Learning generative texture models with extended Fields-of-Experts

Nicolas Heess[1]
n.m.o.heess@sms.ed.ac.uk
Christopher K.I. Williams[1]
c.k.i.williams@inf.ed.ac.uk
Geoffrey E. Hinton[2]
hinton@cs.toronto.edu

[1] University of Edinburgh
School of Informatics
Edinburgh, UK

[2] Department of Computer Science
University of Toronto,
Toronto, Canada

Abstract

We evaluate the ability of the popular Field-of-Experts (FoE) to model structure in images. As a test case we focus on modeling synthetic and natural textures. We find that even for modeling single textures, the FoE provides insufficient flexibility to learn good generative models – it does not perform any better than the much simpler Gaussian FoE. We propose an extended version of the FoE (allowing for bimodal potentials) and demonstrate that this novel formulation, when trained with a better approximation of the likelihood gradient, gives rise to a more powerful generative model of specific visual structure that produces significantly better results for the texture task.

1 Introduction

Much effort has been devoted to the development of prior models of generic image structure. Such models are important for many image processing and synthesis tasks, and as a building block of more comprehensive probabilistic models of natural scenes. One successful example is the Field-of-Experts (FoE) framework, recently proposed by Roth and Black [9]. The FoE defines a probability distribution over images in the form of a homogeneous high-order Markov random field (MRF). The clique potentials of this MRF are defined in terms of the responses of linear filters. This MRF-based model is translation invariant and can be applied to images of arbitrary size. The model is parametric and all parameters can be learned from training data, so it can be directly adapted to the statistics of natural images. Several studies have demonstrated that the FoE performs very well in tasks which require a generic image prior, such as image denoising, inpainting and novel view synthesis [5, 9, 16]. However, while the FoE's suitability for such tasks is certainly encouraging, other results, such as the very smooth nature of samples drawn from a FoE model trained on natural image patches ([8], Fig. 4.9) and e.g. the analyses of related models ([10, 14]; c.f. section 2.1 below) suggest that the FoE might still be a relatively limited model of natural images and accounts predominantly for their piecewise smoothness.

In this paper we evaluate the FoE's generative power and argue for a more structured approach to image modeling. Natural images are extremely complex as can be appreciated by considering the image shown in Fig. 1a. The image contains different regions with very

different visual characteristics. Attempting to learn these different characteristics with a single, generic model will most likely lead to the model learning only the most generic properties – such as piecewise smoothness in the case of the FoE [10, 14]. While this might be sufficient for certain image processing tasks (such as denoising or simple inpainting), it is not very satisfying in terms of a more comprehensive model of natural image structure. As an alternative approach we therefore propose to (a) focus on models that are good at capturing *specific* structure in natural images and (b) use these models as building blocks of more comprehensive, hierarchical models that can then account for more complex properties of natural images. For example, a number of texture models can be composed together to model an image comprised of multiple regions such as the one in Fig. 1a. This can, for instance, be achieved using an additional discrete-valued random field which switches between the different texture models. Melas and Wilson [6] used such a scheme although their individual texture models were limited to be Gaussian random fields which are rather weak models of textures (and natural image structure in general). Suitable "component models" are clearly an important prerequisite for such hierarchical models. However, many of the most powerful methods for generating specific structure that have been proposed in the past (e.g. [2, 13]), are not formulated as probabilistic models and it is therefore not clear how they could be used in this context.

Below we first demonstrate that the FoE in its original form as a generic image model suffers from similar limitations to Gaussian random fields when it comes to modeling specific structures. It is not able to model single textures and does not perform any better than the simpler Gaussian FoE on this task. We provide an intuitive explanation as to why this is the case. We then develop a model that can be used as described in the previous paragraph: A modification of the model structure and of the learning algorithm can substantially increase the generative power, giving rise to a compact parametric model of textures (the FoE with bimodal potentials or BiFoE) that can be fully learned from training data and the performance of which is comparable with state-of-the-art nonparametric approaches such as [2] in our experiments. The rest of the paper is structured as follows: In section 2 we describe in detail the FoE as proposed by Roth and Black [9], we discuss related models and we present our extension that allows modeling individual visual textures. Section 3 contains the experimental evaluation of these models. We assess the generative power of the models on a texture synthesis and on an inpainting task. We provide some insight into the large differences in performance of the FoE and BiFoE in section 4 and conclude with a discussion in section 5.

