Chapter 2

Mathematical MRF Models

This chapter introduces foundations of MRF theory and describes important mathematical MRF models for modeling image properties. The MRF models will be used in the subsequent chapters to derive MAP-MRF image analysis models and for MRF parameter estimation.

2.1 Markov Random Fields and Gibbs Distributions

Markov random field theory is a branch of probability theory for analyzing the spatial or contextual dependencies of physical phenomena. It is used in visual labeling to establish probabilistic distributions of interacting labels. This section introduces notations and results related to MRF's.

2.1.1 Neighborhood System and Cliques

The sites in $\mathcal{S}$ are related to one another via a neighborhood system (Section 2.12). A neighborhood system for $\mathcal{S}$ is defined as

$$\mathcal{N} = \{\mathcal{N}_i \mid \forall i \in \mathcal{S}\} \tag{2.1}$$

where $\mathcal{N}_i$ is the set of sites neighboring i. The neighboring relationship has the following properties:

(1) A site is not neighboring to itself: $i \notin \mathcal{N}_i$.

(2) The neighboring relationship is mutual: $i \in \mathcal{N}_{i'} \Longleftrightarrow i' \in \mathcal{N}_i$.

S.Z. Li, *Markov Random Field Modeling in Image Analysis*,
Advances in Pattern Recognition, DOI: 10.1007/978-1-84800-279-1_2,

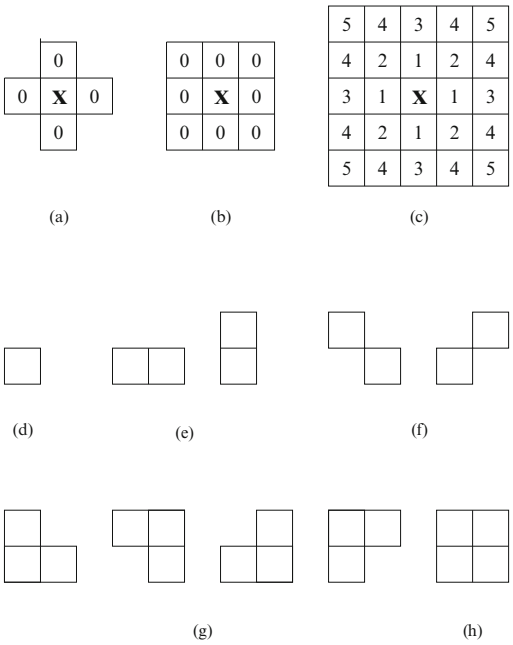

Figure 2.1: Neighborhood and cliques on a lattice of regular sites.

For a regular lattice $\mathcal{S}$, the set of neighbors of i is defined as the set of sites within a radius of $\sqrt{r}$ from i

$$\mathcal{N}_i = \{i' \in \mathcal{S} \mid [\text{dist}(\text{pixel}_{i'}, \text{pixel}_i)]^2 \leq r,\ i' \neq i\} \tag{2.2}$$

where $\text{dist}(A, B)$ denotes the Euclidean distance between A and B, and r takes an integer value. Note that sites at or near the boundaries have fewer neighbors.

In the first-order neighborhood system, also called the 4-neighborhood system, every (interior) site has four neighbors, as shown in Fig. 2.1(a) where x denotes the site considered and zeros its neighbors. In the second-order neighborhood system, also called the 8-neighborhood system, there are eight neighbors for every (interior) site, as shown in Fig. 2.1(b). The numbers $n = 1, \ldots, 5$ shown in Fig. 2.1(c) indicate the outermost neighboring sites in the nth-order neighborhood system. The shape of a neighbor set may be described as the hull enclosing all the sites in the set.

When the ordering of the elements in $\mathcal{S}$ is specified, the neighbor set can be determined more explicitly. For example, when $\mathcal{S} = \{1, \ldots, m\}$ is an ordered set of sites and its elements index the pixels of a 1D image, an interior site $i \in \{2, \ldots, m-1\}$ has two nearest neighbors, $\mathcal{N}_i = \{i-1, i+1\}$, and a site at the boundaries (the two ends) has one neighbor each, $\mathcal{N}_1 = \{2\}$ and $\mathcal{N}_m = \{m-1\}$. When the sites in a regular rectangular lattice

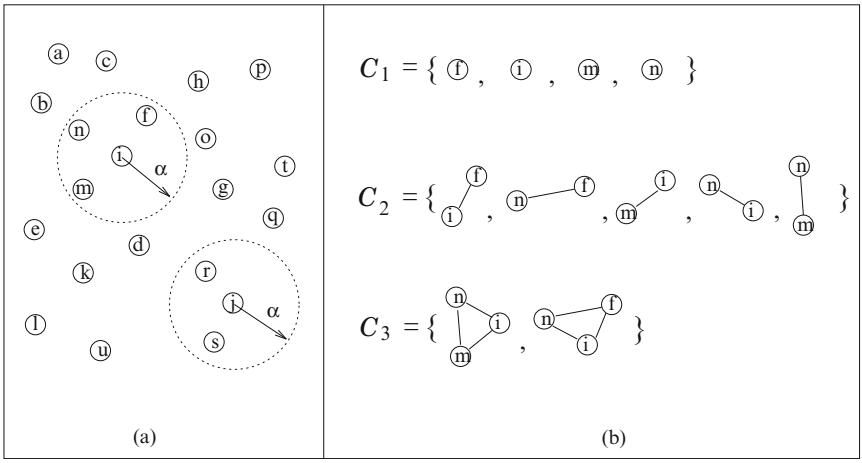

Figure 2.2: Neighborhood and cliques on a set of irregular sites.

$\mathcal{S} = \{(i,j) \mid 1 \le i,j \le n\}$ correspond to the pixels of an $n \times n$ image in the 2D plane, an internal site (i,j) has four nearest neighbors as $\mathcal{N}_{i,j} = \{(i-1,j),(i+1,j),(i,j-1),(i,j+1)\}$, a site at a boundary has three, and a site at the corners has two.

For an irregular $\mathcal{S}$, the neighbor set $\mathcal{N}_i$ of i is defined in the same way as (2.2) to comprise nearby sites within the radius of $\sqrt{r}$

$$\mathcal{N}_i = \{i' \in \mathcal{S} \mid [\text{dist}(\text{feature}_{i'}, \text{feature}_i)]^2 \le r,\ i' \ne i\} \tag{2.3}$$

The $\text{dist}(A,B)$ function needs to be defined appropriately for non-point features. Alternatively, the neighborhood may be defined by the Delaunay triangulation,[1] or its dual, the Voronoi polygon, of the sites (Besag 1975). In general, the neighbor sets $\mathcal{N}_i$ for an irregular $\mathcal{S}$ have varying shapes and sizes. Irregular sites and their neighborhoods are illustrated in Fig. 2.2(a). The neighborhood areas for sites i and j are marked by the dotted circles. The sizes of the two neighbor sets are $\#\mathcal{N}_i = 3$ and $\#\mathcal{N}_j = 2$.

The pair $(\mathcal{S},\mathcal{N}) \stackrel{\triangle}{=} \mathcal{G}$ constitutes a graph in the usual sense; $\mathcal{S}$ contains the nodes and $\mathcal{N}$ determines the links between the nodes according to the neighboring relationship. A *clique* c for $(\mathcal{S},\mathcal{N})$ is defined as a subset of sites in $\mathcal{S}$. It consists of either a single-site $c = \{i\}$, a pair of neighboring sites $c = \{i,i'\}$, a triple of neighboring sites $c = \{i,i',i''\}$, and so on. The collections of single-site, pair-site, and triple-site cliques will be denoted by $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$, respectively, where

$$\mathcal{C}_1 = \{i \mid i \in \mathcal{S}\} \tag{2.4}$$

[1] Algorithms for constructing a Delaunay triangulation in $k \ge 2$ dimensional space can be found in (Bowyer 1981; Watson 1981).

$$\mathcal{C}_2 = \{\{i, i'\} \mid i' \in \mathcal{N}_i,\ i \in \mathcal{S}\} \tag{2.5}$$

and

$$\mathcal{C}_3 = \{\{i, i', i''\} \mid i, i', i'' \in \mathcal{S} \text{ are neighbors to one another}\} \tag{2.6}$$

Note that the sites in a clique are *ordered* and $\{i, i'\}$ is not the same clique as $\{i', i\}$, and so on. The collection of all cliques for $(\mathcal{S}, \mathcal{N})$ is

$$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cdots \tag{2.7}$$

where "$\cdots$" denotes possible sets of larger cliques.

The type of a clique for $(\mathcal{S}, \mathcal{N})$ of a regular lattice is determined by its size, shape, and orientation. Figures 2.1(d)–(h) show clique types for the first- and second-order neighborhood systems for a lattice. The single-site and horizontal and vertical pair-site cliques in (d) and (e) are all those for the first-order neighborhood system (a). The clique types for the second-order neighborhood system (b) include not only those in (d) and (e) but also diagonal pair-site cliques (f) and triple-site (g) and quadruple-site (h) cliques. As the order of the neighborhood system increases, the number of cliques grows rapidly and so do the computational expenses involved.

Cliques for irregular sites do not have fixed shapes like those for a regular lattice. Therefore, their types are essentially depicted by the number of sites involved. Consider the four sites f, i, m, and n within the circle in Fig. 2.2(a), in which m and n are supposed to be neighbors to each other and so are n and f. Then the single-site, pair-site, and triple-site cliques associated with this set of sites are shown in Fig. 2.2(b). The set $\{m, i, f\}$ does not form a clique because f and m are not neighbors.

