文章编号: 1001-0920(2020)02-0469-05 DOI: 10.13195/j.kzyjc.2018.0609

基于变分贝叶斯推断的字典学习算法

刘 连, 王孝通†

(海军大连舰艇学院 航海系，辽宁 大连 116018)

摘 要: 传统的字典学习算法在对训练图像进行学习时收敛速率慢,当图像受到噪声干扰时学习效果变差. 对此,提出一种基于变分推断的字典学习算法. 首先设定模型中各参数的共轭稀疏先验分布;然后基于贝叶斯网络求出所有参数的联合概率密度函数;最后利用变分贝叶斯推断原理计算出各参数的最优边缘分布,训练出自适应学习字典. 利用该字典进行图像去噪实验以及压缩感知重构实验,仿真结果表明,所提出的算法可显著提高字典学习效率,对测试图像的去噪效果和重构精度有很大改善.

关键词: 贝叶斯网络；变分推断；字典学习；图像去噪；压缩感知

中图分类号: TN911.7 **文献标志码:** A

Dictionary learning algorithm based on variable Bayes inference

LIU Lian, WANG Xiao-tong†

(Department of Navigation，Dalian Naval Academy，Dalian 116018，China)

Abstract: The traditional dictionary learning algorithms have slow convergence rate when learning the training image. And the effect of dictionary learning becomes worse if the images are corrupted by noise. Therefore, a dictionary learning algorithm based on variational inference is proposed to solve this problem. The algorithm firstly sets the conjugate sparse prior distribution of the parameters in the model, and then the joint probability density function of all parameters is calculated based on the Bayesian network. Finally, the optimal edge distribution of the parameters is calculated by the variational Bayesian inference, and the adaptive dictionary training is completed. The image denoising experiment and the compressed sensing image reconstruction experiment are carried out by the adaptive dictionary. The simulation results show that the algorithm can significantly increase the efficiency of dictionary learning, and the visual effect of the denoising and the reconstruction of the test images are improved.

Keywords: Bayesian network；variational inference；dictionary learning；image denoising；compressed sensing

0 引 言

随着信息处理技术的不断发展,研究对象的数据维度越来越高,给信号的采集、压缩及存储等环节带来了巨大压力. 信号的稀疏表示是解决上述问题的关键技术,利用一组空间向量基对原始信号进行正交投影,使投影系数中的非零元素个数远小于信号的维度,并以此逼近原信号,该方法在压缩感知、图像去躁以及超分辨率重构等领域得到了广泛应用. 目前,稀疏表示方法大致可分为固定基稀疏化和自适应字典学习[1-3]:固定基稀疏化由于向量基一定,在针对不同类信号处理时可能稀疏表示效果并不理想;字典学习法通过提取信号的训练数据进行学习,自适应地产生一组超完备字典,稀疏化效果较好,是稀疏化表示领域的研究热点.

字典学习法根据原理不同可分为综合字典学习法[4]、解析字典学习法[5]、盲字典学习法[6]等. 综合字典学习法基于训练数据所在空间寻找字典原子,通过稀疏性以及相关性确立正则项,建立优化问题并进行求解,该类算法的学习字典稀疏表示性能较好,但计算复杂度较高. 常见的综合字典学习法有KSVD字典学习[7]、在线字典学习法[8-9]以及非参数贝叶斯字典学习法[10]等. 解析字典学习法从与训练数据正交的空间中确立字典原子,以字典与训练数据的乘积和稀疏系数矩阵的差为保真项进行优化求解,该类算法的学习效率较高,对图像去噪及超分辨率重构等应用效果较好. 常见的解析字典学习法包括解析KSVD字典学习法[11]、稀疏变换字典学习法[12]以及结构化字典学习法[13-14]. 盲字典学习是近年来提出的基于

收稿日期: 2018-05-09；**修回日期:** 2018-08-05.

基金项目: 国家自然科学基金项目(61471412, 61771020, 61373262).

