The Handbook of Brain Theory and Neural Networks, M. Arbib, editor, 564-567, MIT Press, 1995

Markov random field models in image processing

Anand Rangarajan[†] and Rama Chellappa[‡]

[†]Department of Computer Science
Yale University, New Haven, CT

[‡]Department of Electrical Engineering and
Center for Automation Research
University of Maryland, College Park, MD

RUNNING HEAD: Markov random field models in image processing

Correspondence:
Anand Rangarajan
Department of Computer Science, Yale University
51 Prospect Street
New Haven, CT 06520-8285
Phone: (203) 432 1219
Fax: (203) 432 0593
email: rangarajan-anand@cs.yale.edu

1. INTRODUCTION

Markov random field models have become useful in several areas of image processing. The success of Markov random fields (MRFs) can be attributed to the fact that they give rise to good, flexible, stochastic image models. The goal of image modeling is to find an adequate representation of the intensity distribution of a given image. What is adequate often depends on the task at hand and MRF image models have been versatile enough to be applied in the areas of image and texture synthesis (Chellappa and Kashyap, 1985), image compression, restoration (Geman and Geman, 1984), tomographic reconstruction (Geman and Graffigne, 1987), image and texture segmentation (Rangarajan et al., 1991), texture classification, (Derin and Elliott, 1987), and surface reconstruction (Geiger and Girosi, 1991). Our aim is to highlight the central ideas of this field using illustrative examples and provide pointers to the many applications.

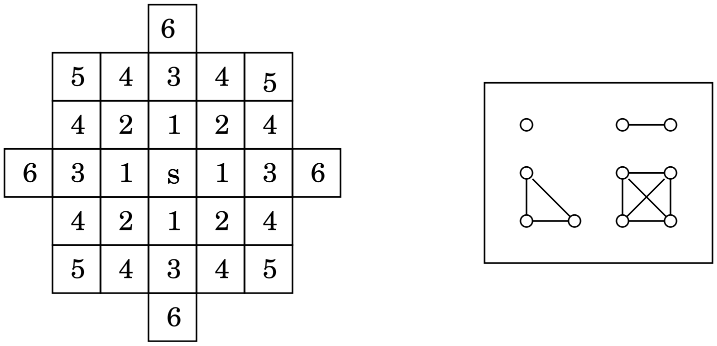
A guiding insight underlying most of the work on MRFs in image processing is that the information contained in the local, physical structure of images is sufficient to obtain a good, global image representation. This notion is captured by means of a local, *conditional* probability distribution. Here, the image intensity at a particular location depends only on a *neighborhood* of pixels. The conditional distribution is called an MRF. For example, a typical MRF model assumes that the image is locally smooth except for relatively few intensity gradient discontinuities corresponding to region boundaries or edges. The MRF image models are defined on the image intensities and on a further set of *hidden* attributes (edges, texture and region labels). The observed quantities are usually noisy, blurred images, feature vectors or projection data in the case of emission tomography. The intensity image underlying the observations is needed in applications like restoration and tomographic reconstruction, whereas, region, boundary and texture labels are sought in applications like texture segmentation.

Once the local, conditional probability distribution of the MRF is specified, there are five remaining steps involved. First, the joint distribution of the MRF is obtained. In this way, the image is represented in one global, joint probability distribution. Next, the process by which the observations are generated from the image is captured in a *degradation* probability distribution. In image restoration, for example, the degradation corresponds to a (typically uniform) blur. Then, Bayes' theorem is invoked to obtain the posterior probability distribution of the image given the observations. The posterior distribution gives us the probability that an image (with smooth regions and sharp region boundaries for example) could have been degraded to obtain the particular observed noisy, blurred image. Once the posterior probability distribution is obtained, we can associate a cost with each configuration in the posterior. For example, if only the true underlying image will do, the cost penalizes all other images equally. The cost is formulated keeping in mind the task at hand. A measure of the cost is minimized with respect to the image intensities (in image recovery tasks) or image attributes (in labeling tasks). Finally, since the MRFs are specified with model parameters, these are estimated from a training set (if one exists) or adaptively along with the cost minimization phase alluded to earlier. The overall MRF framework fits well within a Bayesian estimation/inference paradigm. In the next section, we step through all five phases of MRF modeling.

2. A FRAMEWORK FOR ESTIMATION AND INFERENCE

As mentioned in the Introduction, MRF image models represent knowledge in terms of "local" probability distributions. Specifically, the kinds of probability distributions generated by MRFs have a local neighborhood structure. Neighborhood systems commonly used by MRFs are depicted in Figure 1(a). In our exposition, we have adopted much of the notation found in (Geman and Graffigne, 1987).