2.1.2 Markov Random Fields

Let $F = \{F_1, \ldots, F_m\}$ be a family of random variables defined on the set $\mathcal{S}$ in which each random variable F_i takes a value f_i in $\mathcal{L}$. The family F is called a random field. We use the notation $F_i = f_i$ to denote the event that F_i takes the value f_i and the notation $(F_1 = f_1,\ \ldots,\ F_m = f_m)$ to denote the joint event. For simplicity, a joint event is abbreviated as $F = f$, where $f = \{f_1, \ldots, f_m\}$ is a *configuration* of F corresponding to a realization of the field. For a discrete label set $\mathcal{L}$, the probability that random variable F_i takes the value f_i is denoted $P(F_i = f_i)$, abbreviated $P(f_i)$ unless there is a need to elaborate the expressions, and the joint probability is denoted $P(F = f) = P(F_1 = f_1, \ldots, F_m = f_m)$ and abbreviated $P(f)$. For a continuous $\mathcal{L}$, we have probability density functions (p.d.f.s) $p(F_i = f_i)$ and $p(F = f)$.

F is said to be a Markov random field on $\mathcal{S}$ w.r.t. a neighborhood system $\mathcal{N}$ if and only if the following two conditions are satisfied:

$$P(f) > 0, \quad \forall f \in \mathbb{F} \qquad \text{(positivity)} \tag{2.8}$$

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = P(f_i \mid f_{\mathcal{N}_i}) \qquad \text{(Markovianity)} \qquad (2.9)$$

where $\mathcal{S} - \{i\}$ is the set difference, $f_{\mathcal{S}-\{i\}}$ denotes the set of labels at the sites in $\mathcal{S} - \{i\}$, and

$$f_{\mathcal{N}_i} = \{f_{i'} \mid i' \in \mathcal{N}_i\} \qquad (2.10)$$

stands for the set of labels at the sites neighboring i. The positivity is assumed for some technical reasons and can usually be satisfied in practice. For example, when the positivity condition is satisfied, the joint probability $P(f)$ of any random field is uniquely determined by its local conditional probabilities (Besag 1974). The Markovianity depicts the local characteristics of F. In MRF's, only neighboring labels have direct interactions with each other. If we choose the largest neighborhood in which the neighbors of any sites include all other sites, then any F is an MRF w.r.t. such a neighborhood system.

An MRF can have other properties, such as homogeneity and isotropy. It is said to be homogeneous if $P(f_i \mid f_{\mathcal{N}_i})$ is independent of the relative location of the site i in $\mathcal{S}$. So, for a homogeneous MRF, if $f_i = f_j$ and $f_{\mathcal{N}_i} = f_{\mathcal{N}_j}$, there will be $P(f_i|f_{\mathcal{N}_i}) = P(f_j|f_{\mathcal{N}_j})$ even if $i \neq j$. The isotropy will be illustrated in the next subsection with clique potentials.

In modeling some problems, we may need to use several *coupled* MRF's; each of the MRF's is defined on one set of sites, and the sites due to different MRF's are spatially interwoven. For example, in the related tasks of image restoration and edge detection, two MRF's, one for pixel values ($\{f_i\}$) and the other for edge values ($\{l_{i,i'}\}$), can be defined on the image lattice and its dual lattice, respectively. They are coupled to each other, for example, via conditional probability $P(f_i \mid f_{i'}, l_{i,i'})$ (see Section 3.3.1).

The concept of MRF's is a generalization of that of Markov processes (MPs), which are widely used in sequence analysis. An MP is defined on a domain of time rather than space. It is a sequence (chain) of random variables $\ldots, F_1, \ldots, F_m, \ldots$ defined on the time indices $\{\ldots, 1, \ldots, m, \ldots\}$. An nth-order unilateral MP satisfies

$$P(f_i \mid \ldots, f_{i-2}, f_{i-1}) = P(f_i \mid f_{i-1}, \ldots, f_{i-n}) \qquad (2.11)$$

A bilateral or noncausal MP depends not only on the past but also on the future. An nth-order bilateral MP satisfies

$$P(f_i \mid \ldots, f_{i-2}, f_{i-1}, f_{i+1}, f_{i+2}, \ldots) = P(f_i \mid f_{i+n}, \ldots, f_{i+1}, f_{i-1}, \ldots, f_{i-n}) \qquad (2.12)$$

It is generalized into MRF's when the time indices are considered as spatial indices.

There are two approaches for specifying an MRF, that in terms of the conditional probabilities $P(f_i \mid f_{\mathcal{N}_i})$ and that in terms of the joint probability $P(f)$. Besag (1974) argued for the joint probability approach in view of the disadvantages of the conditional probability approach. First, no obvious method is available for deducing the joint probability from the associated

conditional probabilities. Second, the conditional probabilities themselves are subject to some non obvious and highly restrictive consistency conditions. Third, the natural specification of an equilibrium of a statistical process is in terms of the joint probability rather than the conditional distribution of the variables. Fortunately, a theoretical result about the equivalence between Markov random fields and Gibbs distributions (Hammersley and Clifford 1971; Besag 1974) provides a mathematically tractable means of specifying the joint probability of an MRF.

2.1.3 Gibbs Random Fields

A set of random variables F is said to be a *Gibbs random field* (GRF) on $\mathcal{S}$ w.r.t. $\mathcal{N}$ if and only if its configurations obey a *Gibbs distribution*. A Gibbs distribution takes the form

$$P(f) = Z^{-1} \times \mathrm{e}^{-\frac{1}{T}U(f)} \tag{2.13}$$

where

$$Z = \sum_{f \in \mathbb{F}} \mathrm{e}^{-\frac{1}{T}U(f)} \tag{2.14}$$

is a normalizing constant called the *partition function*, T is a constant called the *temperature*, which shall be assumed to be 1 unless otherwise stated, and $U(f)$ is the *energy function*. The energy

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) \tag{2.15}$$

is a sum of *clique potentials* $V_c(f)$ over all possible cliques $\mathcal{C}$. The value of $V_c(f)$ depends on the local configuration on the clique c. Obviously, the Gaussian distribution is a special member of this Gibbs distribution family.

A GRF is said to be homogeneous if $V_c(f)$ is independent of the relative position of the clique c in $\mathcal{S}$. It is said to be isotropic if V_c is independent of the orientation of c. It is considerably simpler to specify a GRF distribution that is homogeneous or isotropic than one without such properties. The homogeneity is assumed in most MRF vision models for mathematical and computational convenience. The isotropy is a property of direction-independent blob-like regions.

To calculate a Gibbs distribution, it is necessary to evaluate the partition function Z, which is the sum over all possible configurations in $\mathbb{F}$. Since there are a combinatorial number of elements in $\mathbb{F}$ for a discrete $\mathcal{L}$, as illustrated in Section 1.1.2, the evaluation is prohibitive even for problems of moderate size. Several approximation methods exist for solving this problem (see Chapter 8).

$P(f)$ measures the probability of the occurrence of a particular configuration, or "pattern", f. The more probable configurations are those with lower energies. The temperature T controls the sharpness of the distribution. When

the temperature is high, all configurations tend to be equally distributed. Near zero temperature, the distribution concentrates around the global energy minima. Given T and $U(f)$, we can generate a class of "patterns" by sampling the configuration space $\mathbb{F}$ according to $P(f)$; see Section 3.4.1.

For discrete labeling problems, a clique potential $V_c(f)$ can be specified by a number of *parameters*. For example, letting $f_c = (f_i, f_{i'}, f_{i''})$ be the local configuration on a triple clique $c = \{i, i', i''\}$, f_c takes a finite number of states and therefore $V_c(f)$ takes a finite number of values. For continuous labeling problems, f_c can vary continuously. In this case, $V_c(f)$ is a (possibly piecewise) continuous function of f_c.

Sometimes, it may be convenient to express the energy of a Gibbs distribution as the sum of several terms, each ascribed to cliques of a certain size, that is,

$$\begin{aligned} U(f) &= \sum_{\{i\}\in\mathcal{C}_1} V_1(f_i) + \sum_{\{i,i'\}\in\mathcal{C}_2} V_2(f_i, f_{i'}) + \\ &\quad \sum_{\{i,i',i''\}\in\mathcal{C}_3} V_3(f_i, f_{i'}, f_{i''}) + \cdots \end{aligned} \tag{2.16}$$

The above implies a homogeneous Gibbs distribution because V_1, V_2, and V_3 are independent of the locations of i, i' and i''. For nonhomogeneous Gibbs distributions, the clique functions should be written as $V_1(i, f_i)$, $V_2(i, i', f_i, f_{i'})$, and so on.

An important special case is when only cliques of size up to two are considered. In this case, the energy can also be written as

$$U(f) = \sum_{i\in\mathcal{S}} V_1(f_i) + \sum_{i\in\mathcal{S}} \sum_{i'\in\mathcal{N}_i} V_2(f_i, f_{i'}) \tag{2.17}$$

Note that in the second term on the right-hand side, $\{i, i'\}$ and $\{i', i\}$ are two distinct cliques in $\mathcal{C}_2$ because the sites in a clique are *ordered*. The conditional probability can be written as

$$P(f_i \mid f_{\mathcal{N}_i}) = \frac{\mathrm{e}^{-\left[V_1(f_i) + \sum_{i'\in\mathcal{N}_i} V_2(f_i, f_{i'})\right]}}{\sum_{f_i\in\mathcal{L}} \mathrm{e}^{-\left[V_1(f_i) + \sum_{i'\in\mathcal{N}_i} V_2(f_i, f_{i'})\right]}} \tag{2.18}$$

By incorporating (2.17) into (2.13), we can write the joint probability as the product

$$P(f) = Z^{-1} \prod_{i\in\mathcal{S}} r_i(f_i) \prod_{i\in\mathcal{S}} \prod_{i'\in\mathcal{N}_i} r_{i,i'}(f_i, f_{i'}) \tag{2.19}$$

where $r_i(f_i) = \mathrm{e}^{-\frac{1}{T}V_1(f_i)}$ and $r_{i,i'}(f_i, f_{i'}) = \mathrm{e}^{-\frac{1}{T}V_2(f_i, f_{i'})}$.