责任编委: 孙秋野.

†通讯作者. E-mail: 602993590@qq.com.

压缩感知的字典学习算法,该类方法不提取训练数据,而是以压缩测量数据为训练样本进行字典学习,具有更强的自适应性,但不能保证得到最优解,往往只能求得次优解. 目前,该类算法主要有压缩字典学习法[15]、自适应稀疏基学习法[16]等.

以上各类字典学习算法在训练样本受到噪声干扰时学习效果并不理想,并且用于去噪的字典学习算法普遍都存在计算复杂度高,收敛速度慢的缺点. 为了解决上述问题,本文提出一种基于变分推断的字典学习法. 该方法针对系数矩阵的稀疏性引入贝塔伯努利先验,根据所建立的最优化模型中各参数的边缘分布,利用平均域原理计算模型各参数的后验分布最优逼近,当算法收敛时,得到最终的自适应学习字典. 以字典的学习效率、图像去噪以及重构实验来验证学习字典的稀疏表示性能,实验结果表明,本文算法在保证重构精度的前提下,学习效率较一般算法有显著提高.

1 模型描述

令数据集矩阵 $X \in R^{M\times N}$,字典矩阵 $D \in R^{M\times K}$,系数矩阵 $S \in R^{K\times N}$,利用字典表示数据集的模型为

$$X = DS + n. \tag{1}$$

其中:n 为服从均值为零、协方差矩阵为 $\gamma_\varepsilon^{-1}I_M$ 的高斯白噪声;初始设定字典矩阵 D 中各原子服从多元高斯分布 $N(d_j|0, M^{-1}I_M), j = 1,2,\cdots,K$;系数矩阵 S 中各列向量服从多元高斯分布 $N(s_i|0, (\tau_0\gamma_\varepsilon)^{-1}I_K), i = 1,2,\cdots,N$. 引入贝塔伯努利稀疏先验分布,原始模型改为

$$X = D(S \odot Z) + n. \tag{2}$$

其中:$\odot$ 为克罗内克积,表示矩阵间对应位置元素的乘积;Z 为隐变量矩阵,$Z \in R^{K\times N}$,每个隐变量服从0与1的伯努利分布 $\text{Bernoulli}(\beta_j)$. 式(2)可化简为

$$X = D\hat{S} + n, \tag{3}$$

其中:$\hat{s}_i = s_i \odot z_i$. 隐变量控制了稀疏矩阵 S 中各列向量的稀疏性,其分布参数 β_j 设定服从贝塔分布 $\text{Beta}(a_0, b_0)$,超参数 γ_ε 服从伽马分布 $\text{Gamma}(\gamma_\varepsilon|c_0, d_0)$. 通过式(3)可进一步得出所有参数的贝叶斯网络,如图1所示.

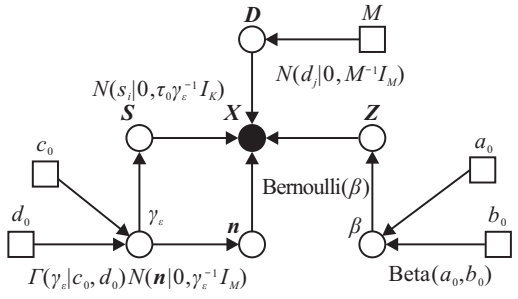


图1 贝叶斯网络模型

2 变分贝叶斯推断

根据采样后的数据集,利用贝叶斯定理各参数的后验分布[17-18]可表示为

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x,\theta)\mathrm{d}\theta}. \tag{4}$$

