Modeling Image Analysis Problems Using Markov Random Fields

Stan Z. Li

School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore 639798

1 Introduction

Modeling problems in this article are addressed mainly from the computational viewpoint. The primary concerns are how to define an objective function for the optimal solution for an image analysis problem and how to find the optimal solution. The reason for defining the solution in an *optimization* sense is due to various uncertainties in imaging processes. It may be difficult to find the perfect solution, so we usually look for an optimal one in the sense that an objective, into which constraints are encoded, is optimized.

Contextual constraints are ultimately necessary in the interpretation of visual information. A scene is understood in the spatial and visual context of the objects in it; the objects are recognized in the context of object features at a lower level representation; the object features are identified based on the context of primitives at an even lower level; and the primitives are extracted in the context of image pixels at the lowest level of abstraction. The use of contextual constraints is indispensable for a capable image analysis system.

Markov random field (MRF) theory provides a convenient and consistent way for modeling context dependent entities such as image pixels and correlated features. This is achieved through characterizing mutual influences among such entities using conditional MRF distributions. The practical use of MRF models is largely ascribed to a theorem stating the equivalence between MRFs and Gibbs distributions which was established by Hammersley and Clifford [55] and further developed by Besag [12]. This is because the joint distribution is required in most applications but deriving the joint distribution from conditional distributions turns out to be very difficult for MRFs. The MRFs-Gibbs

[1] Email address: szli@szli.eee.ntu.edu.sg
[2] Homepage: http://markov.eee.ntu.edu.sg:8000/~szli/

Article published in Handbook of Statistics – Vol.20 ©Elsevier Science (2000) 1–43

equivalence theorem points out that the joint distribution of an MRF is a Gibbs distribution, the latter taking a simple form. This gives us a not only mathematically sound but also mathematically tractable means for statistical image analysis [50,46]. From the computational perspective, the local property of MRFs leads to algorithms which can be implemented in a local and massively parallel manner. Furthermore, MRF theory provides a foundation for multi-resolution computation [49].

For the above reasons, MRFs have been widely employed to solve image analysis problems at all levels. Most of the MRF models are for low level processing. These include image restoration and segmentation [61,56,23,36,46,21,27,85,89,92], surface reconstruction [8,53,98,17,99,24,43], edge detection [103,131,44], texture analysis [23,30,40,34,35], optical flow [64,63,78,118,60], shape from X [9,68], active contours [74,3,123], deformable templates [97,95,70] data fusion [26], visual integration, and perceptual organization [2,125]. The use of MRFs in high level, such as for object matching and recognition, has also emerged in recent years [100,29,51,4,42,76,28,87,90,91].

MRF theory tells us how to model the *a priori* probability of contextual dependent patterns, such as textures and object features. A particular MRF model favors the class of patterns encoded by itself by associating them with larger probabilities than other pattern classes. MRF theory is often used in conjunction with statistical decision and estimation theories, so as to formulate objective functions in terms of established optimality principles. *Maximum a posteriori* (MAP) probability is one of the most popular statistical criteria for optimality and in fact, has been the most popular choice in MRF modeling for image analysis. MRFs and the MAP criterion together give rise to the MAP-MRF framework adopted in this book as well as in most other MRF works. This framework, advocated by Geman and Geman (1984) [46] and others, enables us to develop algorithms for a variety of problems systematically using rational principles rather than relying on *ad hoc* heuristics [22,96,88].

An objective function is completely specified by its *form*, *i.e.* the parametric family, and the involved *parameters*. In the MAP-MRF framework, the objective is the joint posterior probability of the MRF labels. Its form and parameters are determined, in turn, according to the Bayes formula, by those of the joint prior distribution of the labels and the conditional probability of the observed data. "A particular MRF model" referred in the previous paragraph means a particular probability function (of patterns) specified by the functional form and the parameters. Two major parts of the MAP-MRF modeling is to derive the form of the posterior distribution and to determine the parameters in it, so as to completely define the posterior probability. Another important part is to design optimization algorithms for finding the maximum of the posterior distribution.

