Learning High-Order MRF Priors of Color Images

Julian J. McAuley   J.MCAULEY@STUDENT.UNSW.EDU.AU

National ICT Australia, Canberra, ACT 0200, Australia
University of New South Wales, Sydney, NSW 2052, Australia

Tibério S. Caetano   TIBERIO.CAETANO@NICTA.COM.AU
Alex J. Smola   ALEX.SMOLA@NICTA.COM.AU

National ICT Australia, Canberra, ACT 0200, Australia
RSISE, Australian National University, Canberra, ACT 0200, Australia

Matthias O. Franz   MATTHIAS.FRANZ@TUEBINGEN.MPG.DE

Max Planck Institute for Biological Cybernetics, Tuebingen 72012, Germany

Abstract

In this paper, we use large neighborhood Markov random fields to learn rich prior models of color images. Our approach extends the monochromatic *Fields of Experts* model (Roth & Black, 2005a) to color images. In the *Fields of Experts* model, the curse of dimensionality due to very large clique sizes is circumvented by parameterizing the potential functions according to a product of experts. We introduce simplifications to the original approach by Roth and Black which allow us to cope with the increased clique size (typically 3x3x3 or 5x5x3 pixels) of color images. Experimental results are presented for image denoising which evidence improvements over state-of-the-art monochromatic image priors.

1. Introduction

The problem of characterizing a prior distribution of natural images is at the foundation of many low-level vision problems such as image denoising, super-resolution, inpainting (i.e., filling-in of missing image regions), as well as stereo disparity estimation. Whenever a "scene" must be inferred from noisy, degraded or partial image information, a natural image prior is required (Freeman et al., 2000). The inferred scene should be consistent with the relevant information in

the deteriorated input image data, but at the same time it should make sense visually. Examples of a "scene" may be a denoised image (inferred from a noisy one), a depth map (inferred from local stereo disparities), or a high-resolution image (inferred from a low-resolution one). Possibly the simplest example of an image prior is pairwise smoothness: neighboring pixels are much more likely to have similar grey values than very different grey values. This is a typical definition of how natural scenes "should behave".

In a recent paper, Roth and Black (Roth & Black, 2005a) introduced a generic Markov random field image prior over extended neighborhoods—the *Fields of Experts* model—which is capable of representing a richer prior structure for images than pairwise smoothness. In order to cope with the large clique sizes involved, they parameterize the potential functions in terms of filters and associated coefficients. Learning in the high dimensional space of the clique then becomes merely learning the filters and the coefficients. With the resulting prior image model, they present results for image denoising and inpainting at the level of the current state-of-the-art. In a subsequent paper, the same authors obtained cutting-edge results on optical flow estimation by using the same parametric high-order Markov random field model (Roth & Black, 2005b).

In this paper we build on their results by generalizing the *Fields of Experts* model to a prior model for *color* images. This means that tasks like image denoising, inpainting or super-resolution do not need to be done independently for each channel, but can be performed directly over a multiband image in such a way that

the correlations between the different channels can be exploited by the model. As far as we know, this is the first time that a high-order Markov random field prior for color images has been reported. Although we use a naïve learning procedure, we present experimental results that show improvements of the color image prior over the state-of-the-art monochromatic prior reported in Roth and Black (2005a). We expect these improvements to become more significant after optimizing the learning algorithm.

2. Background

Natural scenes consist of surfaces that are often formed by homogeneous materials. As a result, the spectral and spatial structures of natural images are highly correlated. This fact has been confirmed by a number of colorimetric studies of natural scenes (Burton & Moorhead, 1987; Ruderman et al., 1998). These correlations constitute the basis of the color constancy capabilities of natural visual systems (Maloney, 1986). A straightforward idea is to use them for image processing tasks such as the denoising or super-resolution of color images. Here, we intend to develop an approach in the context of high-order Markov random fields that is capable of coding the localized spatial and spectral dependencies of color images in a very natural manner. To this end, we will extend the monochromatic model in Roth and Black (2005a) to multiband images.