由图1中的贝叶斯网络模型可知,该模型所含参数较多,并且总体的联合概率密度函数较复杂,对式(4)中最右侧一项的分母进行积分十分困难. 本文利用变分贝叶斯推断算法,根据平均域原理,假设各参数或组合参数间相互独立,为了分别得到其后验概率分布,引入KL散度,利用各参数的真实后验分布寻找一个简单可求的概率分布来近似,即

$$\begin{aligned}&\mathrm{KL}(p(\theta)\|p(\theta|x)) =\\ &\int p(\theta)\ln\frac{p(\theta)}{p(\theta|x)}\mathrm{d}\theta =\\ &\int p(\theta)\ln\frac{p(\theta)p(x)}{p(\theta,x)}\mathrm{d}\theta =\\ &\int p(\theta)\ln\frac{p(\theta)}{p(\theta,x)}\mathrm{d}\theta + \ln p(x).\end{aligned} \tag{5}$$

其中:$\int p(\theta)\ln\frac{p(\theta)}{p(\theta|x)}\mathrm{d}\theta$ 称为变分自由能;$p(x)$ 为数据集的边缘似然分布,是大于零的定值;而KL散度恒大于零. 因此,通过不断增大变分自由能,使得KL散度逐渐接近于零,从而使估计的参数概率分布 $p(\theta)$ 逐渐逼近该参数的真实后验分布 $p(\theta|x)$. 换言之,可以把变分自由能视为 $\ln p(x)$ 的下界,利用类似期望最大化算法的思想使自由能逐渐逼近 $\ln p(x)$. 对变分自由能进行关于 $p(\theta)$ 的泛函求导,得到各参数的通解表达式,即

$$\ln p(\theta_i) = \mathrm{E}_{\{\theta_{mb}\}}\langle\ln p(x,\theta)\rangle + \text{const}, \tag{6}$$

其中 $\{\theta_{mb}\}$ 为关于参数 θ_i 的马尔科夫毯中所含参数[19]. 式(6)为对联合概率密度函数的对数求指定参数的期望,由于该式为关于 θ_i 的函数,其他函数间求期望的值可用常数项表示. 式(6)两边取指数后进行归一化处理可得到 $p(\theta_i)$ 的估计分布.

3 变分字典学习算法

首先根据平均域原理,将原始模型的联合概率分布函数分解为各待求参数的边缘分布,设定待求参数集合为 $\lambda = \{Z, S, D, \gamma_\varepsilon, \beta\}$,联合概率密度可表示为

$$p(X, Z, S, D, \gamma_\varepsilon, \beta) =$$

$$\prod_{i}^{N} p(x_i)p(s_i,\gamma_\varepsilon)\cdots$$
$$\prod_{j}^{K} p(d_j)p(\beta_j)\prod_{i}^{N}\prod_{j}^{K} p(z_{ij}). \tag{7}$$

根据式(6)计算出各组参数的后验概率估计表达式如下.

1) 字典矩阵的各列原子d_j.

$$\ln q(d_j) = \mathrm{E}_{\lambda/\{d_j\}}\langle \ln p(X,\lambda)\rangle + \text{const} =$$
$$\mathrm{E}_{\lambda/\{d_j\}}\langle \ln p(X|d_j,Z,S,n)\rangle + \cdots +$$
$$\mathrm{E}_{\lambda/\{d_j\}} \ln p(d_j) + \text{const} =$$
$$\mathrm{E}_{\lambda/\{d_j\}}\Big\langle -\frac{1}{2}\sum_{i}^{N}(x_i - D(s_i\odot z_i))^{\mathrm{T}}\gamma_\varepsilon I_M(x_i - \cdots -$$
$$D(s_i\odot z_i))\Big\rangle + \ln p(d_j) + \text{const}.$$

化简后得到服从多元高斯分布d_j,均值向量和协方差矩阵分别为

$$\begin{cases}\mu_{d_j} = \Sigma_{d_j}\sum_{i=1}^{N}\mathrm{E}(s_{ij}^2)\mathrm{E}(z_{ij}^2)\hat{x}_i,\\ \Sigma_{d_j} = \Big(M^{-1}I_M + \mathrm{E}(\gamma_\varepsilon)\sum_{i=1}^{N}\mathrm{E}(s_{ij}^2)\mathrm{E}(z_{ij}^2)\Big)^{-1}.\end{cases} \tag{8}$$

