For example, for noisy observation $\mathbf{y}$, we do denoising via a simple Wiener filter (Sonka et al. 2014) or Gaussian low-pass filter. Then we extract DSIFT-BoW feature from $\hat{\mathbf{x}}(\mathbf{y})$ and let $\mathbf{b}_{\hat{\mathbf{x}}(\mathbf{y})}$ represent the corresponding feature of the latent clear image. The encoder of DSIFT-BoW is denoted as D. Note that, a clear image can be roughly recovered using some simple methods for extracting the BoW feature as the initialization. But it is not good enough to show many pixel-level details.

Link the scene coefficient $f(\mathbf{x})$ and low-level statistics through $\mathbf{w}(f(\mathbf{x}), \boldsymbol{\theta})$

Images containing similar scene contents usually follow similar low-level distributions. Based on this observation, for simpleness, we first assume that all images $\mathbf{x}$ can be clustered into K clusters w.r.t. $f(\mathbf{x})$. Accordingly, we assign the images to the clusters of which the centroids are the closest based on Euclidean distance. The images belonging to the k-th cluster is given a tag k and represented as $\mathbf{x}^k$. We assume that the training images are from a distribution of which the parameters are the combination of K sets of parameters $\{\mathbf{w}_i^k\}$. In learning process, we assume each $\{\mathbf{w}_i^k\}$ for all k can be learned by fitting the observations belonging to the k-th cluster, i.e., $\{\mathbf{x}_i^k\}$. We define the centroids of the each clusters as $\mathbf{x}'^k$. Following this, we approximate each $\mathbf{w}_i$ as a linear combination of K principle $\mathbf{w}_i^k$:

$$\mathbf{w}_i(f(\mathbf{x}); \boldsymbol{\theta}) = \sum_{k=1}^{K} \kappa_k(f(\mathbf{x}), f(\mathbf{x}'_k))\mathbf{w}_i^k \tag{3}$$

where $\kappa_k(\cdot, \cdot)$ is a similarity measurement of $f(\mathbf{x})$. We let $\kappa_k(\cdot, \cdot)$ be a simple Gaussian kernel:

$$\kappa(f(\mathbf{x}), f(\mathbf{x}'^k)) = \exp\left\{-\frac{1}{2}\left[\frac{\|f(\mathbf{x}) - f(\mathbf{x}'^k)\|_2^2}{\sigma_k^2}\right]\right\} \tag{4}$$

where σ_k^2 is the band width of the kernel regarding to the k-th cluster. Benefiting from this clustering-based representation, the distribution model (1) can be learned efficiently (see "Learning algorithm" section).

Learning algorithm

In this section, we will introduce an efficient learning algorithm that estimates model parameters from high-quality training samples, and inference algorithm for image restoration.

Given a set of training images, $\{\mathbf{x}_t\}_{t=1}^{T}$, the parameters of the model $\boldsymbol{\Theta}$ and low-level features $\{\mathbf{F}_i\}$ are estimated by maximizing the likelihood on the training data. We maximize the likelihood through minimizing the Kullback–Leibler divergence (KLD) between the model and empirical distribution of training data.

Substituting (2) into (1), the log-likelihood of observations is formulated as

$$\begin{aligned} E(\{\mathbf{x}_t\}, \{\mathbf{F}_i\}, \boldsymbol{\Theta}) &= -\sum_{t}^{T}\sum_{c\in\mathcal{C}}\sum_{i=1}^{N}\log(\phi(\mathbf{F}_i^T\mathbf{x}_{t,c}; \boldsymbol{\Theta})) \\ &= -\sum_{t}^{T}\sum_{c\in\mathcal{C}}\sum_{i=1}^{N}\log\left\{\sum_{j=1}^{J}\left[\sum_{k=1}^{K}\kappa(f(\mathbf{x}_t),f(\mathbf{x}'^k))w_{ij}^k\right]\mathcal{N}(\mathbf{F}_i^T\mathbf{x}_{t,c}; \frac{\sigma_i^2}{s_j})\right\}. \end{aligned} \tag{5}$$

Note that in our model, the filters $\mathbf{F}_i$'s are shared by all images in K different clusters, while the weights $\mathbf{w}_i^k$'s are different for K different clusters.

