A Fast Algorithm for Learning the Overcomplete Image Prior

Zhe WANG†
, Nonmember, Siwei LUO†, Member, and Liang WANG†, Nonmember

SUMMARY
In this letter, we learned overcomplete filters to model rich priors of nature images. Our approach extends the Gaussian Scale Mixture Fields of Experts (GSM FOE), which is a fast approximate model based on Fields of Experts (FOE). In these previous image prior models, the overcomplete case is not considered because of the heavy computation. We introduce the assumption of quasi-orthogonality to the GSM FOE, which allows us to learn overcomplete filters of nature images fast and efficiently. Simulations show these obtained overcomplete filters have properties similar with those of Fields of Experts (FOE) and denoising experiments also show the superiority of our model.

key words: overcomplete, Fields of Experts, GSM FOE, image denoising

1. Introduction
How to learn rich, generic image priors that capture the statistics of natural scenes has been the subject of intense study. Many low-level vision algorithms assume a prior probability over nature images and then use this image prior to predict unseen quantities. The inferred “quantities” should be consistent with the relevant information in the “dirty” input image data, and at the same time it should be in accordance with the statistics of natural scenes, in other words it should make sense visually. Examples of tasks that have been tackled by using this approach include image denoising and inpainting [2], [7], [9].

Recently, Roth and Black introduced Fields of Experts model (FOE) to represent a richer prior structure for images than pairwise smoothness [2]. The main idea of FOE is that the distribution of image \( x \) can be regarded as the product of the potential over all maximum cliques (3 × 3 or 5 × 5 pixels) in MRF model. The form is defined as

\[
p(x) = \frac{1}{Z_{	ext{FOE}}(\theta)} \prod_{i,k} \phi_k (\omega_{ik}^T x) \tag{1}
\]

and

\[
Z_{	ext{FOE}}(\theta) = \int x \prod_{i,k} \phi_k (\omega_{ik}^T x) dx \tag{2}
\]

where \( \omega_{ik}^T \) is result of applying the linear filter \( \omega_k \) to image \( x \) at clique i. Function \( \phi \) is a parametric, student T distribution[4] for the potentials to model the well known property of nature images, which is non-Gaussian. The partition function \( Z_{	ext{FOE}}(\theta) \) is an explicit normalization constant. Figure 1 (a) shows these learned filters using maximum likelihood. They do not resemble derivative filters at all. Nevertheless, Roth and Black showed that the learned filters gave superior performance compared to simple derivative filters on a range of image-restoration problems.

However, these MRF-based models in general are extremely difficult to learn and they require high computational cost. In both [1], [2], learning is performed by gradient ascent which includes the gradient of the partition function (2) that is intractable (requires Monte-Carlo sampling at every step). Despite Roth and Black using an approximate sampling method called “contrastive divergence” [3], these MCMC-based methods were unacceptably slow.

In a recent paper, Weiss and Freeman presented tractable lower and upper bounds on the partition function of models, which was based on recent results in machine learning that dealt with Gaussian potentials [7], [8]. They extended these results to non-Gaussian potentials and derived a novel basis rotation algorithm for approximating the maximum likelihood filters. The experimental result showed that the learning can be accomplished in a matter of minutes, and denoising experiments approved the learnt filters represented a true property of natural images.

Nevertheless, overcompleteness (i.e., the number of filters is greater than the dimensions of observed data) is an important property of low-level vision [10] and none of these image prior algorithm is able to concern this. Although the Fields of Experts model can learn more filters than its data dimensions in theory, the high computational costs make it impossible in practice. In this letter, we aim at developing a fast algorithm for learning the overcomplete image prior. Our approach extends the Gaussian Scale Mixture Fields of Experts (GSM FOE) and combines it with the quasi-orthogonal estimation used in ICA [5], [12]. In the following simulations show that these learnt overcomplete filters resemble oriented receptive fields and outperform previous image prior models [2], [7], [9] in denoising experiment.