Let us associate an image with a random process X whose element is X_s, where $s \in S$



			6			
	5	4	3	4	5	
	4	2	1	2	4	
6	3	1	s	1	3	6
	4	2	1	2	4	
	5	4	3	4	5	
			6			

(a) (b)

Figure 1:

refers to a site in the image. The local conditional distribution can be written as follows:

$$\Pr(X_s = x_s | X_t = x_t, t \neq s, t \in S) = \Pr(X_s = x_s | X_t = x_t, t \in G_s). \tag{1}$$

where X and x denote the random field and a particular realization respectively and G_s is the local neighborhood at site s. Note that in general, G_s can be large or small, but it is usually a local neighborhood in keeping with the spirit of MRF modeling.

Let s be the site (i, j) and let the local neighborhood be a first-order neighborhood ($G(s)$ is the collection $(i, j+1)$, $(i, j-1)$, $(i+1, j)$, $(i-1, j)$). Then, let the conditional density take the form

$$p(X_s = x_s | X_t, t \in G_s) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(x_{ij} - \frac{1}{4}\left[x_{i,j+1} + x_{i,j-1} + x_{i+1,j} + x_{i-1,j}\right]\right)^2\right] \tag{2}$$

This is a very simple special case of the first order Gauss-Markov model (Woods, 1972; Besag, 1974). The Gauss-Markov model has been widely used in image processing tasks (Dubes and Jain, 1989).

The MRF model consists of a set of *cliques*. A clique is a collection of sites such that any two sites are neighbors. Different orders of cliques are shown in Figure 1(b). The order of a clique refers to the number of distinct sites that appear multiplicatively. We now calculate the clique energies involving the site x_{ij} by expanding the conditional probability density and collecting the terms. There are cliques of order one and two. They are

$$\frac{x_{ij}^2}{2}, \ -\frac{x_{ij}x_{i,j+1}}{4}, \text{ and } -\frac{x_{ij}x_{i+1,j}}{4}. \tag{3}$$

The first term in (3) is of order one and the latter two terms are of order two.

2.1. MRF–Gibbs equivalence

We now ask the following question: given the conditional probability structure $\Pr(X_s = x_s|X_t = x_t, t \in G_s)$, what is the joint probability distribution $\Pr(X = x)$? This is of utmost importance since it is the joint probability distribution and not the conditional that contains the complete image representation.

Before relating the conditional and joint distributions, we introduce the concept of a Gibbs distribution which will turn out to be crucial in specifying the relationship. A Gibbs distribution is specified by an "energy function" $E(x)$ and can be written as

$$\Pr(X = x) = \frac{1}{Z}\exp(-E(x)) \tag{4}$$

where the *partition function*

$$Z = \sum_x \exp(-E(x)) \tag{5}$$

is a normalizing constant and involves a summation over all possible configurations of X. Energy functions have been widely used in spin glass models of statistical physics. The minimum energy configuration corresponds to an ordered system of spins. $E(x)$ cannot take infinite values.

Our detour into Gibbs distributions is justified for the following reason. The Hammersley-Clifford theorem (Besag, 1974; Geman and Geman, 1984) states that any conditional distribution has a joint distribution which is Gibbs (Kinderman and Snell, 1980; Dubes and Jain, 1989) if the following conditions hold.

Positivity: $\Pr(X = x) \ > \ 0$.

Locality: $\Pr(X_s = x_s|X_t = x_t, t \neq s, t \in S) = \Pr(X_s = x_s|X_t = x_t, t \in G_s)$.

Homogeneity: $\Pr(X_s = x_s|X_t = x_t, t \in G_s)$ is the same for all sites s.

The locality condition is the same as the Markov property (1). The Hammersley-Clifford theorem allows us to shuttle between the conditional probability structure in (1) and the joint probability in (4).

The recipe for obtaining the joint density function is as follows: (i) assemble the different clique energies from the conditional probability, and (ii) compute the energy function by adding up the clique energies.

