# Estimating Markov Random Field Potentials for Natural Images

Urs Köster[1,2], Jussi T. Lindgren[1,2] and Aapo Hyvärinen[1,2,3]

[1] Department of Computer Science, University of Helsinki, Finland
[2] Helsinki Institute for Information Technology, Finland
[3] Department of Mathematics and Statistics, University of Helsinki, Finland

**Abstract.** Markov Random Field (MRF) models with potentials learned from the data have recently received attention for learning the low-level structure of natural images. A MRF provides a principled model for whole images, unlike ICA, which can in practice be estimated for small patches only. However, learning the filters in an MRF paradigm has been problematic in the past since it required computationally expensive Monte Carlo methods. Here, we show how MRF potentials can be estimated using Score Matching (SM). With this estimation method we can learn filters of size $12 \times 12$ pixels, considerably larger than traditional "hand-crafted" MRF potentials. We analyze the tuning properties of the filters in comparison to ICA filters, and show that the optimal MRF potentials are similar to the filters from an overcomplete ICA model.

## 1 Introduction

Probabilistic models of natural images are useful in a wide variety of applications, such as denoising and inpainting[1], novel view synthesis[2], texture modelling [3], and in modelling the early visual system [4]. Such models can also provide controllable test stimuli for experiments in neurophysiology and psychophysics.

Two approaches that have received significant interest with relation to image modelling are Markov Random Fields (MRF, e.g. [5]) and Independent Component Analysis (ICA [6], in images context see e.g. [4]). Traditionally, in the MRF framework the model parameters have been selected by hand (e.g. [3]) rather than learned, whereas in the ICA approach the model parameters are learned from the data. Only recently Roth and Black have shown that MRF performance can be improved by fitting the model parameters to natural image data [1].

In ICA, the observed data vector $\mathbf{x}$ is assumed to be generated as a linear superposition of features, $\mathbf{x} = A\mathbf{s}$, where the distribution of the sources is usually assumed to be a known supergaussian probability density function (pdf). Due to the assumption that sources are independent, we can write $p(\mathbf{s}) = \prod_i p_i(s_i)$ or for the log-probability $\log p(\mathbf{s}) = \sum_i \log p_i(s_i)$. If the mixing matrix $A$ is invertible and has inverse $W$, consisting of vectors $\mathbf{w}_i$, we can make a transformation of density to obtain the pdf for the data as $\log p(\mathbf{x}) = \sum_i \log p_i(\mathbf{w}_i^T \mathbf{x}) + \log |\det W|$. This model can easily be estimated with maximum likelihood.

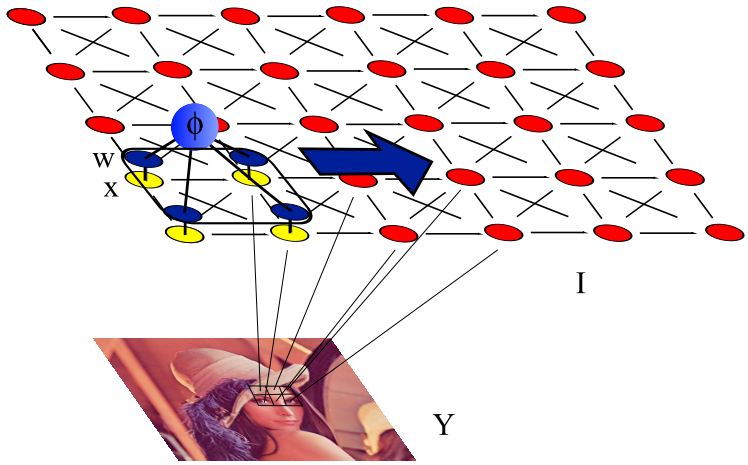



**Fig. 1.** Sketch of a Markov Random Field: The MRF has maximal cliques of size $2 \times 2$ pixels; one clique $\mathbf{x}_i$ is highlighted. Each unit of the field is associated to a pixel of the underlying image $Y$. The potential energy $V$ for each clique is defined as a function of the inner product of the image patch corresponding to the clique with a bank of filters of the same size as the clique, $V(\mathbf{x}) = \phi(\mathbf{w}^T\mathbf{x})$. This is visualized by the filter vector $\mathbf{w}$ that is scanned over the whole image, and the product is computed with each clique. In general there will be several filters in a filter matrix $W$, but for visualization purposes only one is shown.