2) 为了简化运算,将系数矩阵各行向量及其协方差参数合并求解,即求参数组(s_j,γ_ε)的概率分布函数,有

$$\ln q(s_j,\gamma_\varepsilon) =$$
$$\mathrm{E}_{\lambda/\{s_j,\gamma_\varepsilon\}}\langle \ln p(X|d_j,Z,s_j)\rangle + \cdots +$$
$$\ln p(s_j|\gamma_\varepsilon) + \ln p(\gamma_\varepsilon) + \text{const}.$$

经化简后得到正态伽马分布为

$$(s_{ij},\gamma_\varepsilon) \sim \text{Normal-gamma}(\mu_{s_{ij}},\tau_{s_{ij}},c_j,d_j),$$

该分布概率密度函数可表示为

$$p(s_{ij},\gamma_\varepsilon) = \frac{d_0^{c_0}}{\mathrm{Gamma}(c_0)\sqrt{2\pi}}\gamma_\varepsilon{}^{c_0-1}\cdots$$
$$\exp\Big(-d_0\gamma_\varepsilon - \frac{\tau_0\gamma_\varepsilon(s_{ij}-\mu_{s_{ij}})}{2}\Big). \tag{9}$$

其中各参数为

$$\mu_{s_{ij}} = \mathrm{E}(d_j^{\mathrm{T}})\mathrm{E}(z_{ij})\hat{x}_i\tau_{s_{ij}}^{-1},$$
$$\tau_{s_{ij}} = \mathrm{E}(z_{ij}^2)\mathrm{E}(d_j^{\mathrm{T}})\mathrm{E}(d_j) + \tau_0,$$
$$c_i = K + N + c_0,$$
$$d_i = \frac{1}{2}\tau_{s_{ij}}\hat{x}_i^{\mathrm{T}}\hat{x}_i + d_0 - \frac{1}{2}\tau_{s_{ij}}\mu_{s_{ij}}^{\mathrm{T}}\mu_{s_{ij}}.$$

3) 隐变量矩阵各元素z_{ij}.

$$\ln q(z_{ij}) = \mathrm{E}_{\lambda/\{z_{ij}\}}\langle \ln p(X|D,z_{ij},S)\rangle + \cdots +$$
$$\ln p(z_{ij}|\beta_j) + \text{const}.$$

由于z_{ij}的取值始终为0或1,直接求边缘分布较为困难. 令$z_{ij}=1$,有

$$\ln p(z_{ij}=1) \propto p(1) =$$
$$-\frac{1}{2}\tau\gamma_\varepsilon\mathrm{E}(s_{ij}^2)\mathrm{E}(d_j^{\mathrm{T}})\mathrm{E}(d_j)\cdots$$
$$-\tau\gamma_\varepsilon\mathrm{E}(s_{ij})\mathrm{E}(d_j^{\mathrm{T}})\hat{x}_i + \ln\beta_j.$$

令$z_{ij}=0$,有

$$\ln p(z_{ij}=0) \propto \ln p(0) = \ln(1-\beta_j).$$

因此得到z_{ij}服从伯努利分布,有

$$z_{ij} \sim \text{Bernoulli}\Big(\frac{p(1)}{p(0)+p(1)}\Big). \tag{10}$$

4) 隐变量元素分布参数β_j.

$$\ln q(\beta_j) = \mathrm{E}_Z\langle \ln p(Z|\beta_j)\rangle + \ln p(\beta_j) + \text{const}.$$

化简后得到β_j服从贝塔分布,有

$$\beta_j \sim \text{Beta}(a_j,b_j). \tag{11}$$

其中

$$a_j = a_0 + \sum_{i=1}^{N}\mathrm{E}(z_{ij}),$$
$$b_j = b_0 + N - \sum_{i=1}^{N}\mathrm{E}(z_{ij}).$$

4　算法步骤

综上所述,变分字典学习算法的实现步骤如下.