In Roth and Black (2005a), the authors have merged the ideas of learning in Markov random fields (Zhu et al., 1998) and sparse image coding (Olshausen & Field, 1997) in order to develop a high-order Markov random field model where the cliques are square image patches of typically 3x3 and 5x5 pixels. The potential functions over these cliques are then assumed to be *products of experts* (Hinton, 1999), i.e. products of individual functions ϕ_f (with a parameter α_f) of the response of a filter J_f to the image patch $\mathbf{x}_c$:

$$\phi_c(\mathbf{x}_c; J, \alpha) = \prod_{f=1}^{F} \phi_f(\mathbf{x}_c; J_f, \alpha_f). \tag{1}$$

These are assumed to be stationary ($\phi_c := \phi$ henceforth), i.e. every clique in the image will have the same parameter vector $\theta = \{J_f, \alpha_f : 1 \leq f \leq F\}$.

The particular form they postulate for the expert is related to the Student-T distribution, and reads

$$\phi_f(\mathbf{x}_c; J_f, \alpha_f) = (1 + \frac{1}{2}\langle J_f, \mathbf{x}_c \rangle^2)^{-\alpha_f}. \tag{2}$$

It is interesting that, for positive α's, this expression has high value when the patch $\mathbf{x}_c$ and the filter J_f are *less* coincident, which is the opposite of typical definitions for potential functions. This is in principle not a problem, though, since the parameters will be learned accordingly (i.e. smaller α's will be associated to those filters which correspond to more likely configurations).

By invoking the Hammersley-Clifford theorem, which states that the joint probability distribution of a Markov random field with clique set $\mathcal{C}$ can be written as

$$p(\mathbf{x}) = \frac{1}{Z(\Theta)} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x}_c), \tag{3}$$

the model becomes

$$p(\mathbf{x}) = \frac{1}{Z(\Theta)} \prod_{c \in \mathcal{C}} \prod_{f=1}^{F} (1 + \frac{1}{2}\langle J_f, \mathbf{x}_c \rangle^2)^{-\alpha_f}. \tag{4}$$

Both the J_f's and α_f's are learned from a training set of grey images using contrastive divergence (for details, see Roth and Black (2005a)). Once the model is learned, inference is performed by gradient ascent on the log-posterior. Roth and Black reported state-of-the-art results for both denoising and inpainting with the obtained image prior.

3. The Color Image Prior

The contribution of this paper consists of extending the model of Eq. 4 to *color* images. In the following, we describe how this can be accomplished despite the significant increase in the dimensionality of the space in which learning must be performed.

3.1. Learning the Model

The assumptions made in our model are as follows: first, we assume the parametric form given in Eq. 4 as defining the class of probability distributions which we use to model the color image data. This itself defines an exponential family model. Second, we restrict ourselves to modeling potential functions $\phi(\mathbf{x}_c; J, \alpha)$ in cliques of sizes 3x3x3 and 5x5x3 pixels (the last digit refers to the 3 color channels—the model was trained on the RGB color space). Third, we choose a fixed basis of filters J_f's by performing a singular value decomposition of the covariance matrix of the training data. Finally, we select the parameters α by maximum-likelihood learning.

The strongest of these assumptions is the use of the

filters J as the eigenvectors of the covariance matrix. Ideally, the filters themselves should be learned by maximizing the likelihood. However, given the dimensionality of the cliques that arise in this color model—respectively 27 and 75 for the two clique sizes—we found this to be unfeasible in practice (at least in the 5x5 case). On the other hand, this assumption seems prudent given the lack of a feasible maximum-likelihood solution. Principal components describe the directions (color patches) in which real color patches in images vary the most. Eigenvalues give the relative variance across the different directions. Indeed, our results indicate that even under this assumption significant improvements are obtained over monochromatic image priors.