On the other hand, a MRF is a graphical model that is defined as a 2-D lattice of units with undirected links, as illustrated in Fig. 1. The maximal cliques formed by these connections play an important role as the potential energy of the field is given as a function of these cliques. The key property of a MRF is conditional independence, so the state of each unit on the field depends only on those units it is directly linked to and the unit is independent of all other units in the field. While ICA is limited to modelling small image patches, the MRF provides a principled model for *whole* images of arbitrary size, even if the clique size is limited.

The paper is structured as follows: In Section 2 we present the MRF model and how it can be estimated with Score Matching. In Section 3 we discuss the application of the model to natural images, show the filters that are obtained, and analyze the properties of the filters compared to ICA models. Finally we discuss the implications of this work in Section 4.

## 2 The MRF Model and Estimation

In contrast to ICA, where filter responses are computed by a simple inner product, the energy (i.e. the negative logarithm of the non-normalized pdf) of a MRF is given by a convolution of the image $I$ with potential functions $U_k$

$$V(\mathbf{I},\theta) = \sum_{k,x,y} \phi\left(\mathbf{U}_k * \mathbf{I}\right) = \sum_{k,x,y} \phi\left(\sum_{x',y'} \mathbf{U}_{k,x',y'} \mathbf{I}_{x-x',y-y'}\right) \tag{1}$$

where the convolution (denoted by $*$) runs over pixels indices $x$ and $y$. The elementwise nonlinearity $\phi$ gives the energy of the cliques of the field, which are simply summed up to obtain the energy $V$ of the field. As it is customary in ICA to work on whitened data, we insert a whitening filter $\mathbf{Q}$ in the convolution so it becomes $V(\mathbf{I},\theta) = \sum_{k,x,y} \phi(\mathbf{U}_k * \mathbf{Q} * \mathbf{I})$. The whitening filter can be absorbed into the image, which corresponds to estimating a model for white data, but it can also be viewed as a part of the potential function when the model is applied to non-whitened data. It is important to use a whitening filter rather than an arbitrary whitening matrix for this to hold.

The unnormalized probability of the model is given by the exponential of the negative energy, and must be normalized by the partition function $Z$. Since $Z$ cannot be computed in closed form, we estimate the model using Score Matching [7], which works on the non-normalized distribution. To estimate the model with Score Matching we need to compute the Score function $\Psi_j = \frac{\partial V}{\partial \mathbf{I}_j}$, the Score Matching objective function $J$ and its derivatives w.r.t. the parameter vectors.

$$J = \sum_j \left(\frac{1}{2}\Psi_j^2 + \Psi_j'\right), \qquad \frac{\partial J}{\partial \mathbf{w}_k} = \sum_j \left(\Psi_j \frac{\partial \Psi_j}{\partial \mathbf{w}_k} + \frac{\partial \Psi_j'}{\partial \mathbf{w}_k}\right) \tag{2}$$

For further analysis it is convenient to rewrite the convolution as a discrete sum of inner products. We rewrite the convolution $\mathbf{I} * \mathbf{U}_k = \mathbf{X}\mathbf{w}_k$ where $\mathbf{X}$ is a matrix containing vectorized patches $\mathbf{x}_i$ from the image, and $\mathbf{w}_k$ is a vectorized filter. Similarly we write $\mathbf{X}_j$ as a subset of $\mathbf{X}$ containing only those patches which include the image pixel $I_j$. Thus the energy becomes

$$V(I,\theta) = \sum_{k,i} \phi\left(\mathbf{w}_k^T \mathbf{x}_i\right) \tag{3}$$