2.1.4 Markov-Gibbs Equivalence

An MRF is characterized by its local property (the Markovianity) whereas a GRF is characterized by its global property (the Gibbs distribution). The Hammersley-Clifford theorem (Hammersley and Clifford 1971) establishes the equivalence of these two types of properties. The theorem states that *F is an MRF on $\mathcal{S}$ w.r.t. $\mathcal{N}$ if and only if F is a GRF on $\mathcal{S}$ w.r.t. $\mathcal{N}$*. Many proofs of the theorem exist, e.g., in (Besag 1974), (Moussouris 1974) and (Kindermann and Snell 1980).

A proof that a GRF is an MRF is given as follows. Let $P(f)$ be a Gibbs distribution on $\mathcal{S}$ w.r.t. the neighborhood system $\mathcal{N}$. Consider the conditional probability

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = \frac{P(f_i, f_{\mathcal{S}-\{i\}})}{P(f_{\mathcal{S}-\{i\}})} = \frac{P(f)}{\sum_{f'_i \in \mathcal{L}} P(f')} \tag{2.20}$$

where $f' = \{f_1, \ldots, f_{i-1}, f'_i, \ldots, f_m\}$ is any configuration that agrees with f at all sites except possibly i. Writing out $P(f) = Z^{-1} \times \mathrm{e}^{-\sum_{c \in \mathcal{C}} V_c(f)}$ gives[2]

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = \frac{\mathrm{e}^{-\sum_{c \in \mathcal{C}} V_c(f)}}{\sum_{f'_i} \mathrm{e}^{-\sum_{c \in \mathcal{C}} V_c(f')}} \tag{2.21}$$

Divide $\mathcal{C}$ into two sets $\mathcal{A}$ and $\mathcal{B}$ with $\mathcal{A}$ consisting of cliques containing i and $\mathcal{B}$ cliques not containing i. Then (2.21) can be written as

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = \frac{\left[\mathrm{e}^{-\sum_{c \in \mathcal{A}} V_c(f)}\right] \left[\mathrm{e}^{-\sum_{c \in \mathcal{B}} V_c(f)}\right]}{\sum_{f'_i} \left\{\left[\mathrm{e}^{-\sum_{c \in \mathcal{A}} V_c(f')}\right] \left[\mathrm{e}^{-\sum_{c \in \mathcal{B}} V_c(f')}\right]\right\}} \tag{2.22}$$

Because $V_c(f) = V_c(f')$ for any clique c that does not contain i, $\mathrm{e}^{-\sum_{c \in \mathcal{B}} V_c(f)}$ cancels from both the numerator and denominator. Therefore, this probability depends only on the potentials of the cliques containing i,

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = \frac{\mathrm{e}^{-\sum_{c \in \mathcal{A}} V_c(f)}}{\sum_{f'_i} \mathrm{e}^{-\sum_{c \in \mathcal{A}} V_c(f')}} \tag{2.23}$$

that is, it depends on labels at i's neighbors. This proves that a Gibbs random field is a Markov random field. The proof that an MRF is a GRF is much more involved; a result to be described in the next subsection, which is about the uniqueness of the GRF representation (Griffeath 1976), provides such a proof.

The practical value of the theorem is that it provides a simple way of specifying the joint probability. One can specify the joint probability $P(F = f)$ by specifying the clique potential functions $V_c(f)$ and choosing appropriate

[2]This also provides a formula for calculating the conditional probability $P(f_i \mid f_{\mathcal{N}_i}) = P(f_i \mid f_{\mathcal{S}-\{i\}})$ from potential functions.

potential functions for the desired system behavior. In this way, one encodes the a priori knowledge or preference about interactions between labels.

How to choose the forms and parameters of the potential functions for a proper encoding of constraints is a major topic in MRF modeling. The forms of the potential functions determine the form of the Gibbs distribution. When all the parameters involved in the potential functions are specified, the Gibbs distribution is completely defined. Defining the functional forms is the theme in Chapters 3 and4, while estimating parameters is the subject in Chapters 7 and 8.

To calculate the joint probability of an MRF, which is a Gibbs distribution, it is necessary to evaluate the partition function (2.14). Because it is the sum over a combinatorial number of configurations in $\mathbb{F}$, the computation is usually intractable. The explicit evaluation can be avoided in maximum-probability-based MRF models when $U(f)$ contains no unknown parameters, as we will see subsequently. However, this is not true when the parameter estimation is also a part of the problem. In the latter case, the energy function $U(f) = U(f \mid \theta)$ is also a function of parameters θ and so is the partition function $Z = Z(\theta)$. The evaluation of $Z(\theta)$ is required. To circumvent the formidable difficulty therein, the joint probability is often approximated in practice. Several approximate formulae will be introduced in Chapter 7, where the problem of MRF parameter estimation is the subject.

2.1.5 Normalized and Canonical Forms

It is known that the choices of clique potential functions for a specific MRF are not unique; there may exist many equivalent choices that specify the same Gibbs distribution. However, there exists a unique normalized potential, called the *canonical potential*, for every MRF (Griffeath 1976).

Let $\mathcal{L}$ be a countable label set. A clique potential function $V_c(f)$ is said to be *normalized* if $V_c(f) = 0$, whenever for some $i \in c$, f_i takes a particular value in $\mathcal{L}$. The particular value can be any element in $\mathcal{L}$, e.g., 0 in $\mathcal{L} = \{0, 1, \ldots, M\}$. Griffeath (1976) established the mathematical relationship between an MRF distribution $P(f)$ and the unique canonical representation of clique potentials V_c in the corresponding Gibbs distribution (Griffeath 1976; Kindermann and Snell 1980). The result is described below.

Let F be a random field on a finite set $\mathcal{S}$ with local characteristics $P(f_i \mid f_{\mathcal{S}-\{i\}}) = P(f_i \mid f_{\mathcal{N}_i})$. Then F is a Gibbs field with *canonical potential function* defined by

$$V_c(f) = \begin{cases} 0 & c = \phi \\ \sum_{b \subset c} (-1)^{|c-b|} \ln P(f^b) & c \neq \phi \end{cases} \tag{2.24}$$

where ϕ denotes the empty set, $|c - b|$ is the number of elements in the set $c - b$, and

$$f_i^b = \begin{cases} f_i & \text{if } i \in b \\ 0 & \text{otherwise} \end{cases} \tag{2.25}$$

is the configuration that agrees with f on set b but assigns the value 0 to all sites outside of b. For nonempty c, the potential can also be obtained as

$$V_c(f) = \sum_{b \subset c} (-1)^{|c-b|} \ln P(f_i^b \mid f_{\mathcal{N}_i}^b) \tag{2.26}$$

where i is any element in b. Such a canonical potential function is *unique* for the corresponding MRF. Using this result, the canonical $V_c(f)$ can be computed if $P(f)$ is known.

However, in MRF modeling using Gibbs distributions, $P(f)$ is defined after $V_c(f)$ is determined, and therefore it is difficult to compute the canonical $V_c(f)$ from $P(f)$ directly. Nonetheless, there is an indirect way: Use a noncanonical representation to derive $P(f)$ and then canonicalize it using Griffeath's result to obtain the unique canonical representation.

The normalized potential functions appear to be immediately useful. For instance, for the sake of economy, one would use the minimal number of clique potentials or parameters to represent an MRF for a given neighborhood system. The concept of normalized potential functions can be used to reduce the number of nonzero clique parameters (see Chapter 7).

2.2 Auto-models

Contextual constraints on two labels are the lowest order constraints to convey contextual information. They are widely used because of their simple form and low computational cost. They are encoded in the Gibbs energy as pair-site clique potentials. With clique potentials of up to two sites, the energy takes the form

$$U(f) = \sum_{i \in \mathcal{S}} V_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \tag{2.27}$$

where "$\sum_{i \in \mathcal{S}}$" is equivalent to "$\sum_{\{i\} \in \mathcal{C}_1}$" and "$\sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i}$" equivalent to "$\sum_{\{i,i'\} \in \mathcal{C}_2}$". Equation (2.27) is a special case of (2.16), which we call a second-order energy because it involves up to pair-site cliques. It the most frequently used form because it is the simplest in form but conveys contextual information. A specific GRF or MRF can be specified by properly selecting V_1 and V_2. Some important such GRF models will be described subsequently. Derin and Kelly (1989) presented a systematic study and categorization of Markov random processes and fields in terms of what they call strict-sense Markov and wide-sense Markov properties.

When $V_1(f_i) = f_i G_i(f_i)$ and $V_2(f_i, f_{i'}) = \beta_{i,i'} f_i f_{i'}$, where $G_i(\cdot)$ are arbitrary functions and $\beta_{i,i'}$ are constants reflecting the pair-site interaction between i and i', the energy is

$$U(f) = \sum_{\{i\} \in \mathcal{C}_1} f_i G_i(f_i) + \sum_{\{i,i'\} \in \mathcal{C}_2} \beta_{i,i'} f_i f_{i'} \tag{2.28}$$

Such models are called *auto-models* (Besag 1974). The auto-models can be further classified according to assumptions made about individual f_i.

An auto-model is said to be an *auto-logistic* model if the f_i's take on values in the discrete label set $\mathcal{L} = \{0, 1\}$ (or $\mathcal{L} = \{-1, +1\}$). The corresponding energy is of the form

$$U(f) = \sum_{\{i\} \in \mathcal{C}_1} \alpha_i f_i + \sum_{\{i,i'\} \in \mathcal{C}_2} \beta_{i,i'} f_i f_{i'} \tag{2.29}$$

where $\beta_{i,i'}$ can be viewed as the *interaction coefficients*. When $\mathcal{N}$ is the nearest neighborhood system on a lattice (the four nearest neighbors on a 2D lattice or the two nearest neighbors on a 1D lattice), the auto-logistic model is reduced to the *Ising model*. The conditional probability for the auto-logistic model with $\mathcal{L} = \{0, 1\}$ is

$$P(f_i \mid f_{\mathcal{N}_i}) = \frac{\mathrm{e}^{\alpha_i f_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}}}{\sum_{f_i \in \{0,1\}} \mathrm{e}^{\alpha_i f_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}}} = \frac{\mathrm{e}^{\alpha_i f_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}}}{1 + \mathrm{e}^{\alpha_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_{i'}}} \tag{2.30}$$

When the distribution is homogeneous, we have $\alpha_i = \alpha$ and $\beta_{i,i'} = \beta$, regardless of i and i'.