This article, which is an excerpt from Chapter 1 of [88], describes fundamentals of MRF modeling in image analysis. Basic definitions, important theoretical results, and modeling approaches are introduced. The interested reader is referred to [88] for various applications of MRF modeling in image analysis and related issues.

2 Image Labeling

Many image analysis problems can be posed as labeling problems; the solution to a problem is represented by a set of labels assigned to image pixels or features. Labeling is also a natural representation for the study of MRFs [12].

2.1 Sites and Labels

A *labeling problem* is specified in terms of a set of *sites* and a set of *labels*. Let $\mathcal{S}$ index a discrete set of m sites

$$\mathcal{S} = \{1, \ldots, m\} \tag{1}$$

in which $1, \ldots, m$ are indices. A site often represents a point or a region in the Euclidean space such as an image pixel or an image feature such as a corner point, a line segment or a surface patch. A set of sites may be categorized in terms of their "regularity". Sites on a lattice are considered as spatially *regular*. A rectangular lattice for a 2D image of size $n \times n$ can be denoted by

$$\mathcal{S} = \{(i, j) \mid 1 \leq i, j \leq n\} \tag{2}$$

Its elements correspond to the locations at which an image is sampled. Sites which do not present spatial regularity are considered as *irregular*. This is the usual case corresponding to features extracted from images at a more abstract level, such as the detected corners and lines.

We normally treat the sites in MRF models as un-ordered. For an $n \times n$ image, pixel (i, j) can be conveniently re-indexed by a single number k where k takes on values in $\{1, 2, \ldots, m\}$ with $m = n \times n$. This notation of single-number site index is used in this article also for images unless an elaboration is necessary. The inter-relationship between sites is maintained by a so-called *neighborhood system* (to be introduced later).

A label is an event that may happen to a site. Let $\mathcal{L}$ be a set of *labels*. A label set may be categorized as being continuous or discrete. In the continuous case,

a label set may correspond to the real line $\mathbb{R}$ or a compact interval of it

$$\mathcal{L}_c = [X_l, X_h] \subset \mathbb{R} \tag{3}$$

An example is the dynamic range for an analog pixel intensity. It is also possible that a continuous label takes a vector or matrix value, for example, $\mathcal{L}_c = \mathbb{R}^{a \times b}$ where a and b are dimensions.

In the discrete case, a label assumes a discrete value in a set of M labels

$$\mathcal{L}_d = \{\ell_1, \cdots, \ell_M\} \tag{4}$$

or simply

$$\mathcal{L}_d = \{1, \cdots, M\} \tag{5}$$

In edge detection, for example, the label set is $\mathcal{L}$ ={edge,non-edge}.

Besides the continuity, another essential property of a label set is the ordering of the labels. For example, elements in the continuous label set $\mathbb{R}$ (the real space) can be ordered by the relation "smaller than". When a discrete set, say $\{0, \ldots, 255\}$, represents the quantized values of intensities, it is an ordered set because for intensity values we have $0 < 1 < 2 < \ldots < 255$. When it denotes 256 different symbols such as texture types, it is considered to be un-ordered unless an artificial ordering is imposed.

For an ordered label set, a numerical (quantitative) measure of similarity between any two labels can usually be defined. For an unordered label set, a similarity measure is symbolic (qualitative), typically taking a value on "equal" or "non-equal". Label ordering and similarity not only categorize labeling problems but more importantly, affect our choices of labeling algorithms and hence the computational complexity.