Relying on the clustering based definition of $\mathbf{w}_i$, the whole learning scheme comprises two steps: (1) calculating $\kappa(f(\mathbf{x}_t),f(\mathbf{x}'^k))$ for all images, and (2) estimating $\{\mathbf{F}_i\}$ and $\{\mathbf{w}_i^k\}$. Firstly, we extract dense SIFT features from $\{\mathbf{x}_t\}$, build the encoder dictionary D, and generate DSIFT-BoW features $\mathbf{b}_\mathbf{x}$ for all images. Secondly, we cluster images w.r.t. $\{\mathbf{b}_{\mathbf{x}_t}\}$ using Gaussian mixture model (GMM). We let $\mathbf{b}^k$ denote the centers of k-th clusters, and σ_k^2 in (4) be the average of the diagonal of the covariance of the corresponding k-th Gaussian distribution. With the clustering result, $\kappa(f(\mathbf{x}_t),f(\mathbf{x}'^k))$ can be easily computed. Lastly, we estimate $\{\mathbf{F}_i\}$ and $\{\mathbf{w}_i^k\}$ via contrastive divergence learning (CD) with one-step sampling (Hinton 2002) and stochastic gradient descent algorithm (SGD) (Bottou 2010). Taking partial derivatives of the energy function (5) w.r.t. the parameters leads to the following update:

$$\begin{aligned} \delta\mathbf{F}_i &\propto -\frac{1}{T}\sum_{t}^{T}\frac{\partial E(\mathbf{x}_t)}{\partial\mathbf{F}_i} + \mathbf{E}_{p(\mathbf{x};\{\mathbf{F}_i\},\boldsymbol{\Theta})}(\mathbf{x}), \quad \forall i \\ \delta\mathbf{w}_i^k &\propto -\frac{1}{T}\sum_{t}^{T}\frac{\partial E(\mathbf{x}_t)}{\partial\mathbf{w}_i^k} + \mathbf{E}_{p(\mathbf{x};\{\mathbf{F}_i\},\boldsymbol{\Theta})}(\mathbf{x}), \quad \forall k, i \end{aligned} \tag{6}$$

where $\frac{\partial E(\mathbf{x}_t)}{\partial\mathbf{F}_i}$ is the derivative w.r.t. $\mathbf{F}_i$ at $\mathbf{x}_t$, and $\mathbf{E}_{p(\mathbf{x};\{\mathbf{F}_i\},\boldsymbol{\Theta})}(\mathbf{x})$ is the expectation value w.r.t. the model distribution.

An auxiliary-variable-based Gibbs sampler (Schmidt et al. 2010) is used to draw samples from the model distribution. The expectation can be calculated by averaging over the samples. Full learning scheme is illustrated in Algorithm 1.

Algorithm 1: Learning SA-MRF

```
Input: Training image set $\{\mathbf{x}_t\}$, number of cluster $K$
Output: DSIFT-BoW encoder D, $\{\mathbf{b}^k\}$, $\{\mathbf{w}_i^k\}$, $\{\sigma_k^2\}$
Build encoder D and generate $\{\mathbf{b}_{\mathbf{x}_t}\}$ from $\{\mathbf{x}_t\}$;
Cluster $\{\mathbf{b}_{\mathbf{x}_t}\}$ via GMM, get $\{\mathbf{b}_k\}$, $\{\sigma_k^2\}$, and assign $\{\mathbf{x}_t^k\}$;
Estimate $\{\mathbf{F}_i\}$ and $\{\mathbf{w}_i^k\}$ by minimizing (5) through CD and SGD;
return
```

Applications and experiments

To evaluate the modeling ability of the scene-aware prior on real-world image directly, we evaluate the performance of the learned prior on image denoising and image inpainting. Before the evaluation, we will first introduce some implementation details for learning the scene-aware image prior and the learned model in this paper. Following that, we then revisit the standard Bayesian restoration formulation and derive an MMSE estimation approach for our scene-aware image prior.