To determine our filters (J_f's), we randomly selected 50,000 3x3x3 and 5x5x3 patches, cropped from images in the Berkeley Segmentation Database (Martin et al., 2001), and performed a singular value decomposition on their covariance matrices. We ignored the component with highest variance, which corresponded to a uniform grey image (Roth & Black, 2005a)—this is needed to make our model invariant to intensity, and is found to give better results in practice (Welling et al., 2002).

Under this scenario, estimation of the α's in our model takes the following form: Let $\mathcal{D} = \{X_1, X_2 \ldots X_M\}$ be a set of training images, $p(\mathcal{D}|\Theta)$ be the likelihood of the training images given the model, and $\mathcal{R} = \{Y_1, Y_2 \ldots Y_N\}$ be a set of random images sampled from $p(\mathbf{x})$. Let $\Theta = \{\theta_1, \theta_2 \ldots \theta_F\} = \{(J_1, \alpha_1), (J_2, \alpha_2) \ldots (J_F, \alpha_F)\}$ be our set of filters, and their corresponding weights. Then the likelihood of our training images, $p(\mathcal{D}|\Theta)$ is given by

$$p(\mathcal{D}|\Theta) = \prod_{i=1}^{M} \frac{1}{Z(\Theta)} \prod_{c \in X_i} \phi(\mathbf{x}_c^i; J, \alpha), \qquad (5)$$

where an estimate of $Z(\Theta)$ is given by

$$Z(\Theta) \propto \frac{1}{N} \sum_{i=1}^{N} \prod_{c \in Y_i} \phi(\mathbf{x}_c^i; J, \alpha). \qquad (6)$$

From Eq. 5 we obtain the log-likelihood,

$$\log p(\mathcal{D}|\Theta) = \left(\sum_{i=1}^{M} \sum_{c \in X_i} \sum_{f=1}^{F} \alpha_f \psi\left(J_f, \mathbf{x}_c^i\right) \right) - M \log Z, \qquad (7)$$

where $\psi(a, b) = -\log(1 + \frac{1}{2} \langle a, b \rangle^2)$. Taking the derivative with respect to a particular α_k gives

$$\frac{\partial}{\partial \alpha_k} \log p(\mathcal{D}|\Theta) = \left(\sum_{i=1}^{M} \sum_{c \in X_i} \psi\left(J_k, \mathbf{x}_c^i\right) \right) - M \frac{\partial}{\partial \alpha_k} \log Z. \qquad (8)$$

Finally, the derivative of the log partition function is given by

$$\frac{\partial}{\partial \alpha_k} \log Z(\Theta) = \frac{\frac{\partial}{\partial \alpha_k} Z(\Theta)}{Z(\Theta)}$$

$$= \frac{\sum_{i=1}^{N} \left[\left(\sum_{c \in Y_i} \psi(J_k, \mathbf{x}_c) \right) \left(\prod_{c \in Y_i} \phi(\mathbf{x}_c; J, \alpha) \right) \right]}{\sum_{i=1}^{N} \prod_{c \in Y_i} \phi(\mathbf{x}_c; J, \alpha)}, \qquad (9)$$

and the update-equation for gradient ascent on the α's is just

$$\alpha \leftarrow \alpha + \mu \{ \frac{\partial}{\partial \alpha_i} \log p(\mathcal{D}|\Theta) \big| i \in (1 \ldots N) \} \qquad (10)$$

where μ is a user-specified learning rate.

As can be seen from the equations, we actually perform gradient ascent with respect to the α's only (Eqs. 5 – 10), which significantly simplifies the learning problem as previously mentioned. Since our α's had been initialized to zero, we have that $\phi(\mathbf{x}_c; J, \alpha) = 1$ and as a result the posterior distribution is truly flat at this stage. This allows us to compute the right-hand side of Eq. 9 by summing over N truly *random* images (i.e. the Y_i's are generated according to a uniform distribution), meaning that we do not need to perform CPU-intensive sampling. Such a result is only possible during the *first* iteration. After the first iteration, the α's are no longer zero, and as a result the posterior distribution is no longer flat. From this point on, estimating the partition function becomes computationally very intensive because of the need to run a sampler to obtain *each* one of the N samples needed to compute Eq. 9.