An auto-model is said to be an *auto-binomial* model if the f_i's take on values in $\{0, 1, \ldots, M-1\}$ and every f_i has a conditionally binomial distribution of M trials and probability of success q

$$P(f_i \mid f_{\mathcal{N}_i}) = \binom{M-1}{f_i} q^{f_i} (1-q)^{M-1-f_i} \tag{2.31}$$

where

$$q = \frac{\mathrm{e}^{\alpha_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_{i'}}}{1 + \mathrm{e}^{\alpha_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_{i'}}} \tag{2.32}$$

The corresponding energy takes the form

$$U(f) = -\sum_{\{i\} \in \mathcal{C}_1} \ln \binom{M-1}{f_i} - \sum_{\{i\} \in \mathcal{C}_1} \alpha_i f_i - \sum_{\{i,i'\} \in \mathcal{C}_2} \beta_{i,i'} f_i f_{i'} \tag{2.33}$$

It reduces to the auto-logistic model when $M = 1$.

An auto-model is said to be an *auto-normal model*, also called a Gaussian MRF (Chellappa 1985), if the label set $\mathcal{L}$ is the real line and the joint distribution is multivariate normal. Its conditional p.d.f. is

$$p(f_i \mid f_{\mathcal{N}_i}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}[f_i - \mu_i - \sum_{i' \in \mathcal{N}_i} \beta_{i,i'}(f_{i'} - \mu_{i'})]^2} \tag{2.34}$$

It is the normal distribution with conditional mean

$$E(f_i \mid f_{\mathcal{N}_i}) = \mu_i - \sum_{i' \in \mathcal{N}_i} \beta_{i,i'}(f_{i'} - \mu_{i'}) \tag{2.35}$$

and conditional variance

$$\mathrm{var}(f_i \mid f_{\mathcal{N}_i}) = \sigma^2 \tag{2.36}$$

The joint probability is a Gibbs distribution

$$p(f) = \frac{\sqrt{\det(B)}}{\sqrt{(2\pi\sigma^2)^m}}\, \mathrm{e}^{\frac{(f-\mu)^{\mathrm{T}} B (f-\mu)}{2\sigma^2}} \tag{2.37}$$

where f is viewed as a vector, μ is the $m \times 1$ vector of the conditional means, and $B = [b_{i,i'}]$ is the $m \times m$ *interaction matrix* whose elements are unity and the off-diagonal element at (i, i') is $-\beta_{i,i'}$, i.e., $b_{i,i'} = \delta_{i,i'} - \beta_{i,i'}$ with $\beta_{i,i} = 0$. Therefore, the single-site and pair-site clique potential functions for the auto-normal model are

$$V_1(f_i) = (f_i - \mu_i)^2 / 2\sigma^2 \tag{2.38}$$

and

$$V_2(f_i, f_{i'}) = \beta_{i,i'}(f_i - \mu_i)(f_{i'} - \mu_{i'})/2\sigma^2 \tag{2.39}$$

respectively. A field of independent Gaussian noise is a special MRF whose Gibbs energy consists of only single-site clique potentials. Because all higher-order clique potentials are zero, there is no contextual interaction in the independent Gaussian noise. B is related to the covariance matrix Σ by $B = \Sigma^{-1}$. The necessary and sufficient condition for (2.37) to be a valid p.d.f. is that B be symmetric and positive definite.

A related but different model is the simultaneous auto-regression (SAR) model (Woods 1972) Unlike the auto-normal model, which is defined by the m conditional p.d.f.s, this model is defined by a set of m equations

$$f_i = \mu_i + \sum \beta_{i,i'}(f_{i'} - \mu_{i'}) + e_i \tag{2.40}$$

where e_i are independent Gaussian, $e_i \sim N(0, \sigma^2)$. It also generates the class of all multivariate normal distributions, but with joint p.d.f.s, as

$$p(f) = \frac{\det(B)}{\sqrt{(2\pi\sigma^2)^m}}\, \mathrm{e}^{\frac{(f-\mu)^{\mathrm{T}} B^{\mathrm{T}} B (f-\mu)}{2\sigma^2}} \tag{2.41}$$

where B is defined as before. Any SAR model is an auto-normal model with the B matrix in (2.37) being $B = B_2 + B_2^{\mathrm{T}} - B_2^{\mathrm{T}} B_2$, where $B_2 = B_{\mathrm{autoregressive}}$. The reverse can also be done, though in a rather unnatural way, via Cholesky decomposition (Ripley 1981). Therefore, both models can have their p.d.f.s in the form of (2.37). However, for (2.41) to be a valid p.d.f. requires only that $B_{\mathrm{autoregressive}}$ be nonsingular.

2.3 Multi-level Logistic Model

The auto-logistic model can be generalized to a *multilevel logistic* (MLL) model (Elliott et al. 1984; Derin and Cole 1986; Derin and Elliott 1987),

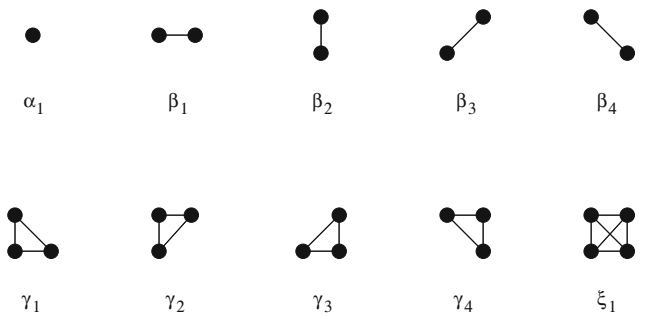

Figure 2.3: Clique types and associated potential parameters for the second-order neighborhood system. Sites are shown as dots and neighboring relationships as joining lines.

also called a Strauss process (Strauss 1977) and generalized Ising model (Geman and Geman 1984). There are M (> 2) discrete labels in the label set $\mathcal{L} = \{1, \ldots, M\}$. In this type of model, a clique potential depends on the type c (related to size, shape, and possibly orientation) of the clique and the local configuration $f_c \stackrel{\triangle}{=} \{f_i \mid i \in c\}$. For cliques containing more than one site ($\#c > 1$), the MLL clique potentials are defined by

$$V_c(f) = \begin{cases} \zeta_c & \text{if all sites on } c \text{ have the same label} \\ -\zeta_c & \text{otherwise} \end{cases} \tag{2.42}$$

where ζ_c is the potential for type c cliques; for single-site cliques, they depend on the label assigned to the site

$$V_c(f) = V_c(f_i) = \alpha_I \quad \text{if } f_i = I \in \mathcal{L}_d \tag{2.43}$$

where α_I is the potential for label value I. Figure 2.3 shows the clique types and the associated parameters in the second-order (8-neighbor) neighborhood system.

Assume that an MLL model is of second order as in (2.27), so that only the α (for single-site cliques) and β (for pair-site cliques) parameters are nonzero. The potential function for pairwise cliques is written as

$$V_2(f_i, f_{i'}) = \begin{cases} \beta_c & \text{if sites on clique } \{i, i'\} = c \in \mathcal{C}_2 \text{ have the same label} \\ -\beta_c & \text{otherwise} \end{cases} \tag{2.44}$$

where β_c is the β parameter for type c cliques and $\mathcal{C}_2$ is set of pair-site cliques. For the 4-neighborhood system, there are four types of pairwise cliques (see Fig. 2.3), and so there can be four different β_c. When the model is isotropic, all

the four neighbors take the same value. Owing to its simplicity, the pairwise MLL model (2.44) has been widely used for modeling regions and textures (Elliott et al. 1984; Geman and Geman 1984; Derin and Cole 1986; Derin and Elliott 1987; Murray and Buxton 1987; Lakshmanan and Derin 1989; Won and Derin 1992).

When the MLL model is isotropic, it depicts blob-like regions. In this case, the conditional probability can be expressed as (Strauss 1977)

$$P(f_i = I \mid f_{\mathcal{N}_i}) = \frac{\mathrm{e}^{-\alpha_I - \beta n_i(I)}}{\sum_{I=1}^{M} \mathrm{e}^{-\alpha_I - \beta n_i(I)}} \tag{2.45}$$

where $n_i(I)$ are the number of sites in $\mathcal{N}_i$ that are labeled I. It reduces to (2.30) when there are only two labels, 0 and 1. In contrast, an anisotropic model tends to generate texture-like patterns. See the examples in Section 3.4.

A hierarchical two-level Gibbs model has been proposed to represent both noise-contaminated and textured images (Derin and Cole 1986; Derin and Elliott 1987). The higher-level Gibbs distribution uses an isotropic random field (e.g., MLL) to characterize the blob-like region formation process. A lower-level Gibbs distribution describes the filling-in in each region. The filling-in may be independent noise or a type of texture, both of which can be characterized by Gibbs distributions. This provides a convenient approach for MAP-MRF modeling. In segmenting noisy and textured images (Derin and Cole 1986; Derin and Elliott 1987; Lakshmanan and Derin 1989; Hu and Fahmy 1987; Won and Derin 1992), for example, the higher-level model determines the prior of f for the region process, while the lower-level Gibbs model contributes to the conditional probability of the data given f. Note that different levels of MRF's in the hierarchy can have different neighborhood systems.