输入: 数据集矩阵X,循环次数r,重构数据集误差阈值T;

输出: 字典矩阵D.

初始化: 设定噪声逆协方差参数,隐变量矩阵初值Z以及超参数τ_0,a_0,b_0,c_0,d_0.

Step 1: 根据式(10)计算出隐变量矩阵的边缘分布函数.

Step 2: 将隐变量期望代入式(8)、(9)、(11)中,更新各参数或参数组的边缘概率分布.

Step 3: 利用式(3)重构数据集矩阵$\tilde{X}$,计算其与原始数据集的均方误差ε.

Step 4: 判定均方误差ε与阈值的大小,若大于阈值则转Step 5,否则转Step 6.

Step 5: 更新循环次数并判定是否达到R,若达到则转Step 6,否则转Step 1.

Step 6: 输出训练字典.

5　实验结果与分析

为了检验本文算法的时效性,选定离散余弦字典、KSVD字典学习以及非参数贝叶斯字典(BPFA),通过训练字典、图像去噪以及压缩感知图像重构实

验进行对比,评价指标为训练字典的时间以及图像处理后的峰值信噪比. 实验运行环境为Intel(R) Core (TM)2 Duo dual-core processor(E4600),内存3 GB,操作软件选用Matlab2011b. 给定测试图像goldhill,尺寸为512×512,对该图像截取图像块并向量化存入训练数据集中,截取尺寸大小设定为9 × 9. 字典中的原子个数过小会影响稀疏表示性能,过大会影响训练效率. 通过实验发现,设定原子数超过某一定值后稀疏表示性能趋于稳定,本文中设定字典原子数为2 000.

5.1 训练字典

除了合理设定字典矩阵的尺寸大小,系数矩阵的稀疏度也影响着训练字典的效率. 本文算法通过设定隐变量服从稀疏先验分布,自适应地确定系数矩阵稀疏度. 各算法的字典训练时间如表1所示.

表 1 各算法字典训练的时间

	各类算法			
	DCT	KSVD	BPFA	本文算法
训练时间/s	1 967	2 365	3 521	1 057

由表1数据可以看出,本文算法在字典学习效率上较其他算法有明显提升,DCT算法与KSVD算法的学习效率相近,而BPFA算法的学习效率最慢.

5.2 图像去噪

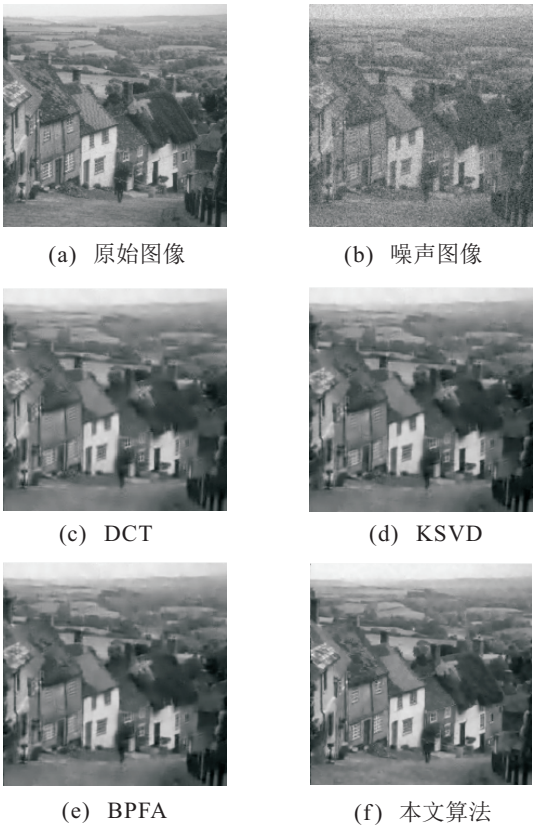
针对本文算法的去噪性能,对各测试图像加入不同标准差的高斯噪声,选定指标为图像去噪后的峰值信噪比,实验结果如表2所示.