Where the sum over $i$ is over the patches contained in the matrix $\mathbf{X}$. Using this notation we can compute the score function w.r.t. to the image pixels $\mathbf{I}_j$

$$\Psi_j = \frac{\partial V}{\partial \mathbf{I}_j} = \frac{\partial}{\partial \mathbf{I}_j} \sum_{k,x,y} \phi\left(\mathbf{U}_k * I\right) = \sum_k \check{\mathbf{w}}_k^T \phi'\left(\mathbf{X}_j \mathbf{w}_k\right) \tag{4}$$

$$\Psi_j' = \frac{\partial^2 V}{\partial \mathbf{I}_j^2} = \sum_k (\mathbf{w}_k \odot \mathbf{w}_k)^T \phi''\left(\mathbf{X}_j \mathbf{w}_k\right) \tag{5}$$

We denote elementwise multiplication of vectors by $\odot$, and $\check{\mathbf{w}}$ indicates reversal of the order of elements in a vector. It is important to note that in order to avoid border effects, the index $j$ does not run over *all* image pixels, but only those that lie in the central region of the image so it can be reached by all pixels in the filter $\mathbf{w}$. The gradient of the objective function is now easily obtained from the gradients

$$\frac{\partial \Psi_j}{\partial \mathbf{w}_k} = \check{\phi}' \left( \mathbf{X}_j \mathbf{w}_k \right) + \mathbf{X}_j \left[ \check{\mathbf{w}}_k \odot \phi'' \left( \mathbf{X}_j \mathbf{w}_k \right) \right] \tag{6}$$

$$\frac{\partial \Psi_j'}{\partial \mathbf{w}_k} = 2\check{\mathbf{w}}_k \phi'' \left( \mathbf{X}_j \mathbf{w}_k \right) + \mathbf{X}_j \left[ \check{\mathbf{w}}_k \odot \check{\mathbf{w}}_k \odot \phi''' \left( \mathbf{X}_j \mathbf{w}_k \right) \right] \tag{7}$$

Thus the Score Matching objective can easily be optimized by gradient descent.