Fortunately, we experimentally found that by performing only *one* step of gradient ascent, it appeared that we had arrived at a maximum likelihood solution. This was found by noticing that, after the first iteration, only the absolute values of the α's changed, whereas their *relative* values appeared to remain approximately the same. This implies that the shape of the posterior distribution remains the same after the first iteration, and so there is no need to iterate further (since the

mode of the posterior remains the same). As a result, in order to train our model, we actually performed only one iteration of gradient-ascent. That is, we set the α's to be the gradient itself (Eq. 8), since their initial values are zero. Hence the learning rate for our gradient-ascent algorithm is in fact arbitrary, as it only serves to scale the resulting α's by a constant multiple. We did not rescale our α's after learning, and in fact found that the first α was negative. Although it may be necessary to set such α's to zero in order to make the potentials normalizable, we found that enforcing this condition made no difference to our experimental results.

Although this solution is clearly sub-optimal, it is the only way we are able to subvert the problem of dimensionality in our model. Notably in the 5x5 case, in which we have 74 filters, each with 75 dimensions, the time taken to estimate the derivative of the log-partition function through sampling would render our problem intractable. Instead, by avoiding sampling altogether, the learning problem becomes trivial in the 3x3x3 and 5x5x3 cases, and even remains feasible with larger clique sizes.

The images from which we learn these clique potentials are slightly larger than the cliques themselves, in order to capture the local correlations due to the superposition of cliques. The image sizes for 3x3x3 and 5x5x3 cliques are, respectively, 7x7x3 and 13x13x3. Ultimately, to train the α's, we randomly cropped 100,000 images of these sizes from the Berkeley Segmentation Database (Martin et al., 2001) and selected the corresponding cliques. A further 50,000 random images were used to estimate the derivative of the log-partition function (Eq. 9). The filters we obtained are shown in Figure 1—it is worth noting that the filters with smaller α's actually correspond to the more common configurations, as already discussed, because of the form of the potential function.

3.2. Inference

In order to perform inference (i.e. denoising in our experiments), we adopted a standard gradient-based approach, as in Roth and Black (2005a). Gradient-ascent is a valid technique in the case of denoising, since the noisy image is 'close to' the original image, meaning that a local maximum is likely to be a global maximum also (the few experiments we performed with a Gibbs sampler gave no better results than the gradient ascent we finally used for all the experiments). In the denoising problem, the purpose is to infer the most likely correction for the image given the image prior and the noise model. The noise model assumed in our

experiments, as in those of Roth and Black (2005a), is i.i.d. Gaussian:

$$p(\mathbf{y}|\mathbf{x}) \propto \prod_j \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}_j - \mathbf{x}_j)^2\right). \quad (11)$$

Here, j ranges over all the pixels in the image, $\mathbf{y}_j$ denotes the real color value of the noisy image at pixel j, and $\mathbf{x}_j$ denotes the color to be estimated at pixel j. Combining the likelihood (Eq. 11) and the prior (Eq. 4), the gradient of the log-posterior becomes

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \sum_{f=1}^{F} \alpha_f J_f^- * \frac{(J_f * \mathbf{x})}{1 + \frac{1}{2}(J_f * \mathbf{x})^2} + \frac{\lambda}{\sigma^2}(\mathbf{y} - \mathbf{x}),$$
$$(12)$$

where $*$ denotes matrix convolution, and the algebraic operations above are performed in an elementwise fashion on the corresponding convolution matrix (using periodic boundary conditions at the image edges). J_f^- denotes the mirror-image of J_f in two dimensions. λ is a critical parameter that gauges the relative importance of the prior and the image terms. Eq. 12 is exactly the same as the gradient given in Roth and Black (2005a), except that each 'coordinate' of the filter becomes 3-dimensional, i.e. a color rather than a grey-value.

The updated image is then simply computed by

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \delta \frac{\partial}{\partial \mathbf{x}} \log p(\mathbf{x}|\mathbf{y}), \quad (13)$$

where δ is the step-size of the gradient ascent.