2.4 The Smoothness Prior

A generic contextual constraint on this world is the *smoothness*. It assumes that physical properties in a neighborhood of space or in an interval of time present some coherence and generally do not change abruptly. For example, the surface of a table is flat, a meadow presents a texture of grass, and a temporal event does not change abruptly over a short period of time. Indeed, we can always find regularities of a physical phenomenon w.r.t. certain properties. Since its early applications in vision (Grimson 1981; Horn and Schunck 1981; Ikeuchi and Horn 1981) aimed at imposing constraints (in addition to those from the data) on the computation of image properties, the smoothness prior has been one of the most popular prior assumptions in low-Level vision. It has been developed into a general framework, called regularization (Poggio et al. 85a; Bertero et al. 1988), for a variety of low-Level vision problems.

Smoothness constraints are often expressed as the prior probability or equivalently an energy term $U(f)$, measuring the extent to which the

smoothness assumption is violated by f. There are two basic forms of such smoothness terms corresponding to situations with discrete and continuous labels, respectively.

Equations (2.42) and (2.44) of the MLL model with negative ζ and β coefficients provide method for constructing smoothness terms for unordered, discrete labels. Whenever all labels f_c on a clique c take the same value, which means the solution f is locally smooth on c, they incur a negative clique potential (cost); otherwise, if they are not all the same, they incur a positive potential. Such an MLL model tends to give a smooth solution that prefers uniform labels.

For spatially (and also temporally in image sequence analysis) continuous MRF's, the smoothness prior often involves derivatives. This is the case with the analytical regularization (to be introduced in Section 1.3.3). There, the potential at a point is in the form of $[f^{(n)}(x)]^2$. The order n determines the number of sites in the cliques involved; for example, $[f'(x)]^2$, where $n = 1$ corresponds to a pair-site smoothness potential. Different orders imply different classes of smoothness.

Let us take continuous restoration or reconstruction of nontexture surfaces as an example. Let $f = \{f_1, \ldots, f_m\}$ be the sampling of an underlying "surface" $f(x)$ on $x \in [a, b]$, where the surface is one-dimensional for simplicity. The Gibbs distribution $P(f)$, or equivalently the energy $U(f)$, depends on the type of surface f we expect to reconstruct. Assume that the surface is flat a priori. A flat surface that has equation $f(x) = a_0$ should have zero first-order derivative, $f'(x) = 0$. Therefore, we may choose the prior energy as

$$U(f) = \int [f'(x)]^2 \mathrm{d}x \tag{2.46}$$

which is called a *string*. The energy takes the minimum value of zero only if f is absolutely flat, or a positive value otherwise. Therefore, the surface which minimizes (2.46) alone has a constant height (gray value for an image).

In the discrete case where the surface is sampled at discrete points $a \leq x_i \leq b$, $i \in \mathcal{S}$, we use the first-order difference to approximate the first derivative and use a summation to approximate the integral, so (2.46) becomes

$$U(f) = \sum_i [f_i - f_{i-1}]^2 \tag{2.47}$$

where $f_i = f(x_i)$. Expressed as the sum of clique potentials, we have

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) = \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \tag{2.48}$$

where $\mathcal{C} = \{(1,2), (2,1), (2,3), \cdots, (m-2, m-1), (m, m-1), (m-1, m)\}$ consists of only pair-site cliques and

$$V_c(f) = V_2(f_i, f_{i'}) = \frac{1}{2}(f_i - f_{i'})^2 \tag{2.49}$$

Its 2D equivalent is

$$\int\int\{[f_x(x,y)]^2+[f_y(x,y)]^2\}\mathrm{d}x\mathrm{d}y \tag{2.50}$$

and is called a *membrane*.

Similarly, the prior energy $U(f)$ can be designed for planar or quadratic surfaces. A planar surface, $f(x)=a_0+a_1x$, has zero second-order derivative, $f''(x)=0$. Therefore, one may choose

$$U(f)=\int[f''(x)]^2\mathrm{d}x \tag{2.51}$$

which is called a *rod*. The surface that minimizes (2.51) alone has a constant gradient. In the discrete case, we use the second-order difference to approximate the second-order derivative, and (2.51) becomes

$$U(f)=\sum_i[f_{i+1}-2f_i+f_{i-1}]^2 \tag{2.52}$$

For a quadratic surface, $f(x)=a_0+a_1x+a_2x^2$, the third-order derivative is zero, $f'''(x)=0$, and the prior energy may be

$$U(f)=\int[f'''(x)]^2\mathrm{d}x \tag{2.53}$$

The surface that minimizes (2.53) alone has a constant curvature. In the discrete case, we use the third-order difference to approximate the second-order derivative and (2.53) becomes

$$U(f)=\sum_i[f_{i+1}-3f_i+3f_{i-1}-f_{i-2}]^2 \tag{2.54}$$

The above smoothness models can be extended to 2D. For example, the 2D equivalent of the rod, called a plate, comes in two varieties, the quadratic variation

$$\int\int\{[f_{xx}(x,y)]^2+2[f_{xy}(x,y)]^2+[f_{yy}(x,y)]^2\}\mathrm{d}x\mathrm{d}y \tag{2.55}$$

and the squared Laplacian

$$\int\int\{f_{xx}(x,y)+f_{yy}(x,y)\}^2\mathrm{d}x\mathrm{d}y \tag{2.56}$$

The surface that minimizes one of the smoothness prior energies alone has either a constant gray level, a constant gradient, or a constant curvature. This is undesirable because constraints from other sources such as the data are not used. Therefore, a smoothness term $U(f)$ is usually utilized in conjunction

with other energy terms. In regularization, an energy consists of a smoothness term and a closeness term, and the minimal solution is a compromise between the two constraints; refer to Section 1.3.3.

The encodings of the smoothness prior in terms of derivatives usually lead to *isotropic* potential functions. This is due to the assumption that the underlying surface is nontextured. *Anisotropic* priors have to be used for texture patterns. This can be done, for example, by choosing (2.27) with direction-dependent V_2. This will be discussed in Section 3.4.

2.5 Hierarchical GRF Model

A hierarchical two-level Gibbs model has been proposed to represent both noise-contaminated and textured images (Derin and Cole 1986; Derin and Elliott 1987). The higher-level Gibbs distribution uses an isotropic random field (e.g., MLL) to characterize the blob-like region's formation process. A lower-level Gibbs distribution describes the filling-in in each region. The filling-in may be independent noise or a type of texture, both of which can be characterized by Gibbs distributions. This provides a convenient approach for MAP-MRF modeling. In segmenting noisy and textured images (Derin and Cole 1986; Derin and Elliott 1987; Lakshmanan and Derin 1989; Hu and Fahmy 1987; Won and Derin 1992), for example, the higher-level model determines the prior of f for the region process, while the lower-level Gibbs model contributes to the conditional probability of the data given f. Note that different levels of MRF's in the hierarchy can have different neighborhood systems.

Various hierarchical Gibbs models result, according to what are chosen for the regions and for the filling-ins. For example, each region may be filled in by an auto-normal texture (Manjunath et al. 1990; Won and Derin 1992) or an auto-binomial texture (Hu and Fahmy 1987); the MLL for the region formation may be substituted by another appropriate MRF. The hierarchical MRF model for textured regions will be further discussed in Section 3.4.1.

A drawback of the hierarchical model is that the conditional probability $P(d_i \mid f_i = I)$ for regions given by $\{i \in \mathcal{S} \mid f_i = I\}$ cannot always be written exactly. For example, when the lower-level MRF is a texture modeled as an auto-normal field, its joint distribution over an irregularly shaped region is not known. This difficulty may be overcome by using approximate schemes such as pseudo-likelihood (to be introduced in Section 7.1) or by using the eigenanalysis method (Wu and Leahy 1993).

2.6 The FRAME Model

The FRAME (filter, random fields and maximum entropy) model, proposed in (Zhu et al. 1997),(Zhu and Mumford 1997) and (Zhu et al. 1998), is a

generalized MRF model that fuses the essence of filtering theory and MRF modeling through the maximum entropy principle. It is generalized in the following two aspects: (1) The FRAME model is defined in terms of statistics (i.e., potential functions) calculated from the output of a filter bank by which the image is filtered, instead of the clique potentials of the image itself. Given an image (a realization of an MRF), the image is filtered by a bank of filters, giving a set of output images. Some statistics are then calculated from the output images. (2) The FRAME model provides a means of learning the model parameters from a set of samples (example images) representative of the MRF to be modeled. Besides, it also gives an algorithm for filter selection.