2 Models

2.1 Field of Experts

The FoE is an extension of the Products-of-Experts (PoE) model [4]. In the PoE framework high-dimensional probability distributions are modeled by taking the product of several distributions (the experts), each of which may be defined on a low-dimensional subspace of the data. In the case of images, a one-dimensional subspace is typically used. Thus, considering an image $\boldsymbol{x}$ as a vector of length N, i.e. $\boldsymbol{x} \in \mathbb{R}^N$, each expert distribution is defined on $\boldsymbol{w}_j^T \boldsymbol{x}$ where $\boldsymbol{w}_j$ defines the subspace of expert $j = 1 \dots M$ (where M is the number of experts). Thus $p(\boldsymbol{x}) = \frac{1}{Z(\Theta)} \prod_{j=1}^{M} \Phi(\boldsymbol{w}_j^T \boldsymbol{x}; \boldsymbol{\alpha}_j)$ where $\Phi(y; \boldsymbol{\alpha}_j)$ is a nonlinear expert function with parameters $\boldsymbol{\alpha}_j$ (typically an unnormalized 1D density function) and Θ is the set of parameters of the model ($\boldsymbol{w}_j$s and $\boldsymbol{\alpha}_j$s); $Z(\Theta) = \int p(\boldsymbol{x}) d\boldsymbol{x}$ is the normalization constant.

In the PoE, $\boldsymbol{w}_j$ has the same size as the image and a PoE is therefore typically limited to small images (image patches). In contrast, in the FoE the $\boldsymbol{w}_j$s are much smaller than the image, and the experts are replicated at each pixel. This allows the application of FoEs to images of arbitrary size while keeping the number of parameters low. Since the experts are replicated at each pixel, the $\boldsymbol{w}_j$s effectively act as linear filters and the FoE is thus a homogeneous high-order Markov random field (MRF) with clique potentials defined in terms of the responses of these linear filters:

$$p(\boldsymbol{x}) = \frac{1}{Z(\Theta)} \prod_{i=1}^{N} \prod_{j=1}^{M} \Phi(\boldsymbol{w}_j^T \boldsymbol{x}_{(i)}; \boldsymbol{\alpha}_j). \tag{1}$$

Here, the index i runs over the pixels in the image, and $\boldsymbol{x}_{(i)}$ is the image patch of the size of filter $\boldsymbol{w}_j$ centered at pixel i.

Roth & Black choose $\Phi(y;\boldsymbol{\alpha})$ to be the one dimensional Student-t potential, i.e. $\Phi(y;v) = (1+\frac{1}{2}y^2)^{-v}$ (with $v>0$) so that (1) can be written in terms of the energy as:

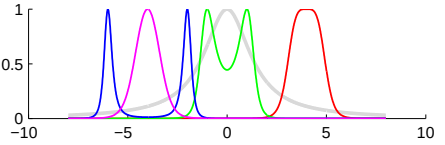
$$p_{FoE}(\boldsymbol{x}) = \frac{1}{Z(\Theta)} \exp(-E_{FoE}(\boldsymbol{x})); \qquad E_{FoE}(\boldsymbol{x}) = \sum_i \sum_j v_j \log\left\{1 + \frac{1}{2}\left(\boldsymbol{w}_j^T \boldsymbol{x}_{(i)}\right)^2\right\}. \tag{2}$$

The choice of the Student-t potential is motivated by the fact that responses of linear filters to natural images typically exhibit highly kurtotic response distributions. Student-t potentials were previously used by Welling et al. [15] in a PoE model of image patches. Sparsity priors have further been proven to be effective in non-translation invariant, directed models of image patches (e.g. [7]).

The gradient of the FoE log likelihood with respect to the parameters v_j and $\boldsymbol{w}_j$ (generically denoted as θ_j below) is given by $\frac{\partial \mathscr{L}(X;\Theta)}{\partial \theta_j} = -\left\langle \frac{\partial E_{FoE}(\boldsymbol{x};\Theta)}{\partial \theta_j} \right\rangle^+ + \left\langle \frac{\partial E_{FoE}(\boldsymbol{x};\Theta)}{\partial \theta_j} \right\rangle^-$ and consists of two terms: the expectation of the gradient of the energy over the data distribution $\langle\cdot\rangle^+$ as well as over the model distribution $\langle\cdot\rangle^-$ (given the current model parameters Θ). The second term cannot be computed analytically. Roth & Black therefore propose to learn the FoE using contrastive divergence (CD; [3, 4]). CD approximates the gradient by replacing $p_{FoE}(\boldsymbol{x};\Theta)$ with $p_T(\boldsymbol{x};\Theta)$ which is obtained by initializing the Markov chain at the data and running MCMC for only a small number of steps T. Roth & Black choose $T=1$ in their experiments. CD is typically used with stochastic gradient ascent (SGA) and mini-batches, i.e. the parameters are updated after computing the gradient for a small subset of the training data.