2.2 The Labeling Problem

The labeling problem is to assign a label from the label set $\mathcal{L}$ to each of the sites in $\mathcal{S}$. Edge detection in an image, for example, is to assign a label f_i from the set $\mathcal{L}$ ={edge,non-edge} to site $i \in \mathcal{S}$ where elements in $\mathcal{S}$ index the image pixels. The set

$$f = \{f_1, \ldots, f_m\} \tag{6}$$

is called a *labeling* of the sites in $\mathcal{S}$ in terms of the labels in $\mathcal{L}$. When each site is assigned a unique label, $f_i = f(i)$ can be regarded as a function with domain $\mathcal{S}$ and image $\mathcal{L}$. Because the support of the function is the whole domain $\mathcal{S}$, it is a *mapping* from $\mathcal{S}$ to $\mathcal{L}$, that is,

$$f: \ \mathcal{S} \longrightarrow \mathcal{L} \tag{7}$$

Mappings with continuous and discrete label sets are demonstrated in Figure 1. A labeling is also called a *coloring* in mathematical programming.

In the terminology of random fields (*cf.* Section 3.2), a labeling is called a *configuration*. In image analysis, a configuration or labeling can correspond to an image, an edge map, an interpretation of image features in terms of object features, or a pose transformation, and so on.

When all the sites have the same label set $\mathcal{L}$, the set of all possible labelings, that is, the configuration space, is the following Cartesian product

$$\mathbb{F} = \underbrace{\mathcal{L} \times \mathcal{L} \cdots \times \mathcal{L}}_{m \text{ times}} = \mathcal{L}^m \tag{8}$$

where m is the size of $\mathcal{S}$. In image restoration, for example, $\mathcal{L}$ contains admissible pixel values which are common to all pixel sites in $\mathcal{S}$ and $\mathbb{F}$ defines all admissible images. When $\mathcal{L} = \mathbb{R}$ is the real line, $\mathbb{F} = \mathbb{R}^m$ is the m dimensional real space. When $\mathcal{L}$ is a discrete set, the size of $\mathbb{F}$ is combinatorial. For a problem with m sites and M labels, for example, there exist a total number of M^m possible configurations in $\mathbb{F}$.

In certain circumstances, admissible labels may not be common to all the sites. Consider, for example, feature based object matching. Supposing there are three types of features: points, lines and regions, then a constraint is that a certain type of image features can be labeled or interpreted in terms of the same type of model features. Therefore, the admissible label for any site is restricted to one of the three types. In an extreme case, every site i may have its own admissible set, $\mathcal{L}_i$, of labels and this gives the following configuration space

$$\mathbb{F} = \mathcal{L}_1 \times \mathcal{L}_2 \cdots \times \mathcal{L}_m \tag{9}$$

This imposes constraints on the search for wanted configurations.

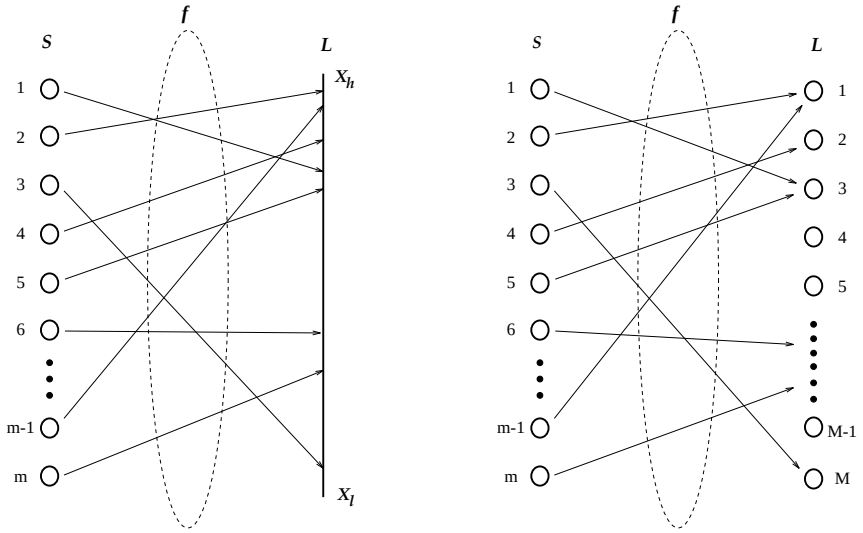

Fig. 1. A labeling of sites can be considered as a mapping from the set of sites $\mathcal{S}$ to the set of labels $\mathcal{L}$. The above shows mappings with continuous label set (left) and discrete label set (right).