The joint distribution of the FRAME model is constrained in such a way that the model can reproduce the statistics of the example images. It is found by solving a constrained maximum entropy problem. Let $G^{(k)}$ $(k = 1, \ldots, K)$ be a bank of K filters (such as Gabor filters), $f^{(k)} = G^{(k)} * f$ the output of filtering f by $G^{(k)}$, and $H^{(k)} \in \mathcal{L}^{\mathcal{S}}$ (the $\mathcal{L}$ is assumed to be the same for all the K filter outputs) the histogram of $f^{(k)}$ defined by

$$H^{(k)}(I) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \delta(I - f_i^{(k)}) \tag{2.57}$$

where $\delta(t) = 1$ if $t = 0$ or 0 otherwise. For the filtered sample images, we denote the averaged histogram of the kth filter output by $\overline{H_{samp}^{(k)}}$ (averaged across all example images). Now, the joint distribution of the FRAME is defined as:

$$p(f) = \arg\max_{p} \left\{ -\int p(f) \log(p(f)) \mathrm{d}f \right\} \tag{2.58}$$

subject to

$$\overline{H_{p(f)}^{(k)}}(I) = \overline{H_{samp}^{(k)}}(I) \qquad \forall k, \forall I \tag{2.59}$$

$$\int p(f) \mathrm{d}f = 1 \tag{2.60}$$

where

$$\overline{H_{p(f)}^{(k)}} = \int H^{(k)}(f) p(f) \mathrm{d}f \tag{2.61}$$

is the expectation of $H^{(k)}$ w.r.t. $p(f)$. By using Lagrange multipliers $\theta_I^{(k)}$ for the constraints of (2.59), we get the Lagrangian

$$\begin{aligned} L(p, \theta) = & -\int p(f) \log(p(f)) \mathrm{d}f \\ & + \int_I \sum_k \theta_I^{(k)} \left\{ \int_f p(f) \sum_i \delta(I - f_i^{(k)}) \mathrm{d}f - |\mathcal{S}| \, \overline{H_{samp}^{(k)}}(I) \right\} \mathrm{d}I \end{aligned} \tag{2.62}$$

(Note that the constraints are multiplied by the factor of $|\mathcal{S}|$.) By setting $\frac{\partial L(p,\theta)}{\partial p} = 0$, the solution to the constrained optimization (ME) problem can be derived as (consider $p(f) = p(f \mid \theta)$ when θ is given)

$$\begin{aligned} p(f \mid \theta) &= \frac{1}{Z(\theta)} \mathrm{e}^{-\sum_{k=1}^{K} \sum_{i \in \mathcal{S}} \{\int \theta^{(k)}(I)\delta(I - f_i^{(k)}) \mathrm{d}I\}} \\ &= \frac{1}{Z(\theta)} \mathrm{e}^{-\sum_{k=1}^{K} \sum_{i \in \mathcal{S}} \{\theta^{(k)}(f_i^{(k)})\}} \end{aligned} \tag{2.63}$$

where $\theta^{(k)}(\cdot)$ are the potential functions of the FRAME model and Z the normalizing factor.

In the discrete form, assume that $I^{(k)} = f_i^{(k)}$ is quantized into L discrete values $I_1^{(k)}, \ldots, I_L^{(k)}$. The solution in (2.63) can be written as

$$\begin{aligned} p(f \mid \theta) &= \frac{1}{Z(\theta)} \mathrm{e}^{-\sum_{k=1}^{K} \sum_{i \in \mathcal{S}} \sum_{\ell=1}^{L} \{\theta_\ell^{(k)} \delta(I_i^{(k)} - f_i^{(k)})\}} \\ &= \frac{1}{Z(\theta)} \mathrm{e}^{-\sum_{k=1}^{K} \sum_{\ell=1}^{L} \theta_\ell^{(k)} H_\ell^{(k)}} \\ &= \frac{1}{Z} \mathrm{e}^{-<\theta, H>} \end{aligned} \tag{2.64}$$

where $\theta_\ell^{(k)} = \theta^{(k)}(I_\ell^{(k)})$, $H_\ell^{(k)} = H^{(k)}(I_\ell^{(k)})$, and $< a, b >$ is the inner product of a and b.

The FRAME model provides a means of modeling complicated high-order patterns in a tractable way. In the traditional MRF model, the neighborhood is usually small to keep the model tractable, and therefore it is difficult to model patterns in which interaction in a large neighborhood is necessary. In contrast, the FRAME model is able to model more complicated patterns by incorporating larger neighborhood and potential functions of higher-order cliques implicitly determined by the filter windows; moreover, it uses an accompanying learning procedure to estimate high-order potential functions from the filter outputs. This makes the high-order model tractable in formulation, albeit expensive in computation. There are two things to learn in the FRAME model: (1) the potential functions $\theta_I^{(k)}$; and (2) the types of filters $G^{(k)}$ to use. These will be described in Section 7.1.7.

Zhu and his colleagues (Wu et al. 2000) have established an equivalence between the FRAME model and another mathematical model of texture, called Julesz ensembles (Julesz 1962), when the size of the image lattice goes to infinity. On the other hand, they also propose fast MCMC algorithms for sampling $p(f \mid \theta)$ which involves hundreds of parameters to estimate in a large neighborhood (Zhu and Liu 2000).

2.7 Multiresolution MRF Modeling

The motivation for multiresolution MRF (MRMRF) modeling here is similar to that of FRAME modeling: to model complex and macro patterns by incorporating interactions in a large neighborhood. The approach used in MRMRF modeling is to build an MRF model based on the outputs of multiresolution filters.

From orthogonal wavelet decomposition, such as in Haar or Daubechies wavelets, nonredundant subbands can be obtained in different scales and directions. They can be used to represent the original image completely. On the other hand, these subbands are downsampled with the discrete wavelet transform. Therefore, the texture structure represented by the information of two faraway pixels in the original image may become that of immediate neighbors in the subband images on the higher-levels. Figure 2.4 shows an example of multiresolution wavelet decomposition. We can see that with the decomposition and downsampling, the subbands of different scales and directions can reveal different characteristics of the original image. The pixel relationship at different scales is not the same, even in the same direction.

Let f be an image defined on a lattice of sites $\mathcal{S}$ indexed by $i \in \mathcal{S}$, $G = \{G^{(1)}, G^{(2)}, ..., G^{(K)}\}$ a set of multiresolution filters such as wavelet filters, and $f^{(k)} = G^{(k)} * f$ the kth subband output of filtering f with $G^{(k)}$. Assume that the pixel values of the K filter outputs are quantized into M levels, giving the label set $\mathcal{L} = \{1, \ldots, M\}$, which is the same for all the K subbands. Then the distribution of image f can be written in the form

$$P(f \mid G) = \frac{1}{Z(G)} \exp(-U(f \mid G)) \tag{2.65}$$

where $Z(G)$ is the normalizing partition function. $U(f \mid G)$ is the energy function, which takes the form

$$U(f \mid G) = \sum_{k=1}^{K} \sum_{c \in \mathcal{C}} V_c^{(k)}(f) \tag{2.66}$$

where $\mathcal{C}$ is the set of all cliques in a neighborhood system and $V_c^{(k)}(f)$ is the clique potential associated with the filter output $f^{(k)}$.

Consider cliques of up to two sites. Let $\theta = \{\theta^{(k)}\} = \{\alpha^{(k)}, \beta^{(k)}\}$ be the set of MRMRF parameters, where $\alpha^{(k)} = \{\alpha^{(k)}(I)\}$ consists of M components for the eight quantized pixel values $I = f_i^{(k)}$, and $\beta^{(k)} = \{\beta_c^{(k)}\}$ consists of four components for cliques $c = (i, i')$ in the 4-neighborhood system. For homogeneous MRMRF's, the potential functions are location independent, though the pair-site clique potentials are direction dependent. The following energy function is used to include cliques of up to two sites:

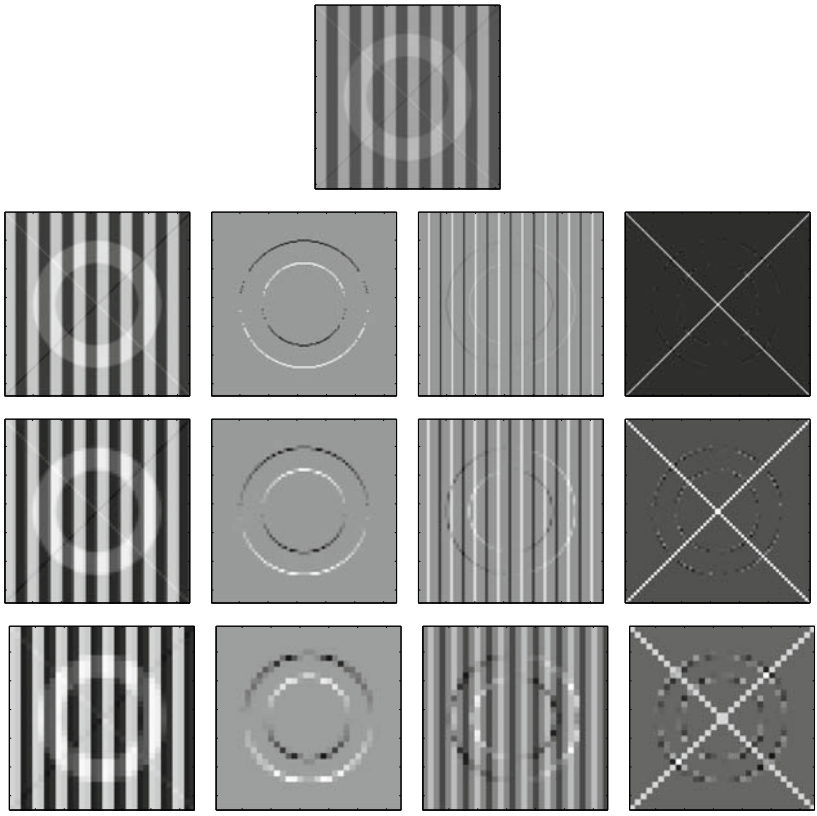

Figure 2.4: The original image (row 0). Haar wavelet decomposition at level 1 (row 1), level 2 (row 2), and level 3 (row 3). From left to right of rows 1–3 are the low-pass, horizontal, vertical, and diagonal subbands.

$$U(f \mid G, \theta) = \sum_{k=1}^{K} \sum_{i \in \mathcal{S}} \left\{ \alpha_i^{(k)}(f_i^{(k)}) + \sum_{i' \in \mathcal{N}_i, c=(i,i')} \beta_c^{(k)} \left[1 - 2\exp(-(f_i^{(k)} - f_{i'}^{(k)})^2) \right] \right\} \tag{2.67}$$

The corresponding potential functions are $V_1^{(k)}(i) = \alpha_i^{(k)}(f_i^{(k)})$ for the single-site clique and $V_2^{(k)}(i, i') = \beta_c^{(k)} \left[1 - 2\exp(-(f_i^{(k)} - f_{i'}^{(k)})^2) \right]$ for $c = (i, i')$.

In this model, the exponential form of the pair-site clique potential is similar to that of the multilevel logistic (MLL) model (see Section 2.3). But it is easier to process and more meaningful in texture representation than the MLL model (Liu and Wang 1999). In MLL, if the two labels are not exactly

the same, they contribute nothing to the potential even if they are similar; in contrast, the use of the exponential term in MRMRF incorporates the label similarity into the potential in a soft way.