Related Work: Although the FoE itself is not easily interpretable, several related models have shed some light on the underlying computational mechanisms. Choosing $\Phi(y) = \exp\{-\frac{1}{2}(y+b)^2\}$ gives rise to the structurally similar but analytically tractable Gaussian Field of Experts (GFoE) with the energy $E_{GFoE}(\boldsymbol{x}) = \frac{1}{2}\sum_i \sum_j \left(\boldsymbol{w}_j^T \boldsymbol{x}_{(i)} + b_j\right)^2$. The GFoE defines a Gaussian distribution over the set of images, i.e. $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$ with $\Sigma = (\sum_j W_j^T W_j)^{-1}$ and $\boldsymbol{\mu} = -(\sum_j W_j^T W_j)^{-1}(\sum_j W_j^T \mathbf{1} b_j)$. Here, W_j is the convolution matrix corresponding to $\boldsymbol{w}_j$ and $\mathbf{1}$ is a vector of ones. The GFoE is a Gaussian MRF (GMRF) model; these have been used for texture modelling for many years, see e.g. [1]. Unlike for the FoE described above, the likelihood of the GFoE and its gradient can be computed exactly. Although the GFoE is structurally similar to the FoE it is well known that it models only the power spectrum of an image and not the phases so its ability to capture image structure is very limited. With this in mind we will use it as a baseline model in our experiments below.

(a)



(b)

Figure 1: (a) Natural images typically contain multiple regions with very different visual characteristics. (b) Different BiFoE potentials. Red: $a = 0$, $b = -4$; Green: $a = -1$, $b = 0$; Blue: $a = -4$, $b = -4$; Magenta: $a = 1$, $b = 4$. Gray: Student-t potential as used in the FoE. $v = 1$ in all cases.

Analyzing a FoE in which the Student-t potential was replaced by a Gaussian scale mixture, made up of a finite number of Gaussians at different scales, Weiss and Freeman [14] recently demonstrated that one would expect the model to learn filters that effectively correspond to whitening filters. Along similar lines Tappen [10] demonstrated the similarity between the FoE and a MRF designed around a fixed set of derivative filters, suggesting that the FoE can be seen as imposing piecewise higher-order continuity on images. Both these results are consistent with the smooth nature of the samples that are generated by a FoE trained on natural images, supporting the idea that the FoE primarily models smoothness constraints with a robust loss function. Furthermore, Zhu and Mumford [17] and Zhu et al. [18] have proposed a model which is based on fixed filters but the potential function is non-parametric, so that there are no restrictions on the shape of the learned potentials. They find that decaying potentials (such as the Student-t potential) are not sufficient in order to obtain good generative models, which raises the question of how the choice of a particular parametric form for the potential function (such as the Student-t potential) affects the expressiveness of the FoE.

2.2 Extended FoE with bimodal potentials

In our experiments we find that the FoE in the form presented by Roth and Black [9] is not able to model natural textures (cf. section 3.2). One possible explanation is that the particular expert function used by Roth & Black is too restrictive. Indeed, we have shown that the distribution defined by the Student-t FoE is unimodal (see Sec. A in the supplementary material) and the findings in section 3.2 and 4 suggest that this might severly affect the generative power of the model. We therefore propose an extension of the original FoE model which allows for bimodal expert functions $\Phi_{Bi}(y;v,a,b) = \left\{1 + \frac{1}{2}\left[(y+b)^2 + a\right]^2\right\}^{-v}$. This choice of Φ gives rise to the following energy:

$$E_{Bi}(\boldsymbol{x}) = \sum_i \sum_j v_j \log \left\{1 + \frac{1}{2}\left[\left(\boldsymbol{w}_j^T \boldsymbol{x}_{(i)} + b_j\right)^2 + a_j\right]^2\right\}.$$