2. Model
2.1 GFOE and GSM FOE

To address the huge computational burden in learning FOE, Weiss and Freeman presented Gaussian Fields of Experts (GFOE) and GSM FOE model to make efficient learning algorithms [7]. GFOE is a FOE model which have the Gaus-
Fig. 1 (a) The 24 $5 \times 5$ filters learned by FOE algorithm. (b) The 25 $15 \times 15$ filters learned by GSM FOE algorithm.

Fig. 2 (a-b) The 15 $15$ whitening filter and its power spectrum (c-d) A typical filter derived by FOE (from Fig. 1 (a)) and its power spectrum.

Gaussian potential function. The form is defined as

$$p(x) = \frac{1}{Z_{GFOE}(\omega_k)} \prod_{i,k} \exp\left(-\left(\omega_{ik}^T x\right)^2\right)$$

(3)

and the partition function is

$$-\ln Z_{GFOE}(\omega_k) = \ln \det \left(\sum_{i,k} \omega_{ik} \omega_{ik}^T\right)$$

(4)

When converting the filters $\{\omega_k\}$ into frequency domain, it can prove that

$$-\ln Z_{GFOE}(\omega_k) = \sum_{\omega} \ln \left(\sum_{k} |W_k(\omega)|^2\right)$$

(5)

where $W_k(\omega)$ is the Fourier transform of filter $\omega_k$. So the maximum likelihood filters should favor the set that maximum $-\ln Z$. When the filters were constrained to be unit norm, they showed that the whitening filter [6] is the optimal filter for $K = 1$ in the GFOE model. Figure 2 shows the whitening filter and a typical filter derived by FOE — These two filters have the same power spectrums that are mostly concentrated at high frequencies. Unfortunately, the nature images are non-Gaussian and it is not appropriate to assume that the potential function is Gaussian. Based on the GFOE, they presented the GSM FOE model and used the GSM (Gaussian Scale Mixture) [11] to approximate the non-Gaussian potential:

$$\phi(y) \propto \sum_{j=1}^{J} \frac{\pi_j}{\sigma_j} \exp\left(-\frac{1}{2\sigma_j^2} y^2\right)$$

(6)

where $\pi_j$ and $\sigma_j$ are fixed to model the non-Gaussian distribution such as student-t [4]. So the GSM FOE has the form of

$$p(x) = \frac{1}{Z_{GSM}(\omega_k)} \prod_{i,k} \phi_k(\omega_{ik}^T x)$$

$$= \frac{1}{Z_{GSM}(\omega_k)} \exp\left(-\sum_{i,k} E_{GSM}(\omega_{ik}^T x)\right)$$

(7)

$Z_{GSM}$ is the partition function of GSM FOE model and still difficult to compute. Fortunately, it can prove that there are tractable lower and upper bounds on the partition function, which is

$$\ln Z_{GFOE} + Na \leq \ln Z_{GSM} \leq \ln Z_{GFOE} + Nb$$

(8)

where $a$, $b$ and $N$ are all constants that do not depend on the value of filters.

From the upper and lower bounds on $\ln Z_{GSM}$, they presented a Basis Rotation Algorithm for learning the optimal filters in GSM FOE model. The main idea is that they search the filters in a restrict set where $\ln Z_{GSM}$ is constant, and learn filters that minimize the energy $E_{GSM}$ in this set. If basis set $B \in \mathbb{R}^{\infty \times K}$ contains $K$ filters for columns, $\{b_i\}_{i=1}^{K}$, and orthogonal matrix $D \in \mathbb{R}^{K \times K}$ contains $K$ vectors for columns, $\{d_i\}_{i=1}^{K}$, from (5) it can prove that

$$\ln Z_{GFOE}(B) = \ln Z_{GFOE}(BD)$$

(9)

It means that rotating the basis set $B$ in the orthogonal space does not change the $\ln Z$. Based on this, they used a variant of the EM algorithm for learning the optimal $d_i$, one by one so that $W = BD$ minimizes the energy $E_{GSM}$. Learning can be accomplished in a matter of minutes and applying these obtain filters (see Fig. 1 (b)) in denoising experiment can lead to state-of-the-art results.

2.2 Rotating in the Overcomplete Case

It seems that the Basis Rotation Algorithm still work in the overcomplete case. When $\hat{B}$ is an overcomplete basis set $(n < K)$, the optimal overcomplete filters can be learnt by rotating $\hat{B}$. However, it is not feasible because such overcomplete basis set $\hat{B}$ is also hard to be found.