Before giving the conditional probability $P(f_i|f_{\mathcal{N}_i}, G, \theta)$, we first define a generalized energy of f_i as

$$U(f_i \mid G,\theta) = \sum_{k=1}^{K}\left\{\alpha^{(k)}(f_i^{(k)}) + \sum_{i'\in\mathcal{N}_i, c=(i,i')} \beta_c^{(k)}\left[1-2\exp(-(f_i^{(k)}-f_{i'}^{(k)})^2)\right]\right\} \tag{2.68}$$

Thus

$$P(f_i|f_{\mathcal{N}_i}, G, \theta) = \frac{\sum_{k=1}^{K}\exp(-U(f_i^{(k)} \mid \theta^{(k)}))}{\sum_{k=1}^{K}\sum_{I^{(k)}=1}^{M}\exp(-U(I^{(k)} \mid \theta^{(k)}))} \tag{2.69}$$

where $I^{(k)} = f_i^{(k)}$ is the value of pixel i in the kth subband image. To define a pseudo-likelihood, we simplify it by

$$P(f_i \mid f_{\mathcal{N}_i}, G, \theta) \approx \prod_{k=1}^{K}\frac{\exp(-U(f_i^{(k)} \mid \theta^{(k)}))}{\sum_{I^{(k)}=1}^{M}\exp(-U(I^{(k)} \mid \theta^{(k)}))} \tag{2.70}$$

Then the pseudo-likelihood can be written as

$$\begin{aligned} PL(f \mid G,\theta) &= \prod_{i\in\mathcal{S}} P(f_i \mid f_{\mathcal{N}_i}, G, \theta) \\ &\approx \prod_{i\in\mathcal{S}}\prod_{k=1}^{K}\frac{\exp(-U(f_i^{(k)} \mid \theta^{(k)}))}{\sum_{I^{(k)}=1}^{M}\exp(-U(I^{(k)} \mid \theta^{(k)}))} \\ &= \prod_{k=1}^{K} PL(f^{(k)} \mid \theta^{(k)}) \end{aligned} \tag{2.71}$$

where

$$PL(f^{(k)} \mid \theta^{(k)}) = \prod_{i\in\mathcal{S}}\frac{\exp(-U(f_i^{(k)} \mid \theta^{(k)}))}{\sum_{I^{(k)}=1}^{M}\exp(-U(I^{(k)} \mid \theta^{(k)}))} \tag{2.72}$$

Thus the pseudo-likelihood $PL(f \mid G, \theta)$ can be approximated by the product of individual pseudo-likelihoods of the subband outputs $f^{(k)}$, in which an assumption is made that the parameters at different subbands are independent of each other. With this simplification, the parameter of the model can be estimated easily.

The parameters are estimated from sample data f_{samp}, which is given. The estimation can be done by maximum likelihood $P(f_{samp}^{(k)} \mid \theta^{(k)})$ or by

maximizing the pseudo-likelihood $PL(f_{samp}^{(k)} \mid \theta^{(k)})$ defined in (2.71). According to that definition, each subband can be considered as an independent MRF model,and hence the parameters of each subband can be estimated independently without considering the other subbands. Any method for MRF parameter estimation can be used to estimate parameters of each subband. The Markov chain Monte Carlo (MCMC) method (Section 7.1.6) used in the parameter estimation for the FRAME model (Section 7.1.7) would be a proper choice.

The MRMRF model attempts to incorporate information in a large neighborhood by fusing filtering theory and MRF models, which is similar to the FRAME model (Zhu et al. 1998). Compared with the traditional MRF model, the MRMRF model can reveal more information contained in the textures since the original images are decomposed into subbands of different scales and directions and downsampled. For this reason, the MRMRF model is more powerful than the traditional MRF model; however, it seems less powerful than the FRAME model since only up to pair-site interactions are considered in MRMRF. Computationally, it is also between the traditional MRF and the FRAME, the FRAME being very expensive.

2.8 Conditional Random Fields

In the MAP-MRF framework, the optimal configuration is the optimum of the posterior probability $P(f \mid d)$, or equivalently that of the joint probability $P(f,\ d) = p(d \mid f)P(f)$. The prior is formulated as an MRF, and the likelihood is due to the observation model. Usually, for tractability reasons $p(d \mid f)$ is assumed to have the factorized form (Besag 1974)

$$p(d \mid f) = \prod_{i \in \mathcal{S}} p(d_i \mid f_i) \tag{2.73}$$

even though the underlying observation model is not as simple.

The conditional random field (CRF) models the posterior probability $P(f \mid d)$ directly as an MRF without modeling the prior and likelihood individually. The label set f is said to be a CRF, given d, if every f_i satisfies the Markovianity (with positivity assumed) (Lafferty et al. 2001)

$$P(f_i \mid d,\ f_{\mathcal{S}-\{i\}}) = P(f_i \mid d,\ f_{\mathcal{N}_i}) \tag{2.74}$$

According to the Markov-Gibbs equivalence, we have

$$P(f \mid d) = \frac{1}{Z} \exp\left(-\frac{1}{T} E(f \mid d)\right) \tag{2.75}$$

where Z is the partition function and $E(f \mid d)$ the energy function. If only up to pairwise clique potentials are nonzero, the posterior probability $P(f \mid d)$ has the form

$$P(f \mid d) = \frac{1}{Z} \exp\left\{ -\sum_{i \in \mathcal{S}} V_1(f_i \mid d) - \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i,\ f_{i'} \mid d) \right\} \tag{2.76}$$

where $-V_1$ and $-V_2$ are called the association and interaction potentials, respectively, in the CRF literature (Lafferty et al. 2001). Generally, these potentials are computed as a linear combination of some feature attributes extracted from the observation.

There are two main differences between the CRF and MRF. First, in a CRF, the unary (or association) potential at site i is a function of all the observation data $d_1, \ldots, d_n$ as well as that of the label f_i; in an MRF, however, the unary potential is a function of the observations f_i and d_i only. Second, in an MRF, the pairwise (or interaction) potential for each pair of sites i and i' is independent of the observation; however, in a CRF, it is also a function of all $d_1, \ldots, d_n$ as well as that of the labels f_i and $f_{i'}$ (Lafferty et al. 2001; Ng and Jordan 2002; Rubinstein and Hastie 1997).

Therefore, a CRF may be suitable for dealing with situations where the likelihood of an MRF is not of a factorized form such that all the d_i ($\forall i \in \mathcal{S}$) can explicitly exist in both unary and pairwise potentials. Moreover, unlike in an MRF, where $d_{i'}$ can influence f_i ($i \neq i'$) indirectly through the neighborhood system, in a CRF, this is done directly by the link between $d_{i'}$ and i. The CRF has so far been used mainly in speech (1D signal) analysis. It can be extended to discriminative random fields (DRF) for image analysis as follows.

2.9 Discriminative Random Fields

While the MRF is a generative model for modeling a spatial pattern such as an image, the discriminative random field (DRF) (Kumar and Hebert 2003; Kumar 2005; Kumar and Hebert 2006) is a discriminative model that has been used for classifying patterns directly (e.g., target vs. non target classification) in images. The DRF is a special type of CRF with two extensions to it. First, a DRF is defined over 2D lattices (such as the image grid), as illustrated in Fig. 2.5. Second, the unary (association) and pairwise (interaction) potentials therein are designed using local discriminative classifiers.

The DRF of Kumar and Hebert (2003) and Kumar and Hebert (2006) defines the potentials in terms of generalized linear models as

$$V_1(f_i \mid d) = -\log\left(\sigma[f_i T_i(d)]\right) \tag{2.77}$$

$$V_2(f_i, f_{i'} \mid d) = \alpha f_i f_{i'} + \beta(2\sigma[\delta(f_i, f_{i'}) T_{i,i'}(d)] - 1) \tag{2.78}$$

where $\sigma[x]$ is the logistic function (e.g., $1/(1+\mathrm{e}^{-x})$), $T_i(d)$ and $T_{i,i'}(d)$ are functions that transform d into the unary and binary feature attributes and then linearly combine them into scalars, α and β are parameters to be learned from training examples, and $\delta(f_i, f_{i'})$ is an indication function

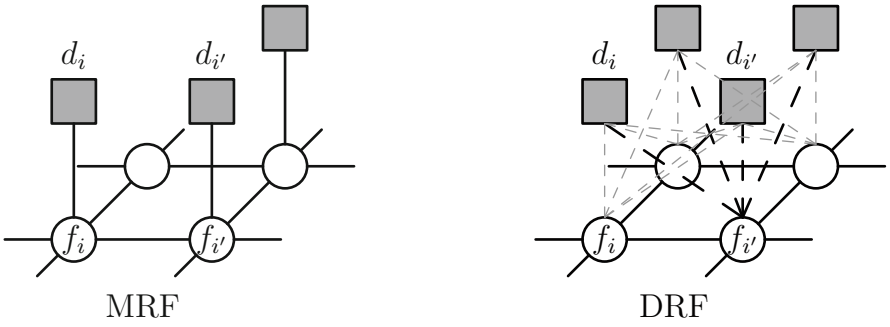

Figure 2.5: MRF vs. DRF. In an MRF, each site is connected to one observation datum. In a DRF, each site is connected to all the observation data.

$$\delta(f_i, f_{i'}) = \begin{cases} 1 & \text{if } f_i = f_{i'} \\ -1 & \text{otherwise} \end{cases} \tag{2.79}$$

In their work, the solution to the target detection problem is based on the maximum posterior marginal principle. Belief propagation and sampling algorithms are employed to find the MPM estimate.

2.10 Strong MRF Model

In addition to the (local) Markovianity (2.9) introduced in Section 2.1.2, there are two other variants of Markov properties, namely pairwise Markovianity and global Markovianity (Lauritzen 1996).