Figure 1b shows the shape of the potential for different settings of the parameters a and b. Φ_{Bi} is bimodal for $a < 0$; b determines the center of the potential. As for the FoE, samples from the model can be obtained by performing MCMC. One efficient way of doing this is hybrid Monte Carlo (HMC), as proposed for the FoE in [9]. Learning, too, could be performed as for the FoE but we found that the approximation of the gradient provided by CD-1 is not sufficient for training this more powerful model. We therefore use an improvement suggested by Tieleman and Hinton [11], Tieleman [12] in the context of Restricted

Boltzmann machines: $p_T(\boldsymbol{x};\Theta)$ is replaced by a set of K *persistent* Markov chains that are initialized at the data *at the beginning of learning* and then updated by one MCMC step for each mini-batch of SGA (instead of re-initializing them at the data for each mini-batch as in CD). One interpretation of this approach is that as long as the model parameters change slowly the persistent Markov chains will provide a K-sample approximation of the current model distribution. Since the learning procedure raises the energies of the current states of the persitent Markov chains, they escape local minima rapidly [11].

3 Experiments

Sections 3.2 and 3.3 describe experiments that evaluate the generative power of the FoE for natural image structure on a texture synthesis task and on a texture inpainting task. We compare the FoE with our extended model (BiFoE) and (as a baseline) with the much simpler GFoE. In section 3.3 we further compare the BiFoE with Efros & Leung's [2] nonparametric texture synthesis method. The dataset used in the experiments and learning of the models is described in section 3.1.

3.1 Setup: Dataset and Learning

Data : We use a range of six textures (Brodatz) as well as two synthetic patterns. All textures were chosen so as to be at reasonable scale given practical filter sizes for the FoE. Digitized versions of the Brodatz textures were downloaded from the web[1] and scaled by a factor 0.75 or 0.5 (preserving all major features of the textures). Examples of the (scaled) textures are shown in Figure 2a. We further generated two sets of synthetic texton patterns by placing white circles and crosses randomly on a black background (the diameter of the circles was 9 pixels, the bars of the crosses were 2×10 pixels; rightmost panels in Fig. 2a).

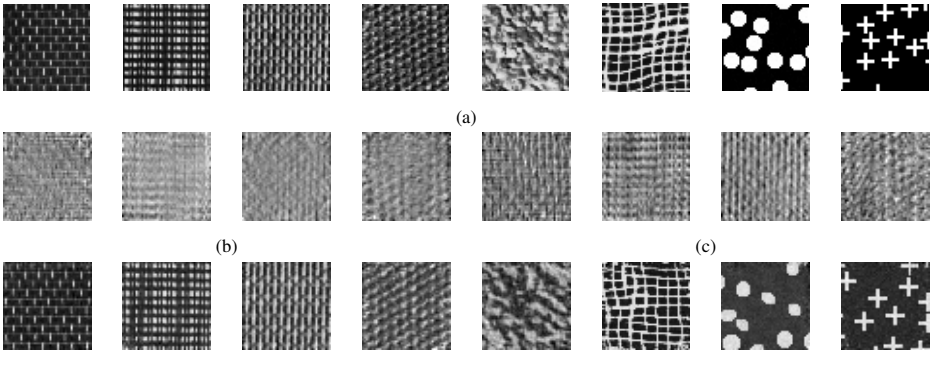


Figure 2: (a) Training data: 50×50 patches of the textures used in our experiments; left to right: D6: woven aluminium wire; D21: french canvas; D53: oriental straw cloth; D77: cotton canvas; D4: pressed cork; D103: loose burlap; circle textons; cross textons. (b) FoE samples: 50×50 samples cropped to 46×46 (to remove boundary pixels) of D6, D21, D53, and D77 from FoE models learned for these textures. (c) as (b) but for the GFoE models. (d) BiFoE samples: samples of all textures in (a) drawn from the BiFoE models.