2.3 Labeling Problems in Image Analysis

In terms of the regularity and the continuity, we may classify a labeling problem into one of the following four categories:

LP1: Regular sites with continuous labels.
LP2: Regular sites with discrete labels.
LP3: Irregular sites with discrete labels.
LP4: Irregular sites with continuous labels.

The first two categories characterize low level processing performed on observed images and the other two do high level processing on extracted token features. The following describes some typical problems in terms of the four categories.

Restoration or smoothing of images having continuous pixel values is an LP1. The set $\mathcal{S}$ of sites corresponds to image pixels and the set $\mathcal{L}$ of labels is a real interval. The restoration is to estimate the true image signal from a degraded or noise-corrupted image.

Restoration of binary or multi-level images is an LP2. Similar to the continuous restoration, the aim is also to estimate the true image signal from the input image. The difference is that each pixel in the resulting image here assumes a discrete value and thus $\mathcal{L}$ in this case is a set of discrete labels.

Region segmentation is an LP2. It partitions an observation image into mutually exclusive regions, each of which has some uniform and homogeneous properties whose values are significantly different from those of the neighboring regions. The property can be, for example, grey tone, color or texture. Pixels within each region are assigned a unique label.

The prior assumption in the above problems is that the signal is smooth or piecewise smooth. This is complementary to the assumption of abrupt changes made for edge detection.

Edge detection is also an LP2. Each edge site, located between two neighboring pixels, is assigned a label in {edge, non-edge} if there is a significant difference between the two pixels. Continuous restoration with discontinuities can be viewed as a combination of LP1 and LP2.

Perceptual grouping [94] is an LP3. The sites usually correspond to initially segmented features (points, lines and regions) which are irregularly arranged. The fragmentary features are to be organized into perceptually more significant features. Between each pair of the features is assigned a label in {connected,disconnected}, indicating whether the two features should be linked.

Feature-based object matching and recognition is an LP3. Each site indexes an image feature such as a point, a line segment or a region. Labels are discrete in nature and each of them indexes a model feature. The resulting configuration is a mapping from the image features to those of a model object.

Pose estimation from a set of point correspondences might be formulated as an LP4. A site is a given correspondence. A label represents an admissible (orthogonal, affine or perspective) transformation. A prior (unary) constraint is that the label of transformation itself must be orthogonal, affine or perspective. A mutual constraint is that the labels $f_1, \cdots, f_m$ should be close to each other to form a consistent transformation.

For a discrete labeling problem of m sites and M labels, there exist a total number of M^m possible labelings. For a continuous labeling problem, there are an infinite number of them. However, among all labelings, there are only a small number of them which are good solutions and may be just a few are optimal in terms of a criterion. How to define the optimal solution for a problem and how to find it are two important topics in the optimization approach to visual labeling.

The use of contextual information is ultimately indispensable in image understanding [107]. The use of contextual information in image analysis and pattern recognition dates back to [25,1]. In [25] character recognition is considered as a statistical decision problem. A nearest neighborhood dependence of pixels on an image lattice is obtained by going beyond the assumption of statistical independence. Information on the nearest neighborhood is used to calculate conditional probabilities. That system also includes parameter estimation from sample characters; recognition is done by using the estimated parameters. The work by [1] is probably the earliest work using the Markov assumption for pattern recognition. There, a Markov mesh model is used to reduce the number of parameters required for the processing using contextual constraints. Fu and Yu (1980) use MRFs defined on an image lattice to develop a class of pattern classifiers for remote sensing image classification. Another development of context-based models is relaxation labeling (RL) [115]. RL is a class of iterative procedures which use contextual constraints to reduce ambiguities in image analysis. A theory is given in [57] to explain RL from a Bayes point of view.