Learning details and learning results of the SA-MRF

To learn the MRF model, 450 images in Berkeley Segmentation Database (BSD500) (Martin et al. 2001) are exploited as the training dataset. The training images are transferred to gray-scale. In learning procedure, the scene coefficients, e.g., dense BoW vectors, are extracted from each whole image for data clustering. Nevertheless, 10 patches with size 50 × 50 are cropped randomly from each image to update the low-level MRF model parameter. Note that, the testing images are excluded from training dataset. The number of GMM mode is set as $K = 4$ in this paper. Before extracting DSIFT for both learning and prediction, images are smoothed with a Gaussian kernel whose standard deviation is set as 1.0.

When we set the number of GMM mode K as 4, the training images are split into four sets. We randomly select several representative samples from each cluster and illustrate them into Fig. 4. As shown in Fig. 4, images within same clusters have closed appearances; conversely, images in different clusters have different visual properties. Although the clustering result does not follow the contents strictly, it reflects the low-level properties properly. For example, in Fig. 4, the cluster on the left contains a lot of clear and flat background areas, and the right bottom one has more complex textures and clutters. The clustering result provides a preferred intermediate result to let the algorithm learn diverse and meaning-full features and distributions. As a result, the learned filters and four sets of experts (potential functions) are shown in Fig. 3. Figure 3a shows the learned filters, and b–e are the learned weights and curves of the potential functions for the four clusters, respectively. Comparing to the learning result in Schmidt et al. (2010), our filters have a wider variant region, and the experts have more spiky peak and heavy tail, which reserve the favored image in a narrower region.

Bayesian image restoration formulation

Given an observed image $\mathbf{y}$, which is assumed to be degenerated from a latent high-quality image $\mathbf{x}$. The distribution of $\mathbf{x}$ follows the model in Eq. (1) conditioned on the scene coefficient $f(\mathbf{x})$. The restoration algorithm consists of two steps: (1) generating

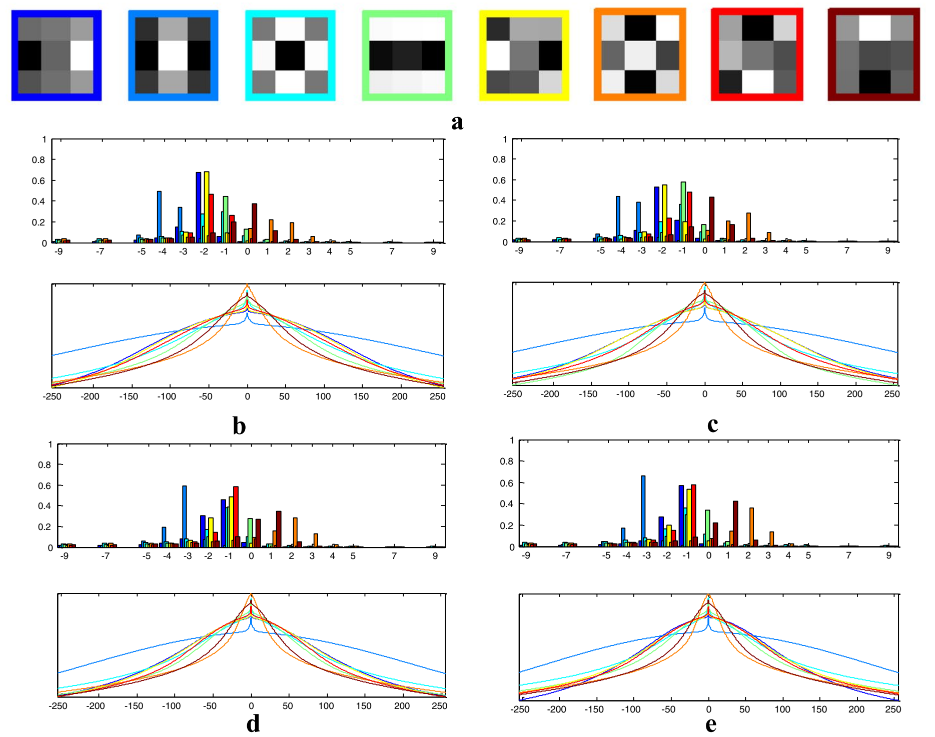

Fig. 3 Learned scene-aware MRF prior. **a** Filters shared by images with different scene contents. **b**–**e** Learned four sets of weights of the potential functions and the shape of each experts