4. Experimental Results

We present denoising results under two different conditions. First, we present results in which the different channels have been corrupted by different amounts of noise. These results reveal the clear need for a color image prior in such situations. Secondly, we present results in which all channels have become equally corrupted, and show that color image priors improve over non-color image priors even in this case.

The results we show use the color model trained in the RGB domain, with denoising done in this domain as well. We compare our results to those of Roth and Black (2005a), in which denoising is performed in the YCbCr domain, on each channel independently. We found that the λ-values used by Roth and Black (2005a) were unsuitable for color images, so we have
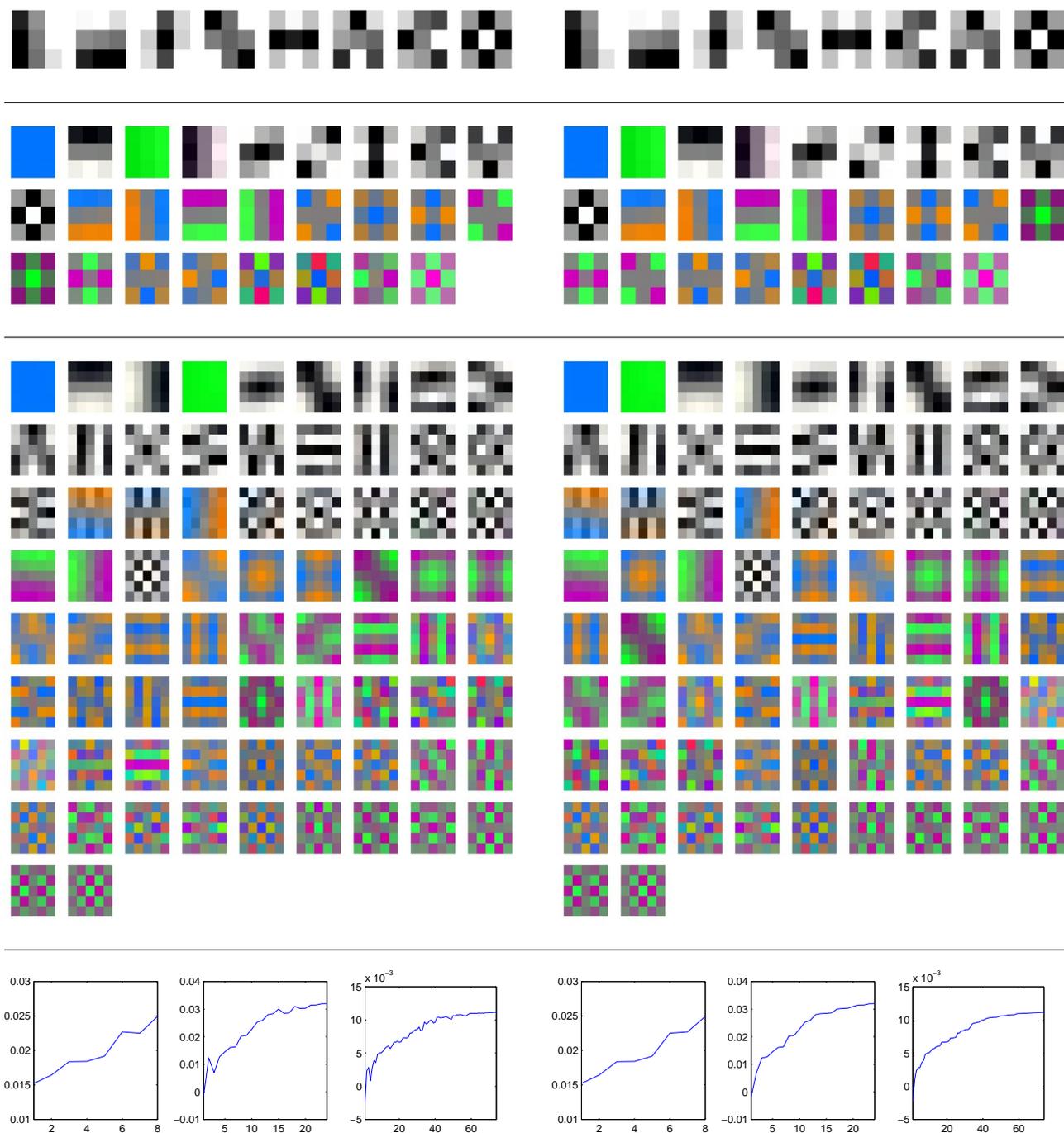
Figure 1. Above are the filters trained on 3x3 intensity, 3x3 color, and 5x5 color, respectively (the filter with highest variance, which corresponds to a flat grey patch in all three cases, is excluded). The left column shows these filters sorted according to their eigenvalues (largest to smallest), while the right column shows them sorted by their importance after learning (α's, smallest to largest). The three graphs on the left plot the α's with respect to the rank of the patches in the left column. The three graphs on the right plot the α's with respect to the rank of the patches in the right column. The fact that the plots are non-monotonic when sorted by eigenvalue demonstrates the necessity for learning the α's by maximum-likelihood. The fact that the two rightmost graphs are approximately concave reveals that, although high-order color features are less prominent than high-order intensity features, they keep being important.