[1] `http://www.ux.uis.no/~tranden/brodatz.html`

Learning: It was not possible to learn BiFoE models with standard CD. To compare models on equal footing we trained both the basic FoE as well as the extended FoE using persistent chains [12]. Learning was performed with SGA with momentum ($\nu = 0.9$). We used 500 25×25 pixel patches, cropped randomly from the texture images of size 480×480 or 320×320 pixels (after scaling). The data was scaled to have overall mean zero and a std.-dev. of 1. The training data was presented in 5 mini batches containing 100 data points each. At the beginning of learning the persistent chains were initialized at the data points in the first mini batch and updated by one "step" of HMC-MCMC[2] for each minibatch. The learning rate was set to $\eta = 0.0001$ and was held fixed for the FoE. For the BiFoE we obtained better results for some textures by reducing η over the course of learning (according to a linear schedule) and for some textures it also seemed advantageous to regularly restart a fraction of the Markov chains. We experimented with different sizes and numbers of filters. Larger and more filters typically improved the visual quality of the models slightly (the maximum size of filters that we tried was 7×7 pixels, the maximum number was $M = 9$ or $M = 15$) but for a reasonable range of values for these two parameters we did not observe large differences in model quality. Unless noted otherwise results shown are obtained with models with $M = 9$ filters of 7×7 pixels. GFoE models were also trained with SGA (same number of mini batches as for the FoE models), but using the exact gradient and with a (fixed) learning rate of 0.001.

3.2 Comparison of the generative power of GFoE, FoE, and BiFoE

The most direct way to evaluate a FoE/GFoE/BiFoE model is to measure the likelihood of a set of image patches under the model. However, the intractability of the partition function $Z(\Theta)$ means that this is not possible. Our main method of evaluation is thus to draw samples from the models and compare these samples with the original textures. In the next section we further discuss texture inpainting experiments. Image denoising has also been used as a test of generic image priors but this is a rather indirect measure of model quality; we do not pursue this here as the above methods are more useful given our modeling goal.

Samples were generated by initializing Markov chains with IID Gaussian noise $N(0,1)$, and performing HMC-MCMC until the chains had settled to equilibrium. For the GFoE samples were drawn from the corresponding Gaussian distribution. For a quantiative assessment of model quality we used a texture similarity score (TSS) based on normalized cross correlation (NCC): Specifically we sampled 100 texture patches of size 25×25 pixels from the models. In order to reduce the influence of boundary pixels we discarded 3 pixels on all four sides, resulting in patches of size 19×19 pixels. We then computed the NCC with the original texture image and used the maximum value across the image as the similarity score for the respective texture sample: $S(\boldsymbol{x}, \boldsymbol{s}) = \max_i \frac{\boldsymbol{x}_{(i)}^T \boldsymbol{s}}{\|\boldsymbol{x}_{(i)}\| \|\boldsymbol{s}\|}$ where $\boldsymbol{s}$ is the texture sample drawn from the model and $\boldsymbol{x}$ is the original texture image. TSSs were computed only for the first four textures in Fig. 2a (D6, D21, D53, and D77), as these four textures were sufficiently regular for the correlation-based similarity score to be meaningful. Below we present statistics of the distributions of TSSs for each model. This score is certainly not perfect, but it corresponds reasonably well with impressions from visual inspection and allows comparisons using a large set of samples from each model. For visual inspection we also generated larger samples (50×50 pixels).

[2] A "step" consisted of choosing a random momentum $N(\mathbf{0}, \boldsymbol{I})$ followed by 30 leapfrog-steps.

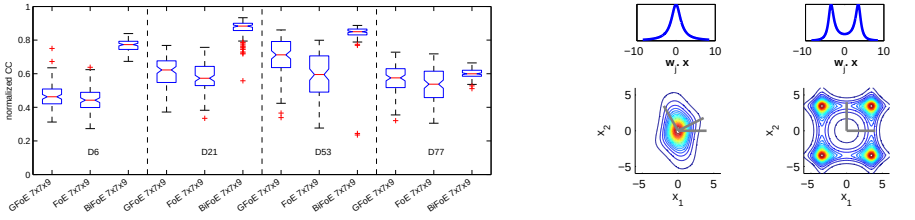


Figure 3: (a) TSSs for the four textures D6, D21, D53, and D77 for the three models considered. The three columns for each texture correspond to (from left to right) the GFoE, FoE, and BiFoE respectively. Boxes indicate the upper and lower quartiles as well as the median (red bar) of the TSS distributions; whiskers show extent of the rest of the data; red crosses: outliers. (100 data points in each case). (b) Contour plot of the pdf of an (overcomplete) PoE model in $\mathbb{R}^2$ obtained from three Student-t experts with zero centered potentials (the expert function is shown above; $\boldsymbol{w}_{1\ldots3}$ are superimposed on the contour in dark gray). (c) Same as in (b) but for two bimodal experts. The pdf in (b) is unimodal, the one in (c) multimodal.