In probability terms, contextual constraints may be expressed locally in terms of conditional probabilities $P(f_i \mid \{f_{i'}\})$, where $\{f_{i'}\}$ denotes the set of labels at the other sites $i' \neq i$, or globally as the joint probability $P(f)$. Because local information is more directly observed, it is normal that a global inference is made based on local properties.

In situations where labels are independent of one another (no context), the joint probability is the product of the local ones

$$P(f) = \prod_{i \in \mathcal{S}} P(f_i) \tag{10}$$

The above implies conditional independence

$$P(f_i \mid \{f_{i'}\}) = P(f_i) \qquad i' \neq i \tag{11}$$

Therefore, a global labeling f can be computed by considering each label f_i locally. This is advantageous for problem solving.

In the presence of context, labels are mutually dependent. The simple relationships expressed in (10) and (11) do not hold any more. How to make a global inference using local information becomes a non-trivial task. Markov random field theory provides a mathematical foundation for solving this problem.

3 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 relevant to image analysis.

3.1 Neighborhood System and Cliques

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

$$\mathcal{N} = \{\mathcal{N}_i \mid \forall i \in \mathcal{S}\} \tag{12}$$

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: $i \in \mathcal{N}_{i'} \iff i' \in \mathcal{N}_i$.

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

$$\mathcal{N}_i = \{i' \in \mathcal{S} \mid [\text{dist}(\text{pixel}_{i'}, \text{pixel}_i)]^2 \leq r,\ i' \neq i\} \tag{13}$$

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(a) where x denotes the considered site and 0's 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(b). The numbers $n = 1, \ldots, 5$ shown in Fig.2(c) indicate the outermost neighboring sites in the n-th 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

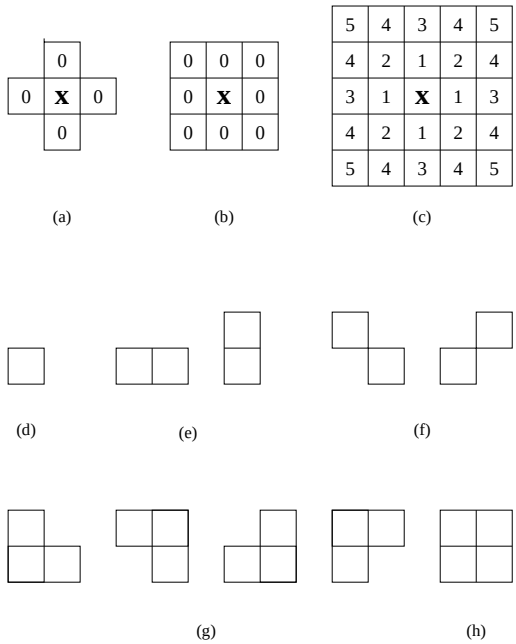

Fig. 2. Neighborhood and cliques on a lattice of regular sites.

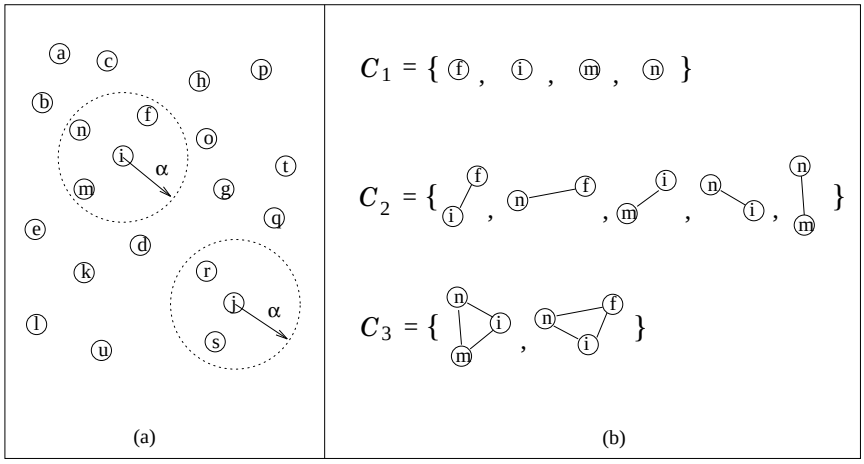

Fig. 3. Neighborhood and cliques on a set of irregular sites.