We calculate the energy function for the simple first order Gauss-Markov model:

$$E(x) = \frac{1}{2}\left(\sum_{ij}\left[x_{ij}^2 - \frac{x_{ij}x_{i,j+1}}{2} - \frac{x_{ij}x_{i+1,j}}{2}\right]\right) = \frac{1}{8}\sum_{ij}\left[(x_{ij} - x_{i,j+1})^2 + (x_{ij} - x_{i+1,j})^2\right]. \quad (6)$$

It can be seen from the energy function $E(x)$ and the conditional density that the essence of the Hammersley-Clifford theorem lies in the clique energies. We examined the conditional density and teased apart the different orders of cliques (first and second order) and the associated clique energies. Then, all clique energies were summed (taking care to count each clique only once) yielding the energy function $E(x)$. Our presentation has been quite terse and further details on cliques and the transition from the conditional to the joint probability distribution can be found in (Besag, 1974; Geman and Geman, 1984; Kinderman and Snell, 1980).

2.2. The prior and degradation models

Naturally, we are not content with merely obtaining MRF–Gibbs image models. These models can be used in a variety of image processing and pattern recognition tasks. As mentioned previously, MRF modeling fits perfectly into a Bayesian estimation/inference paradigm. A Bayesian setup consists of two ingredients—the prior and the degradation model. The prior model is defined on the set of image attributes X that are of interest. In edge preserving image restoration for example, X includes the set of image intensities and a further set of binary-valued edge labels. In texture segmentation, X includes the image intensities and a set of texture labels at each location. The degradation model is a model of the physical process by which the observations are generated. Usually, we are faced with noisy and incomplete observations. Denote the set of observations by Y and let the degradation model also be a Gibbs-Markov distribution:

$$\Pr(Y = y|X = x) = \frac{1}{Z_D(x)} \exp(-E_D(x,y)) \tag{7}$$

where

$$Z_D(x) = \sum_y \exp(-E_D(x,y)). \tag{8}$$

In general, the partition function $Z_D(x)$ is a function of the image attributes x. $E_D(x,y)$ is the energy function corresponding to the degradation model. For example

$$E_D(x,y) = \frac{1}{2}\sum_s (y_s - \sum_t h_{st}x_t)^2$$

yields a Gaussian degradation model wherein Y is obtained by blurring X with a *point spread function* h and adding additive Gaussian noise at each site s. This type of degradation model routinely occurs in image restoration and in tomographic reconstruction.

2.3. A Bayesian posterior energy function

Given the degradation and prior models, Bayesian estimation/inference proceeds as follows. The posterior distribution $\Pr(X = x|Y = y)$ is obtained by using Bayes' theorem:

$$\Pr(X = x|Y = y) = \frac{\Pr(Y = y|X = x)\Pr(X = x)}{\Pr(Y = y)}. \tag{9}$$

Once the posterior distribution is obtained, an estimate ($\hat{X}$) of X is found by minimizing the expected cost which is a measure of distance between the true and estimated values.

$$C = \sum_x C(x, x^*)\Pr(X = x|Y = y) \tag{10}$$

where x^* is the true value. When the familiar squared-error cost is used, the estimator (MMSE) turns out to be the conditional mean $\mathcal{E}(X|Y = y)$ ($\mathcal{E}$ denotes the expectation operator). If the cost penalizes all x different from x^* ($C(x,x^*) = \delta_{x,x^*}$), the maximum *a posteriori* (MAP) estimator results.

When the degradation and prior models are Gibbs, the posterior is Gibbs as well. To see this, assume a prior energy function $E_P(x)$ giving $\Pr(X = x) = \frac{1}{Z_P}\exp(-E_P(x))$. The posterior distribution (using (9)) is

$$\Pr(X = x|Y = y) = \frac{\exp(-E_D(x,y) - \log(Z_D(x)) - E_P(x))}{\sum_x \exp(-E_D(x,y) - \log(Z_D(x)) - E_P(x))} \tag{11}$$

The posterior energy function $E(x) = E_D(x,y) + \log(Z_D(x)) + E_P(x)$. In the case of the MAP estimate, the entire Bayesian estimation engine reduces to minimizing just this posterior energy function $E(x)$ since the partition function of the posterior is independent of x. However, when the MMSE estimate is desired, the expected value of X in the posterior distribution needs to be computed. This computation is usually intractable since it involves computing the partition function of the posterior distribution.