Fig. 3a shows boxplots of correlation scores for the three models for the first four textures in Fig. 2a. Fig. 2b-d show representative samples of the four textures drawn from the FoE, GFoE, and BiFoE models respectively. From the low correlation scores and from comparing samples in Fig. 2b,c with the originals in Fig. 2a it is clear that GFoE and FoE models fail equally to reproduce three of the four textures. For D77 the GFoE / FoE models were able to produce reasonable samples, although not very consistently. Results for the BiFoE are rather different: For the first three textures the TSSs are clearly higher than for the GFoE / FoE. The difference for D77 is smaller, but the quality of the samples from the BiFoE is much more consistent than for the GFoE or FoE models and visually the samples are much more convincing. Fig. 2d shows representative 50×50 samples drawn from the BiFoE model distributions for all 6 textures described above. Especially for the first 5 textures it is difficult to distinguish samples drawn from the model from the original textures. The texton patterns are also modeled well (although the sample quality is somewhat less consistent). For the texton models shown in Figure 2d we used $M = 15$ 7×7 filters as this gave better results. Clearly, the BiFoE is able to model not only the relatively regular textures (D6, D21, D53, D77) but can also handle more variable ones. The fact that we can draw samples that are larger than the patches used for training (twice as large in Fig. 2) confirms that the model is not just memorizing the training data.

3.3 Constrained texture synthesis: Texture Inpainting

We further evaluated the models on a texture inpainting task. This requires the models to generate a texture that is consistent with the given part of the image ("inpainting frame"). Since this imposes constraints on the randomness of the generated texture we can, for sufficiently regular textures, compare the sampled pixels directly to the corresponding region in the original image. For our experiments we used 70×70 texture images with a 50×50 hole in the center. In addition to the probabilistic models discussed in the previous section we also included the synthesis method proposed by Efros and Leung [2] in this experiment

in order to compare the BiFoE to a state-of-the-art non-parametric approach[3].

For the GFoE, FoE, and BiFoE we sampled the "missing" pixels conditioned on the "existing" pixels with HMC-MCMC. Our implementation of Efros & Leung's (E&L) method used 15×15 pixel patches (referred to as "neighborhood windows" in [2]) for infilling extracted from those image patches used to train the BiFoE models; we experimented with different patch sizes for E&L's method and a size 15 seemed to give good results. We used 20 different texture images for each texture, and repeated inpainting 5 times for each image since all methods are stochastic. The quality of the results across repetitions for a given texture image was typically very similar. We performed inpainting for the four regular textures and computed the NCC between the original texture image and the inpainting result (cf. section 3.2). Fig. 4 shows representative results for texture D21, the remaining results can be found in the supplementary material (Fig. S-4). The NCC values (averaged over repetitions and images for each texture and method) for the four regular textures are given in the table below (NCC $\pm$ std-dev):

	GFoE	FoE	BiFoE	Efros & Leung
D6	0.7245 ± 0.0261	0.6686 ± 0.0385	0.8769 ± 0.0163	0.8300 ± 0.0380
D21	0.7862 ± 0.0237	0.7971 ± 0.0283	0.8653 ± 0.0244	0.8330 ± 0.0351
D53	0.7736 ± 0.0208	0.7808 ± 0.0159	0.9145 ± 0.0125	0.8878 ± 0.0300
D77	0.5675 ± 0.0286	0.6102 ± 0.0229	0.6567 ± 0.0205	0.6325 ± 0.0490

Two observations can be made: Firstly, the GFoE & FoE perform better than would be predicted from the results in the previous section. If provided with a reference (the inpainting frame), they can "maintain" the corresponding structure over a certain distance, although the quality of the texture decreases as the distance to the closest reference pixels increases. This can be seen in Figures 4 and S-4: the textures generated by the GFoE and FoE models are quite blurry in the center of the image. For the GFoE this behavior is expected: It models the power spectrum of an image while ignoring the phases, but in the inpainting case the phases are to some extent imposed upon the generated texture by the inpainting frame. Secondly, BiFoE and E&L's method both perform very well on almost all textures - and considerably better than the GFoE / FoE. They do not suffer from the degradation of the texture towards the center of the completed image. The BiFoE seems to perform even slightly better than E&L's method, but more importantly, in contrast to the latter the BiFoE is formulated as an explicit parametric generative model that absorbs the characteristics of a texture into a compact representation (only 9 filters of size 7×7 plus 9×3 parameters for the experts' potentials).