selected new λ-values for their model. Training of all λ's has been done manually, using images other than those shown for our denoising experiments. This was done by denoising a test image with several candidate λ-values, and selecting whichever one yielded the best results. In addition, step-sizes have been chosen to grow linearly with the noise level, which was found to work well in practice. For the model from Roth and Black (2005a), we used step-sizes of $\sigma/100$ or $\sigma/200$ (in the 3x3 and 5x5 case, respectively), whereas we used step-sizes of $\sigma/4$ and $\sigma/2$ for our model. The large difference between these values is in fact entirely superficial, as it is only dependent on the magnitude of the α's[1].

All tables give the PSNR ($= 20 \log_{10}(255/\sigma)$) at the final iteration. Tables 1 and 2 provide both the peak *and* the final PSNR, since any difference between them can often be attributed to a λ-training issue. The images in all figures reflect the final iteration.

In Figure 2 we show an image in which the green channel has been almost completely corrupted ($\sigma = 128$), and an image in which a different amount of noise is applied to each of the three channels. Non-color models are unable to satisfactorily denoise such images, as we show by using a model in which learning has been done independently in each channel. On the other hand, our color model achieves almost complete restoration. A fairly low PSNR for the denoised images can be attributed to the fact that there has been a slight shift in the overall intensity of the final results. Nevertheless, from a visual perspective, the images denoised with the color model appear very satisfactory. The color model appears appropriate whenever the noise is not equal across all channels, with the only difficulty being selection of the weighting coefficients (λ's), and the step-sizes (δ's). We simply used the same λ's and δ's which worked well in the case of equal noise across each channel, and found that this worked reasonably well. A more sophisticated λ-training technique may improve upon these results.

In Figure 3, we show results obtained for denoising an image in which all three channels have been equally corrupted, and compare these to the state-of-the-art. It can be seen that even our 3x3 color model surpasses the 5x5 intensity model in terms of PSNR. This is an important result, since the two models have similar computational complexity (27-dimensional filters for the color model, as opposed to 25-dimensional filters

---

[1]Actually, the magnitude of the alphas *is* important when using inference techniques which rely on sampling (they would need to be rescaled), but it is unimportant for the gradient ascent approach presented in this paper.