## 3 Experiments on Natural Images

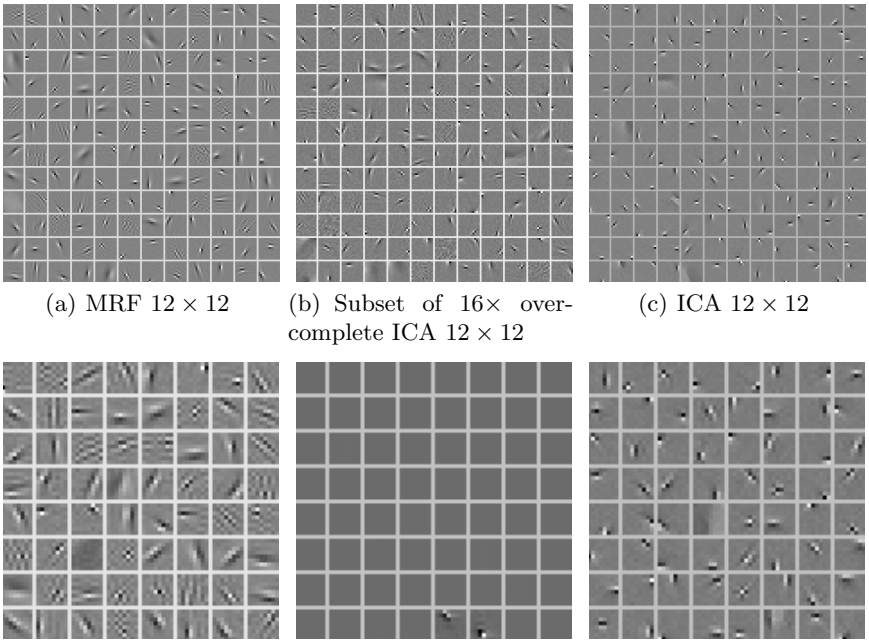
### 3.1 Methods

We performed experiments on natural images from P.O. Hoyer's ImageICA package[4]. For the ICA and overcomplete ICA experiments we randomly sampled 20,000 image patches of $8 \times 8$ and $12 \times 12$ pixels size. For the MRF, we sampled 5,000 larger "images" of size $25 \times 25$ and $36 \times 36$ for use with filters of size $8 \times 8$ and $12 \times 12$ respectively. Now since the main advantage of the MRF model over ICA is that it can be applied to arbitrarily large images, it may seem surprising that we use images that are not significantly larger than the patches ordinarily used in ICA. However, what is important for estimating the model is the size of the filters relative to the images. In particular, since we use only the valid region of the convolution, only the central pixels of the image contribute to the objective function. Thus the full range of dependencies is captured, and the filters should be identical if they were estimated with larger images.

We used less samples for the MRF than for ICA since each of the images is effectively generating more data points due to the convolution. In preprocessing we removed the DC value of the images and normalized them to unit variance. After sampling, we whitened the image vectors with a zero phase whitening filter [8]. We did not reduce the dimensionality with PCA as it is customary in ICA models, since this would destroy the structure that we wish to capture with the MRF. Therefore the highest frequencies containing aliasing artifacts due to the rectangular sampling grid will be boosted, which has to be taken into account in the analysis of the results.

We performed experiments on a complete ICA model with 144 filters, and a 16 times overcomplete ICA model with 2304 filters. The MRF had 144 filters, and all three models were estimated with Score Matching. The filters were initialized randomly and estimated by gradient descent, in case of the MRF a stochastic gradient with a batch size of 20 was used. The experiments were repeated 10 times with different random seed for the sampling of image patches

[4] available at http://www.cs.helsinki.fi/u/phoyer/imageica.tar.gz



(a) MRF $12 \times 12$ (b) Subset of $16\times$ overcomplete ICA $12 \times 12$ (c) ICA $12 \times 12$

(d) MRF $8 \times 8$ (e) MRF $8 \times 8$, no normalization (f) ICA $8 \times 8$

**Fig. 2.** Comparison of the filters obtained for $12 \times 12$ (top) and $8 \times 8$ (bottom) image patches. The complete and overcomplete ICA models shows the well-known Gabor like filters, and the MRF potentials are very similar, sharing the properties of localization and tuning for spatial frequency, phase and orientation. While for the ICA model it is not necessary to normalize the filters, it is interesting to note that in the MRF case almost all the filters go to zero unless the norms of the vectors $\mathbf{w}$ are constrained to be unity.

and initialization of the weight matrix. All the filters were normalized to unit norm, which is necessary to prevent filters from going to zero in the overcomplete ICA and MRF models. Convergence was determined by optical inspection of the filters. Because the estimation of ICA with Score Matching is not widely used, we also estimated the complete ICA model with FastICA, to control for differences that are due to the estimation method.

### 3.2 Results

In most classical MRF work, the potentials that were used were of rather small size such as $3 \times 3$ pixels and typically chosen to be directional derivatives. Thus it is perhaps not surprising that the larger MRF filters we estimated are very similar to ICA filters in appearance, being localized Gabor-like filters with tuning