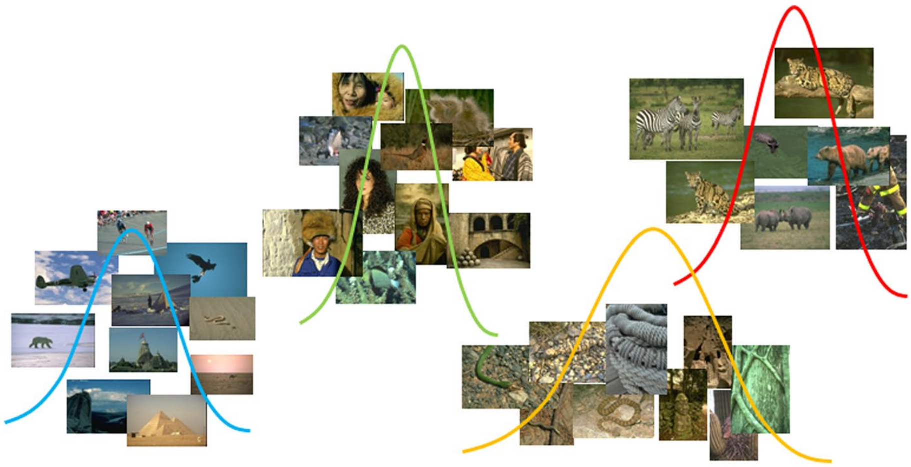

Fig. 4 Samples from the clusters assigned to the different GMM modes in learning process. In this example, the number of GMM mode *K* is set as 4

$\mathbf{b}_{\hat{\mathbf{x}}(\mathbf{y})}$ based on “Potential function conditioned on scene coefficient” section and learned DSIFT-BoW encoder D, and calculating $\kappa(f(\hat{\mathbf{x}}(\mathbf{y})), f(\mathbf{x}'^{k}))$, and (2) estimating the $\hat{\mathbf{x}}$ with the MRF model by computing the *Bayesian minimum mean squared error estimation* (MMSE) through Gibbs sampling:

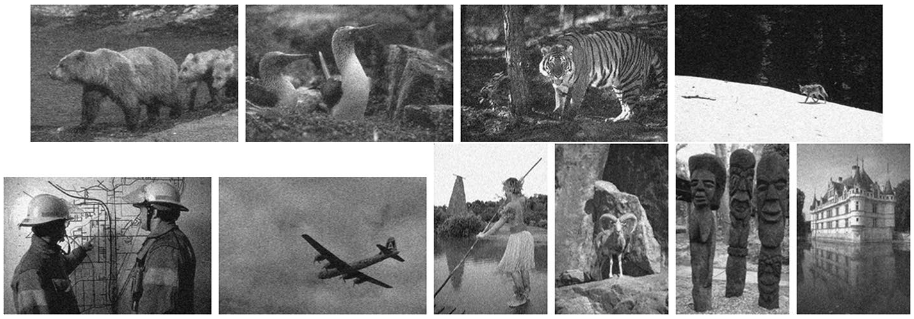

Fig. 5 Testing images with zero-mean Gaussian noise ($\sigma = 25$)

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}'} \int_{\mathbf{x}} \|\mathbf{x}' - \mathbf{x}\|^2 p(\mathbf{x}|\mathbf{y}; \{\mathbf{F}_i\}, \mathbf{w}_i) d\mathbf{x} \tag{7}$$

where $p(\mathbf{x}|\mathbf{y}; \{\mathbf{F}_i\}, \mathbf{w})$ means the image prior distribution from the MRF model in Eq. (1) with calculated $\kappa(f(\hat{\mathbf{x}}(\mathbf{y})), f(\mathbf{x}'^k))$. A posterior version of auxiliary-variable Gibbs sampler is imposed to sample the GSM scale and the latent image iteratively (Schmidt et al. 2010).