表 2 各算法去噪时间与峰值信噪比

噪声标准差	峰值信噪比/dB			
	DCT	KSVD	BPFA	本文算法
10	34.98	34.62	34.91	35.32
15	32.28	32.87	33.05	33.76
20	30.19	30.35	31.26	31.92
25	28.66	28.94	29.24	29.52
30	26.32	26.78	27.41	27.79
35	24.07	24.59	25.32	25.73

由表2可以看出,在噪声标准差为25～35时,本文算法的去噪后图像信噪比在不同条件下均稍高于另外3种对比算法,同时在噪声标准差小于20后,各类去噪算法的峰值信噪比变化趋于平缓,但本文算法依然优于其他算法. 针对测试图像goldhill,各算法的去噪效果图如图2所示.

由图2可以看出,本文算法的去噪效果与BPFA算法效果相近,但比DCT算法和KSVD算法都有明显提高,房子的轮廓和窗框的线条纹理都有了进一步的改善.



(a) 原始图像 (b) 噪声图像

(c) DCT (d) KSVD

(e) BPFA (f) 本文算法

图 2 各算法去噪效果

5.3 压缩感知图像重构

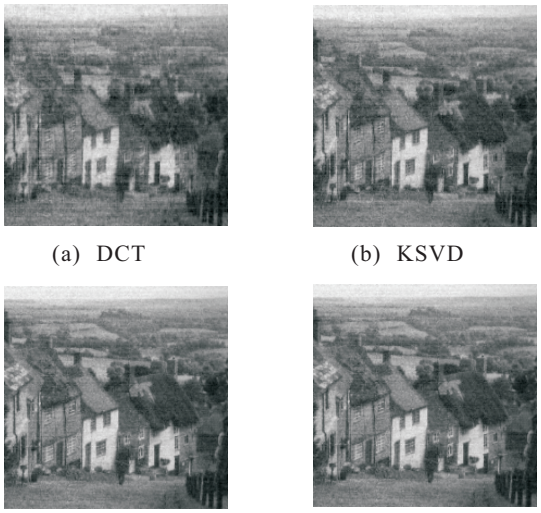
利用学习字典对测试图像进行稀疏表示,选定不同采样率下的高斯随机测量矩阵进行压缩感知,并完成图像的重构实验. 选定指标为重构图像的峰值信噪比,实验结果如表3所示.

表 3 各算法重构时间与峰值信噪比

采样率	峰值信噪比/dB			
	DCT	KSVD	BPFA	本文算法
0.3	24.54	24.75	25.21	25.83
0.4	25.83	26.11	26.42	26.63
0.5	26.56	27.38	27..69	28.01
0.6	28.65	29.07	29.52	29.73
0.7	30.87	31.06	31.12	31.18
0.8	32.16	32.59	33.46	33.67

由表3可以看出: 当采样率大于0.6时,各类算法的重构图像信噪比均大于30 dB;随着采样率的下降,信噪比逐渐减小,本文算法的重构图像信噪比要比DCT算法以及KSVD算法提高1 dB左右,而与BPFA算法的重构效果相近. 针对测试图像goldhill,各算法的重构效果如图3所示.

由图3可以看出,在DCT算法的重构效果中远处的景象较为模糊,KSVD算法稍有改善,而BPFA算法和本文算法的重构图像视觉效果最优,不论是近处的房屋还是远处的树林,重构精度都有了显著提高.



(a) DCT　(b) KSVD　(c) BPFA　(d) 本文算法

图 3　各算法重构效果

6 结　论

本文基于变分贝叶斯推断提出了一种稀疏度自适应的字典学习算法. 该算法通过平均域原理不断迭代逼近各参数的真实边缘分布,自适应地训练出全局字典;针对传统训练字典训练数据受噪声干扰时学习鲁棒性不高和效率较低的缺点,利用本文算法均得到了改善. 最后,通过实验验证了本文算法在字典训练时所用时间较短,在图像去噪和重构效果上较其他算法均有一定的提高.