4 Understanding the differences between the models

From the results presented above it is clear that the BiFoE learns texture models that are markedly superior to the ones learned by the FoE and the GFoE, but there seems to be little difference between the FoE and the GFoE, suggesting that they suffer from similar deficiencies (such as the GFoE's inability to model phases) for the task at hand. Some insight into how these are overcome by the BiFoE can be gained by inspecting the BiFoE parameters: All BiFoE models learn several experts with bimodal expert functions (i.e. $a_j < 0$) and the

[3] We only compared with Efros & Leung's method on the inpainting task because this method requires a "seed" for the synthesized texture which is naturally given by the inpainting frame in the inpainting task.

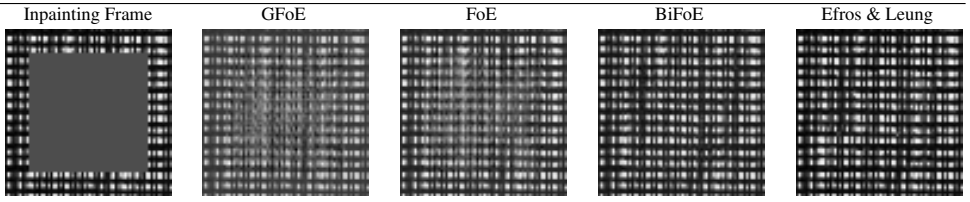


Figure 4: Representative inpainting results for texture D21: The first panel shows the inpainting frame (see supplementary material, Fig. S-4, for the original image). The remaining four panels show texture images completed with the GFoE, FoE, BiFoE and with Efros & Leung's method. Note how the reconstructions obtained with FoE & GFoE become increasingly blurred towards the image center.

interactions between the learned bimodal experts give rise to heavily skewed and in many cases also bimodal response marginals with modes offset from 0. The learned BiFoE models are thus very different from the FoE (and GFoE) models for which the response marginals are almost exclusively centered at zero and roughly symmetric. The supplementary material (Fig. S-3) contains a revealing 1D example that illustrates how the BiFoE uses bimodal potential functions to model patterns that cannot be modeled using the (G)FoE potential functions. The implications of these findings in terms of the resulting probability distributions over images are best understood by considering the possible interactions between the experts and the different types of potential functions. For the GFoE the pdf arising from the replicated experts will always be Gaussian (cf. section 2.1) and is thus inherently unimodal. For the FoE in the form of equation 2 this pdf can take more interesting forms, however, because all expert potentials are centered at zero, the overall pdf will still always be unimodal independent of the number of experts and the parameters they learn. It can be shown that the energy function always has a unique minimum at $\boldsymbol{x} = \mathbf{0}$ (as long as the overall model is complete; proof in supplementary material, Sec. A). The potentials of the BiFoE, in contrast, allow for much more flexibility in shaping a potentially multimodal pdf. This idea is illustrated in Figure 3b,c for the non-translation invariant (PoE) case (see also [4]). For the translation-invariant FoE the situation is more complex but the same considerations hold.

5 Discussion

We have investigated the FoE's ability to model specific visual structure. Our results suggest that in its basic form the FoE is a rather limited model: Although the filters are learned from data, the zero-centered Student-t potential is too restrictive to model even individual textures. We have further demonstrated that introducing more complex bimodal potentials, and using a better learning strategy, gives rise to a considerably more powerful model. The interactions of multiple bimodal experts can flexibly shape the density allowing the BiFoE to learn good generative models of the textures that we considered.

The results above were obtained with a particular form of a bimodal expert. One interesting question is whether this particular form is crucial. One possibility might be to replace the bimodal experts with a mixture of two Gaussians with shared $\boldsymbol{w}_j$ but different biases. Alternatively it might be possible to achieve similar results even with the Student-t potentials when bias terms are included, so that the potentials are no longer necessarily centered at zero (i.e. $\Phi(y;v,b) = (1+\frac{1}{2}(y+b)^2)^{-v}$; note that the bs are missing in equation 2 as they are in Roth and Black [9]; but see Woodford et al. [16]). This formulation can give rise to globally

multimodal distributions. To get two uncentered Student-t potentials to act as a bimodal potential would require, however, that two experts learn the same filters and ν-parameters but different bs. In experiments with 1D patterns which require such bimodal potentials and for which the BiFoE learns a good model without difficulties we found that the FoE with bias terms fails in exactly the same way as the basic FoE (all learned b_js were effectively zero); stable convergence to a bimodal solution (and thus learning a reasonable model) appears very difficult to achieve unless the learning is effectively initialized with the correct solution. For natural textures full bimodality of the potential function may not always be required but preliminary experiments suggest that, here too, the BiFoE is considerably more robust. We are currently investigating this in more detail.