Here, we present an Overcomplete Basis Rotation Algorithm, which combines Basis Rotation Algorithm with the quasi-orthogonal estimation used in ICA [5], [12]. Our approach is inspired by the fact that these filters estimated by Basis Rotation Algorithm vary as $D$ starts with different initial values. The number of the optimal filters seems to exceed the dimension of the orthogonal matrix $D$. So a direct idea is to rotate basis set $B$ with an overcomplete matrix $\tilde{D}$.

One important constraint of the Basis Rotation Algorithm is that the value of the orthogonal matrix $D$ does not change the value $\ln Z$ (9), so what kinds of matrix $\tilde{D}$ does satisfy such constraint?

Defined orthogonal matrix $\tilde{D} \in \mathbb{R}^{K \times (M \times K)}$ contains $M \times K$ vectors for columns, $\{\tilde{d}_j\}_{j=1}^{M \times K}$, which means that $\tilde{D}$ is $M \times$ overcompleteness. From (5) it can prove that

$$\ln Z_{GFOE}(\tilde{B}) \approx \ln Z_{GFOE}(\tilde{B}D) + \ln M$$

(10)

if $\tilde{d}_i^T \tilde{d}_j \approx 0$ for any $i \neq j$

In high dimension space, such a vector set $\tilde{D}$ ($\tilde{d}_i^T \tilde{d}_j \approx 0$) is called quasi-orthogonal. It means that $\ln Z$ doesn’t vary with the value of quasi-orthogonal matrix $\tilde{D}$. 
Matrix $\hat{D}$ does not seem to exist because it is generally thought that there are only $n$ vectors that are orthogonal to each other in an $n$ dimension space. Actually, there is much more place for vectors in the high dimension spaces, and it is possible to have $2n$ vectors that are practically orthogonal in the $n$ dimension space. As $n$ is larger, the number of quasi-orthogonal vectors grow and the angles between these vectors are as close as 90 degrees.

Another improvement of our algorithm is the symmetric estimation [5]. Unlike Basis Rotation Algorithm learnt vectors one by one, our algorithm is parallel estimation, which means that there are no priorities among these vectors. This brings two merits: first, some accumulation errors that are probably caused by orthogonalization procedure in original algorithm will be avoided. Secondly, it will have the advantage of parallel computation.

So it is desired that the overcomplete vectors can be learnt by a parallel algorithm. Such estimation is easy in traditional ICA model [5], [12], and can be accomplished by a quasi-orthogonal process, which is the formula:

\begin{equation}
1. \text{Let } \hat{D} \leftarrow \frac{1}{2} \hat{D} - \frac{1}{2} \hat{D} \hat{D}^T \hat{D} \tag{11}
\end{equation}

2. Normalize each column of $\hat{D}$ to unit norm

This is an iterative procedure that maximizes the angles between vectors in $\hat{D}$, and in our algorithm, it just needs to be performed once after every iteration. Overall, we denote $\{x(t)\}$ as training set and "eig min" as the eigenvector corresponding to the minimum eigenvalue of a matrix. The details of the Overcomplete Basis Rotation Algorithm is summarized as follows:

1. Initialization: Choose a random matrix $\hat{D}$, and quasi-orthogonalize $\hat{D}$ using (11). Set $\hat{W} = \hat{BD}$
2. Repeat $S$ times:
   a. For all $\hat{d}_i \in \hat{D}$:
      i. E step: $q_j \propto \frac{\hat{d}_i^T x(t)}{\sigma_i} \exp \left( -\frac{1}{2 \sigma_i^2} (\hat{d}_i^T x(t)) \right)$
      ii. M step: $\hat{d}_j = \text{eig min} B^T \left( \sum_{t \in T} \frac{q_j}{\sigma_j} x(t)x(t)^T \right) B$
   b. Quasi-orthogonalize $\hat{D}$ using (11), and $\hat{W} = \hat{BD}$

3. Experiment

3.1 Learning

We applied our algorithm on nature images. We used the same training data as those used by Roth and Black [2], Weiss and Freeman [7], which is a subset of the Berkeley segmentation database. Before learning, we initialized $\hat{D}$ randomly but quasi-orthogonalized. Besides, we chose the 25 $15 \times 15$ shifted versions of the whitening filter as the basis set $B$.