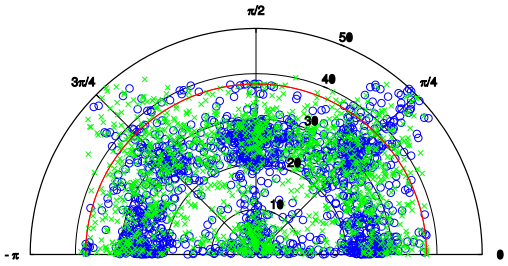


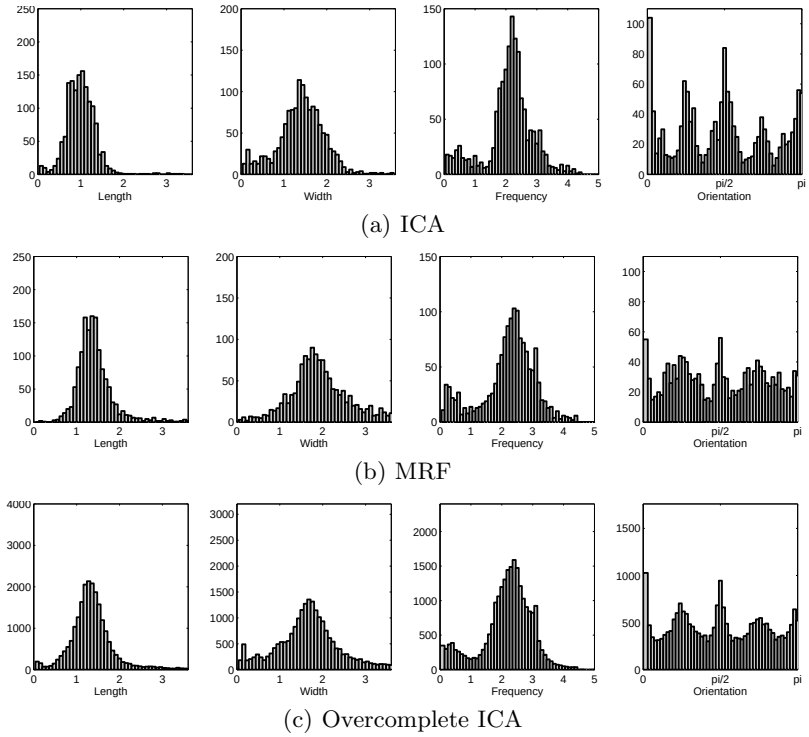

**Fig. 3.** Polar plot of frequency vs. orientation for $12 \times 12$ image patches for ICA (circles) and MRF (crosses). The orientations are not uniformly distributed, with filters preferring to be aligned along the pixels, horizontal or vertical, and at 45 degrees to these directions. Due to the rectangular sampling grid, the maximum frequency is higher along the diagonal, which may also account for the non-uniformity. Usually this problem is alleviated by dimensionality reduction amounting to high-pass filtering, which is not easily possible with the MRF model.

for spatial frequency, phase and orientation. This is shown in Fig. 2, where ICA and MRF filters are compared directly for different image patch sizes.

To analyze the similarity between the two models further, we fit Gabor function to the filters so we can analyze their tuning properties. We used a least squares fit adapted from [9] to parametrize the filters in terms of length and width, frequency, phase and orientation. In Fig. 3 we show a polar plot plotting orientation against frequency.

In Fig. 4 we show histograms of the size and frequency distribution for the three models. The complete ICA model produces very localized filters which cover a relatively narrow band of frequencies. Both overcomplete ICA and the MRF give slightly larger filters with a slightly broadened distribution of frequencies. While the distributions for the MRF and ICA are somewhat different, it is important to note that the filters for overcomplete ICA are also slightly different and in some respects more similar to the MRF (e.g. somewhat larger filters with less peaked frequency tuning). This may suggest that there are no fundamental differences between the filters obtained from the two models.

We performed t-tests to quantify the statistical significance of the difference in mean length, width and frequency of the filters between the four models, FastICA, ICA and overcomplete ICA estimated with Score Matching and the MRF. Only in the comparison between the MRF and overcomplete ICA, there was no sufficient difference in the tuning properties to reject the null hypothesis at the Bonferroni corrected threshold of 0.017. It is interesting to note that estimating the same ICA model with two different estimation methods produces a larger difference in the filters, than the difference between overcomplete ICA and the MRF estimated with Score Matching.