Pairwise Markovianity. A Markovianity is pairwise if for any non-adjacent sites i and i', it satisfies $P(f_i \mid f_{i'}) = P(f_i \mid f_{\mathcal{S}-\{i\}-\{i'\}})$. This means that the labels of two nonadjacent sites are independent given the labels of the other sites.

Global Markovianity. A Markovianity is global if for any disjoint subsets $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ of $\mathcal{S}$, $\mathcal{C}$ separating $\mathcal{A}$ from $\mathcal{B}$, it satisfies $P(f_\mathcal{A} \mid f_\mathcal{B}) = P(f_\mathcal{A} \mid f_\mathcal{C})$; that is, given a set of sites, the labels of any two separated subsets are independent.

Generally, pairwise Markovianity can be deduced from local Markovianity, and local Markovianity can be deduced from global Markovianity, but the reverse does not always hold (Lauritzen 1996). From this viewpoint, the local Markovianity is stronger than the pairwise Markovianity and weaker than the global Markovianity.

A strong MRF is a special case of the standard MRF (Moussouris 1974). Let $\mathcal{G} = (\mathcal{S}, \mathcal{N})$ represent a graph, and suppose F is an MRF defined on $\mathcal{G}$ w.r.t. $\mathcal{N}$. Assuming that $\mathcal{D} \subseteq \mathcal{S}$ is a subset of $\mathcal{S}$, an MRF is a strong MRF, then it satisfies

$$P(f_i \mid f_{\mathcal{D}-\{i\}}) = P(f_i \mid f_{\mathcal{N}_i \cap \mathcal{D}}) \quad \forall \mathcal{A} \subseteq \mathcal{S} \tag{2.80}$$

which is the global Markovianity for i. That is, an MRF is strong if the Markovianity holds not only w.r.t. the neighborhood system but also any of the subsets $\mathcal{D} \subseteq \mathcal{S}$ (Moussouris 1974; Paget 2004). In such a case, if the label of a neighboring site i is undefined, the label of the site i is still conditionally dependent on the labels of its neighboring sites in $\mathcal{N}_i$ that have been labeled.

While in the standard MRF the conditional distribution $P(f_i \mid f_{\mathcal{S}-\{i\}})$ is a Gibbs distribution of clique potentials, the strong MRF, based on the strong Markovianity (2.80) models $P(f_i \mid f_{\mathcal{D}-\{i\}})$ without the potentials. Therefore, it can be used to develop a nonparametric model. It has been used for texture classification in which images contain other textures of unknown origins.

2.11 $\mathcal{K}$-MRF and Nakagami-MRF Models

In a GMRF model, the joint prior distribution $p(f)$ is multivariate normal. In an analysis of ultrasound envelopes of backscattered echo and spatial interaction, the prior $p(f)$ takes the form of a $\mathcal{K}$-distribution or Nakagami distribution. Therefore, the $\mathcal{K}$-MRF (Bouhlel et al. 2004) and Nakagami-MRF (Bouhlel et al. 06a) have been proposed for the modeling problems therein.

A $\mathcal{K}$-distribution (Jakeman and Pusey 1976) with parameters (α, β) has the form

$$\mathcal{K}_{\alpha,\beta}(x) = \frac{2\beta}{\Gamma(\alpha)} \left(\frac{\beta x}{2}\right)^2 B_{\alpha-1}(\beta x) \quad \forall x \in R_+ \tag{2.81}$$

where $\Gamma(\cdot)$ is the Gamma function, α is the shape parameter, $B_{\alpha-1}(\cdot)$ is a modified Bessel function of the second kind of order $(\alpha - 1)$, and β is the scaling parameter of the $\mathcal{K}$-distribution.

The conditional density of a $\mathcal{K}$-MRF model is also a $\mathcal{K}$-distribution (Bouhlel et al. 2004; Bouhlel et al. 06b)

$$p(f_i \mid f_{\mathcal{N}_i}) \propto \mathcal{K}_{\alpha_i,\beta}(f_i) \tag{2.82}$$

where the parameter α_i is given by

$$\alpha_i = a_i + 1 + \sum_{i' \in \mathcal{N}_i} b_{i,i'} \ln f_{i'} \tag{2.83}$$

where the real valued a_i, $b_{i,i'}$, and β can be estimated from examples by solving the following system of equations (Bouhlel et al. 06b)

$$E[f_i \mid f_{\mathcal{N}_i}] = \frac{2\Gamma(\alpha_i + 0.5)}{\beta\Gamma(\alpha_i)}\Gamma(1.5) \tag{2.84}$$

$$E[f_i^2 \mid f_{\mathcal{N}_i}] = 4\frac{\alpha_i}{\beta^2} \tag{2.85}$$

where $E[\cdot]$ is the mathematical expectation.

The Nakagami distribution, with parameters (α, β), has the form

$$\mathcal{N}_{\alpha,\beta}(x) = \frac{2\beta^{\alpha}}{\Gamma(\alpha)} x^{2\alpha-1} \exp(-\beta x^2) \quad \forall x \in R_+ \tag{2.86}$$

where $\Gamma(\cdot)$ is the Gamma function. The conditional density of a Nakagami-MRF model is also a Nakagami distribution (Bouhlel et al. 06a)

$$p(f_i \mid f_{\mathcal{N}_i}) \propto \mathcal{N}_{\alpha_i,\beta}(f_i) \tag{2.87}$$

where the parameter α_i is given by

$$\alpha_i = \frac{1}{2}\left(a_i + 1 + \sum_{j \in \mathcal{N}_i} b_{i,i'} \ln f_{i'}\right) \tag{2.88}$$

where the parameters a_i, $b_{i,i'}$, and β can be estimated from examples by solving the system of equations

$$E[f_i^2 \mid f_{\mathcal{N}_i}] = \frac{\alpha_i}{\beta} \tag{2.89}$$

$$D[f_i^2 \mid f_{\mathcal{N}_i}] = \frac{\alpha_i}{\beta^2} \tag{2.90}$$

where $E[\cdot]$ is the variance.

2.12 Graphical Models: MRF's versus Bayesian Networks

The MAP-MRF approach models MRF problems defined on undirected graphs. The graphical model (GM) (or probabilistic graphical model) approach incorporates the probability theory in the manipulation of more general graphs (Pearl 1988; Jordan 1998; Jensen 2001). The graph theory part represents a complex system by a graph built on many simpler parts linked by relations and provides the data structure required by efficient algorithms. The probability theory part manipulates on the graph, provides interfaces between the model and data, and ensures consistency therein. A GM can be undirected or directed.

An undirected GM, also called a Markov network, is equivalent to a pairwise or second-order MRF. It can be denoted as $\mathcal{G} = (\mathcal{S}, \mathcal{N})$, where each node (site) is associated with a label, with or without an observation on the node, and the relationships between nodes are modeled via the neighborhood system $\mathcal{N}$.

A directed GM is denoted as $\mathcal{G} = (\mathcal{S}, \mathcal{M})$, where $\mathcal{S}$ is the set of nodes and $\mathcal{M}$ is the "parent system". If $i' \in \mathcal{S}$ is a parent node of i, then there is a directed edge from i' to i. All the nodes that i is dependent on constitute the parent set $\mathcal{M}_i$. All the $\mathcal{M}_i$s constitute the parent system $\mathcal{M}$. Such a GM can used to depict causal relationships.

A directed GM is a Bayesian network (BN) or belief network when the graph is acyclic, meaning there are no loops in the directed graph. In a BN, a node i is associated with a random variable taking a discrete or continuous value f_i. The labels and observations can be defined on disjoint subsets of nodes and related through the parent system $\mathcal{M}$. Figure 2.6 illustrates differences between an MRF and a Bayesian network. Both are referred to as inference network in machine learning literature.

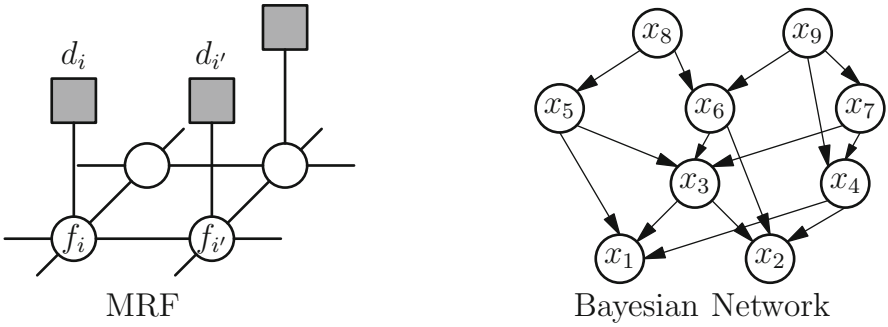

Figure 2.6: MRF vs. BN. Left: A part of an MRF, with sites (circles), labels (f_i's), and observations (squares). Right: A simple instance of a BN, where nodes x_1 and x_2 are the observation variables and the other nodes are latent variables whose values are to be inferred. The arrows show the dependency of the nodes (note that the observation nodes x_1 and x_2 depend on no other nodes).

The relationships in a BN can be described by local conditional probabilities $P(x_i \mid \mathcal{M}_i)$; if i has no parents, as for observation nodes, its local probability is considered unconditional as $P(x_i \mid \mathcal{M}_i) = P(x_i)$. The joint distribution for a BN can be expressed as the product of the local conditional probabilities

$$P(f) = P(f_1, \ldots, f_M) = \prod_{i=1}^{M} P(f_i \mid f_{\mathcal{M}_i}) \tag{2.91}$$

Similar to MRF's, issues in BNs include representation, inference (finding the optimal solution), learning (parameter estimation), decision, and application. The inference can be formulated as a maximum posterior probability or maximum marginal probability problem. Efficient algorithms exist for inference (such as belief propagation; see Section 9.3.3) and learning in BNs (Jordan 1998).