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 $\mathcal{S} = \{(i,j) \mid 1 \leq i,j \leq n\}$ 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 (13) to comprise nearby sites within the radius of $\sqrt{r}$

$$\mathcal{N}_i = \{i' \in \mathcal{S} \mid [\text{dist}(\text{feature}_{i'}, \text{feature}_i)]^2 \leq r,\ i' \neq i\} \tag{14}$$

The dist(A, B) function needs to be defined appropriately for non-point features. Alternatively, the neighborhood may be defined by the Delaunay triangulation, [3] or its dual, the Voronoi polygons, of the sites [13]. 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.3(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 either of a single site $c = \{i\}$, or of a pair of neighboring sites $c = \{i, i'\}$, or of 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

$$\mathcal{C}_1 = \{i \mid i \in \mathcal{S}\} \tag{15}$$

$$\mathcal{C}_2 = \{\{i, i'\} \mid i' \in \mathcal{N}_i,\ i \in \mathcal{S}\} \tag{16}$$

and

$$\mathcal{C}_3 = \{\{i, i', i''\} \mid i, i', i'' \in \mathcal{S} \text{ are neighbors to one another}\} \tag{17}$$

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

$$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cdots \tag{18}$$

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. Fig.2(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 grow rapidly and so the involved computational expenses.

[3] Algorithms for constructing a Delaunay triangulation in $k \geq 2$ dimensional space can be found in [18,133].

Cliques for irregular sites do not have fixed shapes as those for a regular lattice. Therefore, their types are essentially depicted by the number of involved sites. Consider the four sites f, i, m and n within the circle in Fig.3(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.3(b). The set $\{m, i, f\}$ does not form a clique because f and m are not neighbors.

3.2 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 $\mathcal{S}$ with respect to a neighborhood system $\mathcal{N}$ if and only if the following two conditions are satisfied:

$$P(f) > 0, \quad \forall f \in \mathbb{F} \qquad \text{(positivity)} \qquad (19)$$

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = P(f_i \mid f_{\mathcal{N}_i}) \qquad \text{(Markovianity)} \qquad (20)$$

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

$$f_{\mathcal{N}_i} = \{f_{i'} \mid i' \in \mathcal{N}_i\} \qquad (21)$$

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 [12]. The Markovianity depicts the local characteristics of F. In MRFs, 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 with respect to 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* MRFs; each of the MRFs is defined on one set of sites, and the sites due to different MRFs are spatially interwoven. For example, in the related tasks of image restoration and edge detection, two MRFs, 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 *e.g.* via conditional probability $P(f_i \mid f_{i'}, l_{i,i'})$.

The concept of MRFs 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 n-th order unilateral MP satisfies

$$P(f_i \mid \ldots, f_{i-2}, f_{i-1}) = P(f_i \mid f_{i-1}, \ldots, f_{i-n}) \tag{22}$$

A bilateral or non-causal MP depends not only on the past but also on the future. An n-th order bilateral MP satisfies

$$P(f_i \mid \ldots, f_{i-2}, f_{i-1}, f_{i+1}, f_{i+2}, \ldots) = P(f_i \mid f_{i+n}, \ldots, f_{i+1}, f_{i-1}, \ldots, f_{i-n}) \tag{23}$$

It is generalized into MRFs 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) argues for the joint probability approach in view of the disadvantages of the conditional probability approach: Firstly, no obvious method is available for deducing the joint probability from the associated conditional probabilities. Secondly, the conditional probabilities themselves are subject to some non-obvious and highly restrictive consistency conditions. Thirdly, the natural specification of an equilibrium of 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 [55,12] provides a mathematically tractable means of specifying the joint probability of an MRF.