Evaluation on image denoising

We focus on comparing our denoising results to the reconstructs relying on the state-of-the-art generative MRF prior in Schmidt et al. (2010) and another broadly used denoising technique BM3D (Dabov et al. 2007), using peak signal-to-noise ratio (PSNR) and gray-scale structural similarity (SSIM) (Wang et al. 2004). Denoising results on 10 test images in BSD500 are illustrated. We consider the restoration performances on Gaussian noise at two levels: $\sigma = 10$ and $\sigma = 25$ (Schmidt et al. 2010). Note that, the noise levels are corresponding to the images in scale [0, 255]. Considering the space limitation, we only illustrate the 10 noisy images with $\sigma = 25$ in Fig. 5.

Figure 6 illustrates the numerical comparison of denoising results among the BM3D method (Dabov et al. 2007), learned image priors in Schmidt et al. (2010) and Schmidt and Roth (2014) and the proposed scene-aware prior. For fairness, regarding the model in Schmidt and Roth (2014), we use the version with 3 × 3 filters, which is same with the settings for the other learning-based method [MRF model in Schmidt et al. (2010) and the proposed method]. The scene-aware prior has a stable superiority on nearly all test images. And the performances of the two MRF prior-based methods are better than that of BM3D. When the noise level is $\sigma = 25$, the recovering results of scene-aware prior exceed the results of another two. Because the scene-aware prior model captures the statistical characteristic of natural image more preciously. When real details and textures are degenerated heavily, the scene-aware prior gives more benefits for recovering texture details and hallucinating the lost information. Figure 7 shows denoising results on test image *goat*, proposed method recovers the details better and avoids the fake texture which appears in the results of Schmidt et al. (2010). Without comparison with the ground truth, these fake details in the result of Schmidt et al. (2010) can be easily recognized as real details, because they satisfy the average perspective of humans

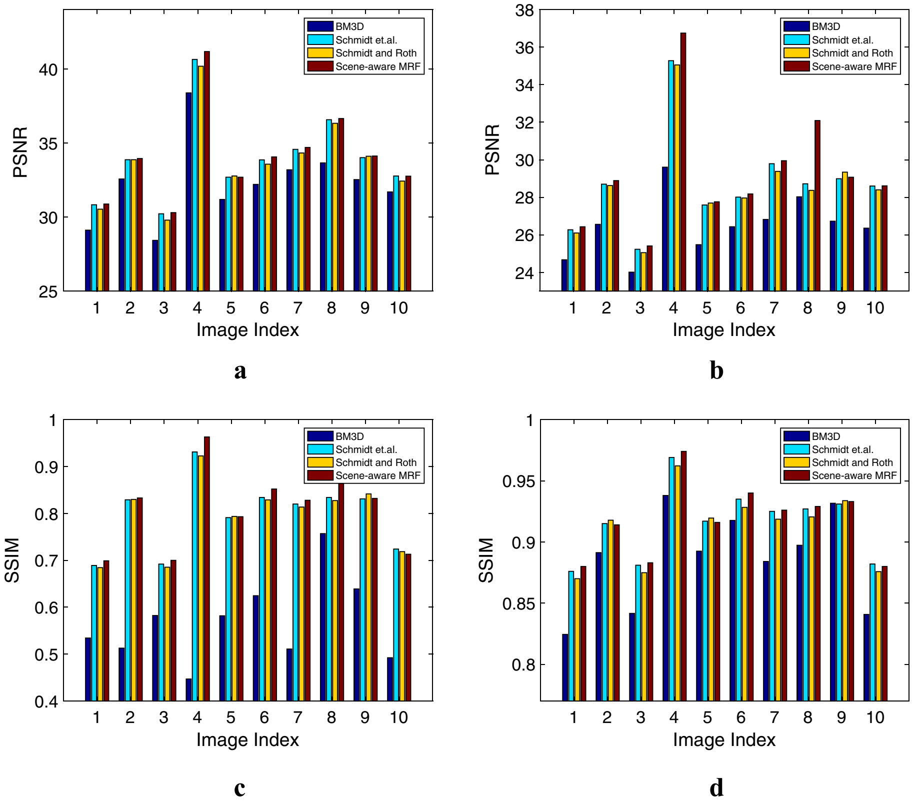

Fig. 6 Image denoising results. **a**, **c** show the PSNR and SSIM values of the results on the images with noise level $\sigma = 10$. **b**, **d** show the PSNR and SSIM values of the results on the images with noise level $\sigma = 25$. BM3D (Dabov et al. 2007), the generative MRF model of Schmidt et al. (2010), shrinkage fields-based method of Schmidt and Roth (2014) and the proposed scene-aware prior are compared