Compared with Weiss and Freeman, we learned $15 \times 15$ filters with $1.6 \times$ and $2 \times$ overcompleteness, respectively.

We performed 200 iterations in each experiment and quasi-orthogonalized the outputs once after each iteration. This algorithm is still fast and can be accomplished in a matter of minutes. Figure 3 show the learnt filters $1.6 \times$ and $2 \times$ overcomplete filters (compare with the complete case in Fig. 1 (b)). Like the filters obtained in [7], these overcomplete filters have the similar mean power spectrum with the Roth and Black’s: they are all predominantly high frequency. On the other hand, these filters exhibit more structures than Roth and Black’s, which means that they are different in the high-order statistics. In Fig. 3, we can see these filters are more clean and localized in space, oriented, and bandpass, like those given by learning receptive fields [6], [10], [12].

To make sure the quasi-orthogonalizion process has worked and these learnt filters are unique, i.e. no identical filters have been estimated several times, we calculated the angles between each vector and the remaining vectors in the overcomplete matrix $\hat{D}$. Figure 4 shows the angles obtained by complete case [7] and our $1.6 \times$ and $2 \times$ overcomplete matrix. We can see that all the angles derived from GSM FOE model are 90 degrees for its orthogonal conditions. Whereas, all the angles derived from our model are quite large and close to 90 degrees, which demonstrates that these vectors are really quasi-orthogonal.

3.2 Denoising

To demonstrate the attractive property of these overcomplete filters, we compare the denoising performance of different image prior models.

We assume that an image has been corrupted with Gaussian noise of known variance. Given an observed noisy image $y$, our goal is to find the true image $x$ that maximizes the posterior probability

![Fig. 3](image324x514to389x579.png)

(a) 40 $15 \times 15$ filters with $1.6 \times$ overcompleteness (b) 50 $15 \times 15$ filters with $1.6 \times$ overcompleteness (c) A typical power spectrum of the overcomplete filter.

![Fig. 4](image400x644to480x714.png)

(a) (b) (c)

The angles between the learned basis vectors (a) complete case (b) $1.6 \times$ overcomplete case (c) $2 \times$ overcomplete case.
Table 1 Denoising PSNR results in [DB].

| $\sigma$ | Lena | Barbara | Boats | House | Peppers | Ours | W&F | DM |
|---|---|---|---|---|---|---|---|---|
| 1 | 47.31 | 47.61 | 44.02 | 48.68 | 46.88 | 46.90 | 47.02 | 47.9 |
| 5 | 37.52 | 36.33 | 35.42 | 37.53 | 36.39 | 36.64 | 36.76 | 36.5 |
| 10 | 34.35 | 31.90 | 32.67 | 34.02 | 33.16 | 33.22 | 33.28 | 32.8 |
| 15 | 32.45 | 29.36 | 30.77 | 32.22 | 31.23 | 31.21 | 31.21 | 30.7 |
| 20 | 31.12 | 27.59 | 29.39 | 30.92 | 29.83 | 28.77 | 29.56 | 29.0 |
| 25 | 30.09 | 26.21 | 28.33 | 29.94 | 28.78 | 28.66 | 28.48 | 27.8 |
| 50 | 27.07 | 22.87 | 25.27 | 26.79 | 25.74 | 25.54 | 24.69 | 24.9 |
| 75 | 25.48 | 22.12 | 23.74 | 24.96 | 23.97 | 24.05 | 22.55 | 23.5 |
| 100 | 24.37 | 21.63 | 22.71 | 23.65 | 22.81 | 23.03 | 21.11 | 22.6 |

1 refers to the complete case by Weiss and Freeman [7].
2 refers to the Directed Model by Domke et al. [9].