Table 1. Comparison of the two 3x3 models, using image 102061 (castle). PSNR results, shown at peak/final.

model / σ	5	10	15	25
Noisy image	34.25	29.29	24.78	20.46
Roth & Black.	39.56/ 39.54	35.28/ 35.24	32.91/ 32.91	29.97/ 29.91
Our model	39.90/ 39.90	35.50/ 35.50	33.13/ 33.13	30.00/ 29.98

Table 2. Comparison of the two 5x5 models, using image 102061 (castle). The model from Roth and Black (2005a) was run with $K = 1000$, whereas ours was run with $K = 250$. PSNR results, shown at peak/final.

model / σ	5	10	15	25
Noisy image	34.25	29.29	24.78	20.46
Roth & Black.	39.73/ 39.64	35.33/ 35.24	32.96/ 32.95	29.95/ 29.82
Our model	40.26/ 40.26	35.91/ 35.91	33.49/ 33.49	30.41/ 30.41

for the intensity model). The results for the 5x5 color model are statistically (if not visually) superior to all others—however, we have only run the 5x5 model for a small number of iterations (250 rather than 1000) due to its increased complexity, meaning that these results are not entirely indicative of its true potential.

Tables 1 and 2 compare the intensity model from Roth and Black (2005a) with our color model in the 3x3 and 5x5 case, respectively. Further results are shown (in the 3x3 case) in Table 3.

5. Discussion

Results indicate that, whenever noise is unevenly distributed across the different channels, the proposed color model delivers significantly better results than the purely monochromatic model (Figure 2). Naturally, what happens in this case is that the color model is able to capture the correlations across the channels and the information available in the less affected channels is spread over the more affected ones. When the

Table 3. Our PSNR results, using the 3x3 model, shown at final.

image / σ	5	10	15	25
kangaroo, 69020	38.30	33.14	30.40	27.48
castle, 102061	39.90	35.50	33.13	29.98
horses, 197017	39.25	34.45	31.90	28.84
mushroom, 208001	38.81	34.16	31.76	28.73
train, 351093	38.90	33.63	30.50	26.72

Figure 2. Top – from left right: The original image, the degraded image with $\sigma = 128$ in the green channel only (PSNR = 13.87), the image restored using our single-channel model (PSNR = 24.1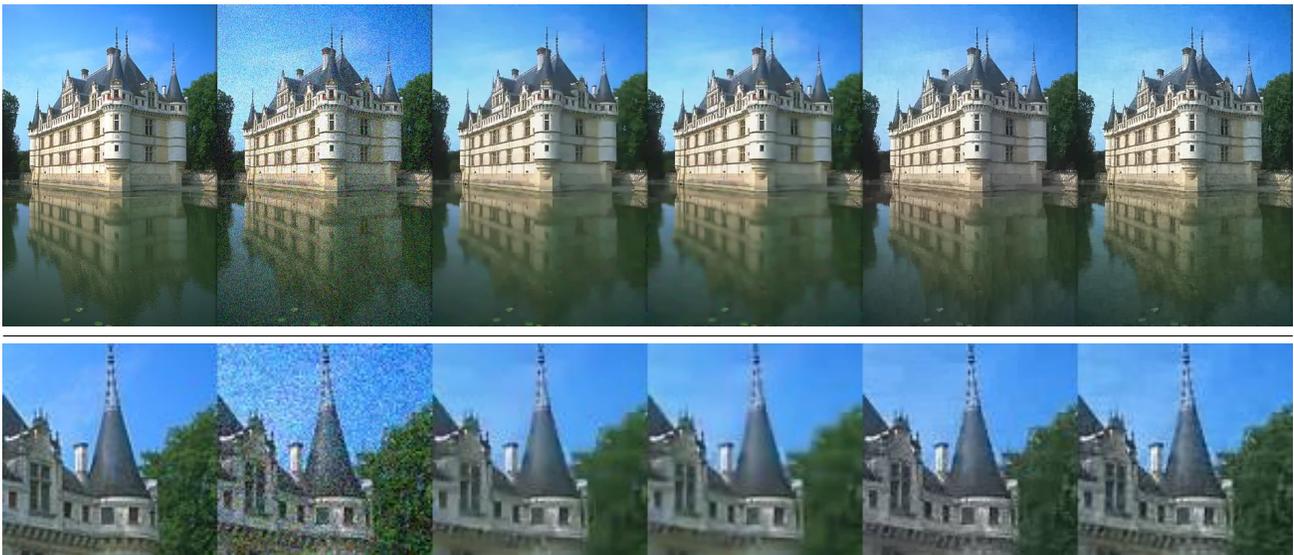5), the image restored using our color model (PSNR = 28.81). Bottom: The original image, the degraded image with $\sigma = 128$ (red), 15 (green), 5 (blue) (PSNR = 13.72), the image restored using our single channel model on each channel separately (PSNR = 23.09), the image restored using our color model (PSNR = 27.91).