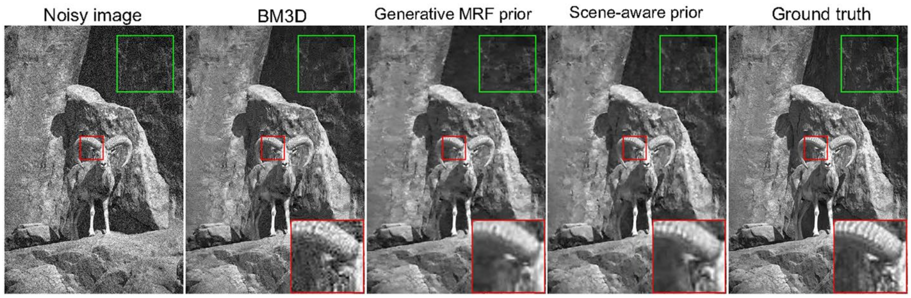

Fig. 7 Denoising results on test image *goat* with $\sigma = 25$. Noisy image: PSNR=19.51dB, SSIM $= 0.42$; result of BM3D (Dabov et al. 2007): PSNR $= 22.43$ dB, SSIM $= 0.51$; Result of generative MRF prior (Schmidt et al. 2010): PSNR $= 26.27$ dB, SSIM $= 0.69$; Our result (scene-aware prior): PSNR $= 26.43$ dB, SSIM $= 0.70$. The proposed method recovers the texture in green box best

on the concept of "natural" image. Figure 8 demonstrates recovering results of the generative MRF prior and our scene-aware prior. Scene-aware prior recovers more detail informations.

When the noise level is low ($\sigma = 10$), the performance of the scene-aware prior is very similar with Schmidt et al. (2010). It can be explained as that there is always noise that is

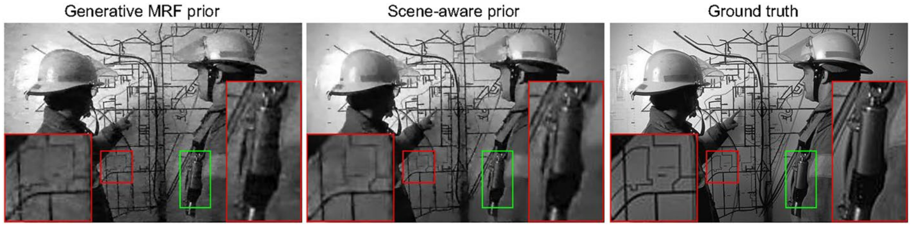

Fig. 8 Denoising results on test image *fireman* with noise level $\sigma = 25$. From left to right: result of generative MRF prior (Schmidt et al. 2010): PSNR = 28.01 dB, SSIM = 0.83; Our result (scene-aware prior): PSNR = 28.18 dB, SSIM = 0.85; and ground truth

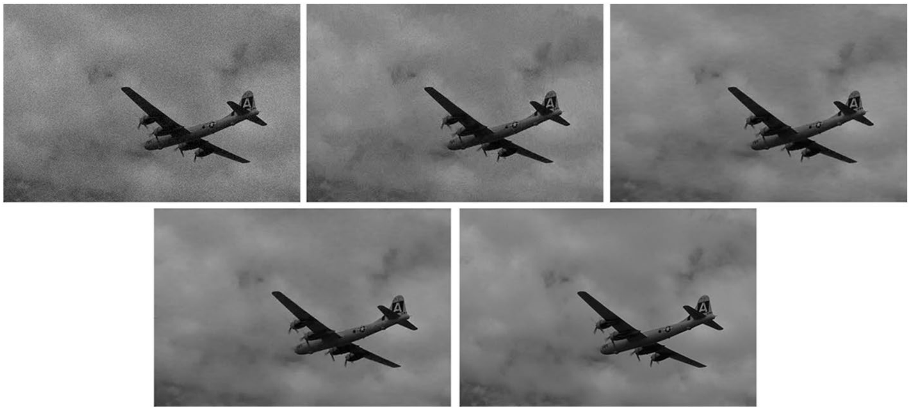

Fig. 9 Denoising results on test image *airplane* with noise level $\sigma = 10$. From left to right: noisy image; result of BM3D: PSNR = 34.89, SSIM = 0.853; result of generative MRF prior (Schmidt et al. 2010): PSNR = 40.64 dB, SSIM = 0.969; our result (scene-aware prior): PSNR = 41.17 dB, SSIM = 0.974; and ground truth