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

$$P(f) = Z^{-1} \times \mathrm{e}^{-\frac{1}{T}U(f)} \tag{24}$$

where

$$Z = \sum_{f \in \mathbb{F}} \mathrm{e}^{-\frac{1}{T}U(f)} \tag{25}$$

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

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) \tag{26}$$

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 if it is homogeneous or isotropic than one without such properties. The homogeneity is assumed in most MRF 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 2.2, the evaluation is prohibitive even for problems of moderate sizes. Several Approximation methods exist for solving this problem.

$P(f)$ 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 the 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 $P(f)$.

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,

$$U(f) = \sum_{\{i\}\in\mathcal{C}_1} V_1(f_i) + \sum_{\{i,i'\}\in\mathcal{C}_2} V_2(f_i, f_{i'}) + \sum_{\{i,i',i''\}\in\mathcal{C}_3} V_3(f_i, f_{i'}, f_{i''}) + \cdots \tag{27}$$

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 non-homogeneous 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

$$U(f) = \sum_{i\in\mathcal{S}} V_1(f_i) + \sum_{i\in\mathcal{S}} \sum_{i'\in\mathcal{N}_i} V_2(f_i, f_{i'}) \tag{28}$$

Note that in the second term on the RHS, $\{i, i'\}$ 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

$$P(f_i \mid f_{\mathcal{N}_i}) = \frac{\mathrm{e}^{-\left[V_1(f_i) + \sum_{i'\in\mathcal{N}_i} V_2(f_i, f_{i'})\right]}}{\sum_{f_i\in\mathcal{L}} \mathrm{e}^{-\left[V_1(f_i) + \sum_{i'\in\mathcal{N}_i} V_2(f_i, f_{i'})\right]}} \tag{29}$$

3.4 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 [55] establishes the equivalence of these two types of properties. The theorem states that *F is an MRF on $\mathcal{S}$ with respect to $\mathcal{N}$ if and only if F is a GRF on $\mathcal{S}$ with respect to $\mathcal{N}$*. Many proofs of the theorem exist, *e.g.* in [12,102,77].

A proof that a GRF is an MRF is given as follows. Let $P(f)$ be a Gibbs

distribution on $\mathcal{S}$ with respect to the neighborhood system $\mathcal{N}$. Consider the conditional probability

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = \frac{P(f_i, f_{\mathcal{S}-\{i\}})}{P(f_{\mathcal{S}-\{i\}})} = \frac{P(f)}{\sum_{f_i' \in \mathcal{L}} P(f')} \tag{30}$$

where $f' = \{f_1, \ldots, f_{i-1}, f_i', \ldots, f_m\}$ is any configuration which agrees with f at all sites except possibly i. Writing $P(f) = Z^{-1} \times \mathrm{e}^{-\sum_{c \in \mathcal{C}} V_c(f)}$ out gives [4]

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = \frac{\mathrm{e}^{-\sum_{c \in \mathcal{C}} V_c(f)}}{\sum_{f_i'} \mathrm{e}^{-\sum_{c \in \mathcal{C}} V_c(f')}} \tag{31}$$

Divide $\mathcal{C}$ into two set $\mathcal{A}$ and $\mathcal{B}$ with $\mathcal{A}$ consisting of cliques containing i and $\mathcal{B}$ cliques not containing i. Then the above can be written as

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = \frac{\left[\mathrm{e}^{-\sum_{c \in \mathcal{A}} V_c(f)}\right] \left[\mathrm{e}^{-\sum_{c \in \mathcal{B}} V_c(f)}\right]}{\sum_{f_i'} \left\{\left[\mathrm{e}^{-\sum_{c \in \mathcal{A}} V_c(f')}\right] \left[\mathrm{e}^{-\sum_{c \in \mathcal{B}} V_c(f')}\right]\right\}} \tag{32}$$