参考文献(References)

[1] Tosic I, Frossard P. Dictionary learning[J]. Signal Processing Magazine of IEEE, 2011, 28(2): 27-38.

[2] Sulam J，Ophir B，Zibulevsky M，et al. Trainlets: Dictionary learning in high dimensions[J]. IEEE Transactions on Signal Processing, 2016, 64(12): 3180-3193.

[3] Lu J, Wang G, Zhou J. Simultaneous feature and dictionary learning for image set based face recognition[J]. IEEE Transactions on Image Process, 2017, 26(8): 4042-4054.

[4] Mehrdad Y, Laurent D, Mike E D. Parametric dictionary design for sparse coding[J]. IEEE Transactions on Signal Processing, 2009, 57(12): 4800-4811.

[5] Ron R, Alfred M B, Elad M. Dictionaries for sparse representation modeling[J]. Proceedings of the IEEE, 2010, 98(6): 1045-1057.

[6] Moshe M, Yonina C E. Blind multiband signal reconstruction: Compressed sensing for analog signals[J]. IEEE Transactions on Signal Processing, 2007, 57(3): 993-1009.

[7] Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11): 4311-4322.

[8] Mairal J, Bach F, Ponce J, et al. Online learning for matrix factorization and sparse coding[J]. Journal of Machine Learning Research, 2009, 11(1): 19-60.

[9] Yashar N, Soosan B, Mohammad A T. Online learning for matrix factorization and sparse coding[J]. IEEE Transactions on Signal Processing, 2015, 64(3): 592-602.

[10] Zhou M, Chen H, Paisley J. Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images[J]. IEEE Transactions on Image Processing, 2011, 21(1): 130-144.

[11] Rubinstein R, Peleg T, Elad M. Analysis K-SVD:A dictionary learning algorithm for the analysis sparse model[J]. IEEE Transactions on Signal Processing, 2013, 61(3): 661-677.

[12] Saiprasad R, Yoram B. Sparsifying transform learning for compressed sensing MRI[C]. IEEE International Symposium on Biomedical Imaging. San Francisco, 2013: 17-20.

[13] Jenatton R, Mairal J, Obozinski G, et al. Proximal methods for hierarchical sparse coding[J]. Journal of Machine Learning Research, 2010, 12(7): 2297-2334.

[14] Boaz O, Michael L, Elad M. Multi-scale dictionary learning using wavelets[J]. IEEE Journal of Selected Topics in Signal Processing, 2011, 5(5): 1014-1024.

[15] Mohammad A, Hayder R. Compressive dictionary learning for image recovery[C]. IEEE International Conference on Image Processing. Orlando, 2012: 661-664.

[16] Jian Z, Chen Z, Debin Z, et al. Image compressive sensing recovery using adaptively learned sparsifying basis via L_0 minimization[J]. Signal Processing, 2014, 103(10): 114-126.

[17] Seeger M W, Wipf D P. Variational Bayesian inference Techniques[J]. IEEE Transactions on Information Theory, 2010, 27(6): 81-91.

[18] Fox C W，Roberts S J. A toturial on variational Bayesian inference[J]. Artificial Intelligence Review, 2012, 38(2): 85-95.

[19] Pravin K R, Jalil T, Markus F L. A variational Bayesian inference framework for multiview depth image enhancement[J]. IEEE International Symposium on Multimedia, 2013, 41(11): 183-190.

作者简介

刘连 (1990−), 男, 博士生, 从事数字图像处理、压缩感知重建算法的研究, E-mail: lljtxy33@163.com;

王孝通 (1962−), 男, 教授, 博士生导师, 从事数字图像处理、雷达导航等研究, E-mail: 602993590@qq.com.

(责任编辑：李君玲)