hard to be removed; and both the prior in Schmidt et al. (2010) and ours reach the latent limitation of similar algorithms. An example for visual illustration is shown in Fig. 9. As shown in Fig. 9, our result is much more closed to the ground truth than others. Hence, although the scene-aware prior and the MRF model learned in Schmidt et al. (2010) are closed to each other on the numerical evaluation, the proposed method can achieve more natural and accurate results, which illustrates the power of the scene-aware concept in prior learning.

Evaluation on image inpainting

Image inpainting is to recover a high-quality image from a degenerated image in which part of the image pixels is lost or deteriorated. Apart from the image denoising task, we also test the proposed method on image inpainting in this section. As shown in Fig. 10, a part of an image is crimped and deteriorated due to folding, which is used as the input image in this experiment. Given a binary mask indicating the deteriorated pixels, the MRF model in Schmidt et al. (2010) and the proposed scene-aware prior both work well on recovering an intact image. The result of the proposed however is more natural, especially on the pixels near the crimps in the original image. Since the ground-truth image for the real-world deteriorated image, only the visual comparison is illustrated in Fig. 10.

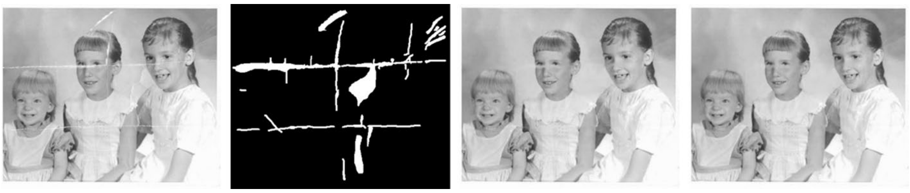

Fig. 10 Evaluation on image inpainting. From left to right: the degenerated image, mask indicating the degenerated part, results of the generative MRF prior (Schmidt et al. 2010), results of the proposed scene-aware prior. Best viewed magnified on screen

Conclusion and future work

The proposed high-order MRF-based scene-aware image prior models the low-level distribution of image conditioned on high-level scene characteristic of observations, and improves the restoration of the degenerated observations. Experimental results demonstrate that the proposed method can generate desirable restoration results.

Our work provides a possible path bridging the low-level prior learning and high-level concept, and proves that the idea can achieve better results than the state-of-the-art methods in similar settings. However, at the same time, it opens up several possible directions for future research:

- Our proposed model learns low-level features in a small local area, and use the simple DSIFT-BoW to express the high-level scene concept, which restricts the expression ability of the model. Embedding the proposed method with the deep convolutional neural network, Bengio et al. (2013) might enable higher expression ability.
- We evaluated the efficiency of the proposed method on image denoising and inpainting tasks. In the future, we may extent this work to more applications tasks, including image superresolution, image deblurring, optical flow, etc.

Authors' contributions

DG drafted the manuscript. YZ, QY, and HL participated in its design and coordination and/or helped to revise the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (61231016, 61572405), China 863 Project 2015AA016402.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 1 August 2017 Accepted: 16 October 2017
Published online: 30 October 2017

References

Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Springer, Berlin, pp 177–186

Chen Y, Yu W, Pock T (2015) On learning optimized reaction diffusion processes for effective image restoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 5261–5269

Cho TS, Joshi N, Zitnick CL, Kang SB, Szeliski R, Freeman WT (2010) A content-aware image prior. In: IEEE conference on computer vision and pattern recognition (CVPR)

Dabov K, Foi A, Katkovnik V, Egiazarian K (2007) Image denoising by sparse 3-D transform-domain collaborative filtering. In: IEEE transactions on image processing

Feng W, Qiao H, Chen Y (2016) Poisson noise reduction with higher-order natural image prior model. SIAM J Imaging Sci 9(3):1502–1524