Figure 3. Top – from left to right: Original image, corrupt image with $\sigma = 25$ in all channels (PSNR = 20.46), denoised image using 3x3 model from Roth and Black (2005a) (PSNR = 29.91), using 5x5 model from Roth and Black (2005a) (PSNR = 29.82), using our 3x3 model (PSNR = 29.98), using our 5x5 model (PSNR = 30.41). Bottom – close-ups of all images.

noise has the same magnitude in all channels, the color prior still outperforms the intensity prior, although the difference is less significant in this case (Figure 3).

The learned filters (Figure 1) reveal some important aspects of natural color images. For example, constant blue and green patches have highest score among all filters, indicating that the DC components of natural images for these colors are very prominent. Immediately after these filters, one notices a predominance of several non-colored filters which encode higher order features, such as edges and primitive textures. The fact that colored edges and textures appear only after their non-colored counterparts evidences that color information plays a less dominant role in encoding high-order features. However, they are nonetheless important, as can be seen from the approximately concave behavior of the two rightmost graphs in Figure 1. Note that the two large clusters of colored high-order filters, namely orange-blue and green-red, were also observed in another study using PCA (Ruderman et al., 1998). However, the actual sequence of the filters of this study differs from our SVD results at several points. This might be due to the different (and uncalibrated) datasets used in our study.

A natural next step consists of attempting to learn the filters from different criteria, possibly maximum-likelihood as well. Although this is computationally very intensive for color images, we believe the use of more sophisticated learning algorithms instead of the naïve approach presented here may render the problem feasible.

6. Conclusion

In this paper, we have studied the statistics of color images under high-order Markov random field models. By collecting a large set of sample color image patches from a standard color image database, we have learned a prior model for color images using a very simple learning algorithm. The resulting prior model can be applied in general scene inference problems, including image denoising, inpainting and super-resolution. Results comparing this color image prior to a state-of-the-art monochromatic prior for denoising problems evidence performance improvements.

References

Burton, G. J., & Moorhead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics*, *26*, 157 – 170.

Freeman, W. T., Pasztor, E. C., & Carmichael, O. T. (2000). Learning low-level vision. *International Journal of Computer Vision*, *40*, 25 – 47.

Hinton, G. E. (1999). Products of experts. *Ninth International Conference on Artificial Neural Networks, ICANN* (pp. 1–6).

Maloney, L. T. (1986). Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *J. Opt.Soc. Am. A*, *3*, 1673 – 1683.

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. *Proc. 8th Int'l Conf. Computer Vision* (pp. 416–423).

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, *37*, 3311 – 3325.

Roth, S., & Black, M. J. (2005a). Fields of experts: A framework for learning image priors. *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 860–867).

Roth, S., & Black, M. J. (2005b). On the spatial statistics of optical flow. *IEEE International Conference on Computer Vision* (pp. 42–49).

Ruderman, D. L., Cronin, T. W., & Chiao, C.-C. (1998). Statistics of cone responses to natural images: implications for visual coding. *J. Opt.Soc. Am. A*, *15*, 2036 – 2045.

Welling, M., Hinton, G. E., & Osindero, S. (2002). Learning sparse topographic representations with products of student-t distributions. *NIPS* (pp. 1359–1366).

Zhu, S. C., Wu, Y., & Mumford, D. (1998). Filters, random fields and maximum entropy (frame): Towards a unified theory of texture modeling. *International Journal of Computer Vision*, *27*, 107 – 126.