(a) ICA

(b) MRF

(c) Overcomplete ICA

**Fig. 4.** Tuning of ICA (top), MRF and overcomplete ICA (bottom) for $12 \times 12$ image patches. We show the size (length and width in pixels) of the Gaussian envelope of the Gabors we fit, and the distribution of frequencies (rad per pixel). Additionally, we show the distribution of orientations, which is clearly not uniform in both cases.

## 4 Discussion

Estimating optimal MRF potentials from natural images has previously been attempted by Roth and Black [1], making use of Contrastive Divergence (CD) [10]. We would like to point out that the filters obtained in the work of Roth and Black have a very different appearance, being disjoint and distributed over the whole image patch rather than the coherent and smooth Gabors that we obtain. The patch size used by those authors was considerably smaller ($3 \times 3$ and $5 \times 5$, which may forces features to spread out more to capture the longer range dependencies of natural images. In addition, it is conceivable that with the particular Monte Carlo method used by the authors, a different local optimum is found or the method encountered some other problems.

It is possible to view the MRF model as a highly overcomplete ICA model with some additional constraints. In particular, the convolution in (1) can be interpreted as keeping the image fixed, and multiplying it with the filters in

different positions.The resulting "filters" would be shifted copies of the original filters at different positions in the image and padded with zeros. Thus, while the model is highly overcomplete, non of the filters model the whole image, but only regions. If we assume natural images to be stationary having copies of the filters at different positions does not have an effect, and the main difference to ICA would be that the size of the filters is restricted to be much smaller than the image. This makes it quite obvious that optimal MRF filters should not be vastly different from ICA filters. It would be interesting to investigate if there are systematic differences in the sets of filters, and how they tile the parameter space of positions, orientations etc. In particular, while an ICA basis may contain nearly identical filters in different positions, this should not be the case with the MRF model. Therefore, if one were to attempt to use ICA filters in place of MRF potentials for e.g. a denoising task, one would face the problem of selecting the correct subset of an ICA basis to form a set of near-optimal MRF potentials.

To conclude, we have show that it is possible to learn the filters used in a non-Gaussian Markov Random Field model. The learning is based on score matching and leads to Gabor-like filters. This gives a well-defined probabilistic model of whole images instead of just small patches.

**Acknowedgements** We wish to thank Jascha Sohl-Dickstein for helpful discussions. Urs Köster is supported by a Scholarship from the *Alfried Krupp von Bohlen und Halbach-foundatation.*

## References

1. Roth, S., Black, M.: Fields of experts: A framework for learning image priors. CVPR, vol. 2 (2005) 860–867.
2. Woodford, O., Reid, I., Torr, P., Fitzgibbon, A.: Fields of experts for image-based rendering. Proceedings British Machine Vision Conference (2006)
3. Zhu, S., Wu, Y., Mumford, D.: FRAME: Filters, random field and maximum entropy – towards a unified theory for texture modeling. International Journal of Computer Vision **27**(2) (1998) 1–20
4. van Hateren, J.H., van der Schaaf, A.: Independent component filters of natural images compared with simple cells in primary visual cortex. Proc.R.Soc.Lond. B **265** (1998) 359–366
5. Li, S.Z.: Markov Random Field modelling in image analysis, 2nd edition. Springer (2001)
6. Comon, P.: Independent component analysis – a new concept? Signal Processing **36** (1994) 287–314
7. Hyvärinen, A.: Estimation of non-normalized statistical models using score matching. Journal of Machine Learning Research **6:695–709** (2005)
8. Atick, J.: Could information theory provide an ecological theory of sensory processing? Network: Computation in neural systems **3** (1992) 213–251
9. Hyvärinen, A., Hoyer, P., Inki, M.: Topographic independent component analysis. Neural Computation. (2001)
10. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Comput. **14**(8) (2002) 1771–1800