Our investigations were motivated by the idea of a more structured approach to natural image modeling, and the BiFoE is a probabilistic generative model that would be a suitable building block in this context. The approach receives further support from the observation that when trained on natural images the filters learned by the BiFoE were qualitatively similar to the ones learned by the basic FoE, and the expert functions were exclusively unimodal and centered at zero. This suggests that, although the BiFoE is a good generative model for specific visual structures, when faced with the task of modeling too heterogeneous a set of patterns (thinking about the structure in a database of natural images as a very large mixture of different textures) it is still not powerful enough and like the basic FoE accounts only for very generic properties such as smoothness (cf. section 2.1). We are currently investigating various possibilities for making the model more powerful. In particular, we are investigating hierarchical formulations of the BiFoE in which the parameters of the BiFoE are modulated by the state of higher-level latent variables to allow efficient learning of multiple textures as well as to model spatially inhomogeneous data such as images with texture boundaries (cf. section 1 and [6]) and images with smoothly varying texture gradients.

Acknowledgements

This work is supported in part by the EU NoE PASCAL2. CW is an Associate Member of the CIfAR NCAP programme. NH is supported by an EPSRC/MRC scholarship from the Neuroinformatics and Computational Neuroscience DTC at the University of Edinburgh. GEH was supported by grants from NSERC and CIfAR.

References

[1] R. Chellappa, S. Chatterjee, and R. Bagdazian. Texture Synthesis and Compression using Gaussian-Markov Random Field Models. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:298–303, 1985.

[2] Alexei A. Efros and Thomas K. Leung. Texture Synthesis by Non-parametric Sampling. In *IEEE International Conference on Computer Vision*, pages 1033–1038, Corfu, Greece, September 1999.

[3] G.E. Hinton, S. Osindero, M. Welling, and Y. Teh. Unsupervised Discovery of Nonlinear Structure using Contrastive Backpropagation. *Cognitive Science*, 30(4):725–731, 2006.

[4] Geoffrey E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Comp.*, 14(8):1771–1800, August 2002.

[5] J.J. McAuley, T.S. Caetano, A.J. Smola, and M.O. Franz. Learning high-order MRF priors of color images. In *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, pages 617–624, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2.

[6] D.E. Melas and S.P. Wilson. Double Markov random fields and Bayesian image segmentation. *IEEE Transactions on Signal Processing*, 50(2):357–365, Feb 2002.

[7] B.A. Olshausen and D.J. Field. Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vision Research*, 37:3311–3325(15), 1997.

[8] S. Roth. *High-Order Markov Random Fields for Low-Level Vision*. PhD thesis, Brown University, 2007.

[9] S. Roth and M.J. Black. Fields of Experts: a framework for learning image priors. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2: 860–867 vol. 2, June 2005. ISSN 1063-6919.

[10] M.F. Tappen. Utilizing Variational Optimization to Learn Markov Random Fields. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[11] T. Tieleman and G.E. Hinton. Using Fast Weights to Improve Persistent Contrastive Divergence. In *ICML '09: Proceedings of the 26th International Conference on Machine Learning*, pages 1033–1040. ACM New York, NY, USA, 2009.

[12] Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML '08: Proceedings of the 25th International Conference on Machine Learning*, pages 1064–1071. ACM New York, NY, USA, 2008.

[13] Li-Yi Wei and Marc Levoy. Fast texture synthesis using tree-structured vector quantization. In *SIGGRAPH '00: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 479–488, 2000.

[14] Y. Weiss and W.T. Freeman. What makes a good model of natural images? *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[15] M. Welling, G. Hinton, and S. Osindero. Learning Sparse Topographic Representations with Products of Student-t Distributions. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1359–1366. MIT Press, Cambridge, MA, 2003.

[16] O.J. Woodford, I.D. Reid, P.H. Torr, and A.W. Fitzgibbon. Fields of Experts for Image-based Rendering. In *BMCV*, 2006.

[17] Song Chun Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11):1236–1250, Nov 1997. ISSN 0162-8828.

[18] Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, Random Fields and Maximum Entropy (FRAME): Towards a unified theory for texture modeling. *Int. J. Comput. Vision*, 27(2):107–126, 1998. ISSN 0920-5691.