2.4. MAP estimation

Restricting our focus to MAP estimation, we observe that MAP estimation reduces to minimizing the posterior energy function $E(x)$. This minimization involves the different kinds of processes which make up X. For example, in edge preserving image restoration (Geman and Geman, 1984) the process X includes both continuous image intensities and binary-valued edge variables. Consequently, the minimization of the posterior objective function is a difficult problem due to the presence of non-trivial local minima. A general technique for finding global minima is Simulated Annealing (SA) but it is usually computationally very intensive. Recently, a lot of effort has been expended in obtaining good sub-optimal solutions to the MAP estimation problem (Geiger and Girosi, 1991; Lee et al., 1993). Deterministic Annealing (DA) is a general method that has emerged recently. Deterministic annealing methods begin with a modified posterior:

$$\Pr(X = x|Y = y) = \frac{1}{Z(\beta)}\exp(-\beta E(x)) \tag{12}$$

where $\beta > 0$ is the inverse temperature. Note that the partition function is now a function of the inverse temperature. The terminology is inherited from statistical physics. The idea of cooling a system slowly to reach a minimum energy configuration has a computational parallel in MRFs. The basic idea is to embed the posterior in a β exponentiated manner and to track the maximum of this posterior through gradual increase of β. In this manner, the posterior energy function is increasingly, closely approximated by a sequence of smooth, continuous energy functions.

The main reason for doing this is based on the following statistical mechanics identity:

$$F(\beta) \stackrel{\text{def}}{=} -\frac{1}{\beta}\log Z(\beta) = \mathcal{E}(E(x)) - \frac{1}{\beta}S(\beta) \tag{13}$$

where S is the entropy (defined as $-\sum_x \Pr(X = x|Y = y)\log(\Pr(X = x|Y = y))$). The entropy is proportional to the logarithm of the total number of configurations and as the temperature is reduced (and fewer configurations become likely), it gradually goes to zero. Also, the expected value of the posterior energy goes to the minimum value of the energy. The key idea in deterministic annealing is to minimize the *free energy* F instead of $E(x)$ while reducing the temperature to zero. The free energy (at low β) is a smooth approximation to the original, non-convex energy function and approaches $E(x)$ as β tends to infinity. However, the free energy involves the logarithm of the partition function which is intractable! An approximation to the free energy (usually called the naive mean field approximation) is minimized instead. While details are beyond the scope of this presentation (the reader is referred to (Geiger and Girosi, 1991; Lee et al., 1993)), we present an example illustrating the method. Let the energy function contain only binary-valued variables and take the following form:

$$E(x) = \sum_{ij} T_{ij}x_i x_j + \sum_i h_i x_i, x_i \in \{0, 1\}. \tag{14}$$

The free energy F is given by

$$F(v) = \sum_{ij} T_{ij}v_i v_j + \sum_i h_i v_i + \frac{1}{\beta}\sum_i [v_i \log(v_i) + (1 - v_i)\log(1 - v_i)] \tag{15}$$

where $v_i \in [0, 1]$. The free energy consists of two terms. The first term can be seen as an approximation to the expected value of the energy once the identification $v_i \approx \mathcal{E}(x_i)$ is made.

Now,

$$\mathcal{E}(E(x)) = \sum_{ij} T_{ij}\mathcal{E}(x_i x_j) + \sum_i h_i \mathcal{E}(x_i). \tag{16}$$

When the expected value of the product $x_i x_j$ is replaced by the product of the expected values $(v_i v_j)$, the naive mean field approximation results. The third term in (15) is an approximation to the entropy. At each setting of β, (15) is minimized w.r.t v after which β is increased. In this manner, a deterministic relaxation network is obtained. There are questions regarding the choice of annealing schedules and the quality of the minima obtained, etc., and for the most part, except for very specific posterior energy functions, there is a dearth of analytical results in this area. However, the method is quite general and has been applied with varying degrees of success in a variety of image processing tasks like restoration, tomographic reconstruction, flow field segmentation and surface reconstruction.

2.5. Parameter estimation

So far, we have concentrated on estimating X given the noisy observations Y. We have emphasized that Gibbs-Markov models are specified by local clique energies (from which the global distribution can be obtained). Consider a prior distribution

$$\Pr(X = x|\theta) = \frac{1}{Z(\theta)} \exp(-\frac{1}{2} \sum_k \sum_{<s,t>_k} \theta_k (x_s - x_t)^2) \tag{17}$$

where θ_k is a parameter associated with clique $< s, t >$. Since pairwise interactions are used, a clique between pixels s and t is denoted by $< s, t >$. This is the general form of the Gauss-Markov model. The model is a generalization of our earlier model (6) since it has the same clique form, albeit with a more general neighborhood structure. The partition function involves a sum over the configurations of X and is a function of θ. Other than the estimation/inference problem, we are also saddled with the problem of parameter estimation.