Fig. 5 Comparison of denoising results on ‘Lena’ images. (a) Original image. (b) Image with additive Gaussian noise ($\sigma = 25$); PSNR = 20.17 dB. (c) Our method; PSNR = 30.11 dB. (d) GSM FOE model by Weiss and Freeman [7]; PSNR = 29.32 dB. (e) Directed Model from [9]; PSNR = 29.48 dB.

$$p(x|y) \propto p(y|x)p(x)$$

(12)

where $p(y|x) \propto \exp(||y - x||^2/(2\sigma^2))$, and $p(x)$ is the overcomplete prior over nature image learnt above. Note that Roth and Black [2] used an additional set of parameters $\lambda(\sigma)$ to tune $p(y|x)\lambda(\sigma)p(x)$, and the parameters were chosen by the denoising performance. Weiss and Freeman reported that GSM FOE got a better performance than FOE without such additional parameters in image denoising. Here, we didn’t consider such parameters either and used conjugate gradient descent to maximize $p(x|y)$.

In Table 1, our results for $2 \times$ overcompleteness are compared with the GSM FOE model [7] and a recent work called Directed Model from [9] by using a standard denoising test set. (The latter method introduced a directed model for computation saving to learn the image prior, and denoising experiments showed that it got a better result in the case of low and high noise levels than GSM FOE.) The table gives results in terms of the peak signal to noise ratio (PSNR). Besides, the results from our experiments and [7] are averaged over 5 experiments for each image and each level of noise. Our approach outperforms both the complete case and Directed Model except the noise levels lower than $\sigma = 15$. Figure 5 provides a visual comparison of each of these methods on ‘Lena’ images. Our approach is seen to provide fewer artifacts as well as better preservation of edges and other details (e.g. the texture pattern on the eyelash).

Acknowledgements

This work is supported by National High Technology Research and Development Program of China (2007AA01Z168), National Nature Science Foundation of China (60805041, 60872082, 60773016) and Doctoral Foundations of Ministry of Education of China (20050004001, 200800041049).

References

[1] S.C. Zhu and D. Mumford, “Prior learning and gibbs reaction-diffusion,” IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.11, pp.1236–1250, 1997.
[2] S. Roth and M.J. Black, “Fields of experts: A framework for learning image priors,” IEEE Conf. on Computer Vision and Pattern Recognition, vol.2, pp.860–867, 2005.
[3] G. Hinton, “Training products of experts by minimizing contrastive divergence,” Neural Comput., vol.14, no.7, pp.1771–1800, 2002.
[4] M. Welling, G. Hinton, and S. Osindero, “Learning sparse topographic representations with products of Student-t distributions,” Advances in Neural Information Processing Systems, pp.1359–1366, 2003.
[5] J. Karhunen, E. Oja, L. Wang, R. Vigrio, and J. Joutsensalo, “A class of neural networks for independent component analysis,” IEEE Trans. Neural Netw., vol.8, no.3, pp.486–504, 1997.
[6] A.J. Bell and T.J. Sejnowski, “Edges are the independent components of natural scenes,” Advances in Neural Information Processing Systems, pp.831–837, 1996.
[7] Y. Weiss and W.T. Freeman, “What makes a good model of natural images?,” IEEE Conf. on Computer Vision and Pattern Recognition, pp.1–8, 2007.
[8] C.K.I. Williams and F.V. Agakov, “Products of gaussians and probabilistic minor component analysis,” Neural Comput., vol.14, no.5, pp.1169–1182, 2002.
[9] J. Domke, A. Karapurkar, and Y. Aloimonos, “Who killed the directed model?,” IEEE Conf. on Computer Vision and Pattern Recognition, pp.1–8, 2008.
[10] B.A. Olshausen and D.J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?,” Vision Research, vol.37, pp.3311–3325, 1997.
[11] J. Portilla, V. Strela, M. Wainwright, and E.P. Simoncelli, “Image denoising using scale mixtures of gaussians in the wavelet domain,” IEEE Trans. Image Process., vol.12, no.11, pp.1338–1351, 2003.
[12] A. Hyvarinen, R. Cristescu, and E. Oja, “A fast algorithm for estimating overcomplete ICA bases for image windows,” Proc. Int. Joint Conf. on Neural Networks, pp.894–899, 1999.