Fergus R, Singh B, Hertzmann A, Roweis ST, Freeman WT (2006) Removing camera shake from a single photograph. In: ACM transactions on graphics (TOG)

Gong D, Tan M, Zhang Y, van den Hengel A, Shi Q (2016) Blind image deconvolution by automatic gradient activation. In: IEEE conference on computer vision and pattern recognition (CVPR)

Gong D, Yang J, Liu L, Zhang Y, Reid I, Shen C et al (2017) From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural computation

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, Cambridge

Krishnan D, Fergus R (2009) Fast image deconvolution using hyper-Laplacian priors. In: NIPS

Krishnan D, Tay T, Fergus R (2011) Blind deconvolution using a normalized sparsity measure. In: IEEE conference on computer vision and pattern recognition (CVPR)

Levin A, Fergus R, Durand F, Freeman WT (2007) Image and depth from a conventional camera with a coded aperture. ACM Trans Gr 26:70

Levin A, Weiss Y, Durand F, Freeman WT (2009) Understanding and evaluating blind deconvolution algorithms. In: IEEE conference on computer vision and pattern recognition (CVPR)

Li LJ, Socher R, Fei-Fei L (2009) Towards total scene understanding: classification, annotation and segmentation in an automatic framework. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE, New York

Li LJ, Su H, Fei-Fei L, Xing EP (2010) Object bank: a high-level image representation for scene classification & semantic feature sparsification. In: Advances in neural information processing systems

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. In: IEEE international conference on computer vision (ICCV)

Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: IEEE international conference on computer vision (ICCV)

McAuley JJ, Caetano TS, Smola AJ, Franz MO (2006) Learning high-order mrf priors of color images. In: Proceedings of the 23rd international conference on machine learning. ACM, New York, pp 617–624

Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175

Ren J, Liu J, Guo Z (2013) Context-aware sparse decomposition for image denoising and super-resolution. IEEE Trans Image Process 22(4):1456–1469

Roth S, Black MJ (2005) Fields of experts: a framework for learning image priors. In: IEEE conference on computer vision and pattern recognition (CVPR)

Samuel KGG, Tappen MF (2009) Learning optimized MAP estimates in continuously-valued MRF models. In: IEEE conference on computer vision and pattern recognition (CVPR)

Schmidt U, Gao Q, Roth S (2010) A generative perspective on MRFs in low-level vision. In: IEEE conference on computer vision and pattern recognition (CVPR)

Schmidt U, Jancsary J, Nowozin S, Roth S, Rother C (2014) Cascades of regression tree fields for image restoration

Schmidt U, Roth S (2014) Shrinkage fields for effective image restoration. In: IEEE conference on computer vision and pattern recognition (CVPR)

Sonka M, Hlavac V, Boyle R (2014) Image processing, analysis, and machine vision. Cengage Learning, Boston

Sun J, Zhu J, Tappen MF (2010) Context-constrained hallucination for image super-resolution. In: IEEE conference on computer vision and pattern recognition (CVPR)

Tappen MF, Liu C (2012) A Bayesian approach to alignment-based image hallucination. In: ECCV

Tappen MF, Liu C, Adelson EH, Freeman WT (2007) Learning Gaussian conditional random fields for low-level vision. In: IEEE conference on computer vision and pattern recognition (CVPR)

Torralba A, Oliva A (2003) Statistics of natural image categories. Netw Comput Neural Syst 14:391–412

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. In: IEEE transactions on image processing

Weiss Y, Freeman WT (2007) What makes a good model of natural images? In: IEEE conference on computer vision and pattern recognition (CVPR)

Wu Z, Lin D, Tang X (2016) Deep Markov random field for image modeling. In: European conference on computer vision. Springer, Berlin, pp 295–312

Xu L, Zheng S, Jia J (2013) Unnatural l0 sparse representation for natural image deblurring. In: IEEE conference on computer vision and pattern recognition (CVPR). IEEE, New York, pp 1107–1114

Zhang H, Wipf D, Zhang Y (2013) Multi-image blind deblurring using a coupled adaptive sparse prior. In: IEEE conference on computer vision and pattern recognition (CVPR)