The parameters can be estimated by maximizing the joint probability of X w.r.t the unknown parameters. In most cases, this computation is intractable in its pure form and approximations have to be devised. An interesting alternative is to maximize the "pseudo-likelihood" with respect to the unknown parameters (Besag, 1977). The pseudo-likelihood takes advantage of the local conditional probability structure of MRFs. The parameters are now estimated by maximizing the product of the conditional distributions at each site s w.r.t the parameters. The availabilty of a suitable training set is critical to both likelihood and pseudo-likelihood parameter estimation. When a training set is not available, parameter estimation and cost minimization proceed in lockstep. There are also issues of consistency and efficiency of these estimates and more details can be found in (Kashyap and Chellappa, 1983).

3. DISCUSSION

In sum, the MRF framework is well suited to a wide variety of image processing problems. Our exposition has been brief and we have ignored important issues like validation, choice of the order of MRF models and size of training sets. Validation, for example, takes us into the bias/variance dilemma (Geman et al., 1992). MRF models being parametric, introduce a certain kind of bias into the image representation. This seems to be the right kind of bias (in terms of reducing variance) for tasks like image restoration, tomographic reconstruction and texture segmentation. However, if the order of the chosen model is incorrect, high bias could result. It is in bias/variance terms that MRF image models should be compared alongside "mechanical" (as opposed to probabilistic) models like splines, generic representations like radial basis functions (RBFs) and *tabula rasa*, feedforward neural networks. Also, there are interesting similarities between Gauss-Markov models and thin-plate splines (Wahba, 1990). For example, the simple case of the first order Gauss-Markov model with the parameters $\theta_1 = \theta_2 = \frac{1}{4}$, is identical to the discrete membrane (first order thin-plate spline in two dimensions). Correspondences of this sort should be expected since MRF models, splines and RBFs impose local smoothness constraints albeit in different ways. Finally, there are interesting relationships between MRFs and recurrent neural networks at both computational and algorithmic levels (Rangarajan et al., 1991).

REFERENCES

*Besag, J., (1974), Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, series B, 36:192–236.

Besag, J., (1977), Efficiency of Pseudo-likelihood estimation for simple Gaussian fields. Biometrika, 64:616–618.

Chellappa, R. and Kashyap, R. L., (1985), Texture synthesis using spatial interaction models. IEEE Trans. Acoust., Speech Sig. Proc., 33:194–203.

Derin, H. and Elliott, H., (1987), Modeling and segmentation of noisy and textured images using Gibbs random fields. IEEE Trans. Patt. Anal. Mach. Intell., 9:39–55.

*Dubes, R. C. and Jain, A. K., (1989), Random field models in image analysis. Journal of Applied Statistics, 16(2):131–164.

Geiger, D. and Girosi, F., (1991), Parallel and deterministic algorithms from MRFs: Surface reconstruction. IEEE Trans. Patt. Anal. Mach. Intell., 13(5):401–412.

Geman, S., Bienenstock, E., and Doursat, R., (1992), Neural networks and the bias/variance dilemma. Neural Computation, 4:1–58.

Geman, S. and Geman, D., (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Patt. Anal. Mach. Intell., 6:721–741.

*Geman, S. and Graffigne, C., (1987), Markov random fields image models and their application to computer vision, In Gleason, A. M., editor, Proc. of the Intl. Congress of Mathematicians, 1986, Amer. Math. Soc, Providence, RI.

Kashyap, R. L. and Chellappa, R., (1983), Estimation and choice of neighbors in spatial interaction models of images. IEEE Trans. Info. theory, 29:60–72.

*Kinderman, R. and Snell, J. L., (1980), Markov Random Fields and their Applications, Amer. Math. Soc., Providence, RI.

Lee, M., Rangarajan, A., Zubal, I. G., and Gindi, G., (1993), A continuation method for emission tomography. IEEE Trans. Nuclear Science, 40:2049–2058.

*Rangarajan, A., Chellappa, R., and Manjunath, B. S., (1991), Markov random fields and neural networks with applications to early vision, In Sethi, I. K. and Jain, A. K., editors, Artificial Neural Networks and Statistical Pattern Recognition: Old and New Connections, Elsevier Science Publishers.

Wahba, G., (1990), Spline models for observational data, volume 59 of CBMS-NSF, regional conference series in applied mathematics, SIAM, Philadelphia, PA.

Woods, J. W., (1972), Two-dimensional discrete Markovian fields. IEEE Trans. Info. theory, 18:101–109.

FIGURE CAPTIONS

Figure 1.(a) Neighborhood systems for MRFs. (b) Cliques in MRFs.