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,

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = \frac{\mathrm{e}^{-\sum_{c \in \mathcal{A}} V_c(f)}}{\sum_{f_i'} \mathrm{e}^{-\sum_{c \in \mathcal{A}} V_c(f')}} \tag{33}$$

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 unique GRF representation [52], 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 potential functions for desired system behavior. In this way, he 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

[4] 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.

the parameters involved in the potential functions are specified, the Gibbs distribution is completely defined.

To calculate the joint probability of an MRF, which is a Gibbs distribution, it is necessary to evaluate the partition function (25). 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.

3.5 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 which specify the same Gibbs distribution. However, there exists a unique normalized potential, called the *canonical potential*, for every MRF [52].

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) [52] establishes the mathematical relationship between an MRF distribution $P(f)$ and the unique canonical representation of clique potentials V_c in the corresponding Gibbs distribution [52,77]. 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 the following:

$$V_c(f) = \begin{cases} 0 & c = \phi \\ \sum_{b \subset c} (-1)^{|c-b|} \ln P(f^b) & c \neq \phi \end{cases} \tag{34}$$

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

$$f_i^b = \begin{cases} f_i \text{ if } i \in b \\ 0 \text{ otherwise} \end{cases} \tag{35}$$

is the configuration which 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

$$V_c(f) = \sum_{b \subset c} (-1)^{|c-b|} \ln P(f_i^b \mid f_{\mathcal{N}_i}^b) \tag{36}$$

where i is any element in b. Such 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 non-canonical 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.

4 Useful MRF Models

The following introduces some useful MRF models for modeling image properties such as regions and textures. We are interested in their conditional and joint distributions, and the corresponding energy functions. The interested reader may refer to Derin and Kelly (1989) [37] for a systematic study and categorization of Markov random processes and fields in terms of what is called there strict-sense Markov and wide-sense Markov properties.

4.1 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

$$U(f) = \sum_{i \in \mathcal{S}} V_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \tag{37}$$

where "$\sum_{i \in \mathcal{S}}$" is equivalent to "$\sum_{\{i\} \in \mathcal{C}_1}$" and "$\sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i}$" to "$\sum_{\{i,i'\} \in \mathcal{C}_2}$". The above is a special case of (27), which we call a second-order energy because it involves up to pair-site cliques. It the most frequently used form owing to the mentioned feature that it is the simplest in form but conveys contextual information. A specific GRF or MRF can be specified by proper selection of V_1's and V_2's. Some important such GRF models are described subsequently. Derin and Kelly (1989) [37] present 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

$$U(f) = \sum_{\{i\} \in \mathcal{C}_1} f_i G_i(f_i) + \sum_{\{i,i'\} \in \mathcal{C}_2} \beta_{i,i'} f_i f_{i'} \tag{38}$$

The above is called *auto-models* [12]. The auto-models can be further classified according to assumptions made about individual f_i's.

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 following form

$$U(f) = \sum_{\{i\} \in \mathcal{C}_1} \alpha_i f_i + \sum_{\{i,i'\} \in \mathcal{C}_2} \beta_{i,i'} f_i f_{i'} \tag{39}$$

where $\beta_{i,i'}$ can be viewed as the *interaction coefficients*. When $\mathcal{N}$ is the nearest neighborhood system on a lattice (4 nearest neighbors on a 2D lattice or 2 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

$$P(f_i \mid f_{\mathcal{N}_i}) = \frac{\mathrm{e}^{\alpha_i f_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}}}{\sum_{f_i \in \{0,1\}} \mathrm{e}^{\alpha_i f_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}}} = \frac{\mathrm{e}^{\alpha_i f_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}}}{1 + \mathrm{e}^{\alpha_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_{i'}}} \tag{40}$$

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

$$P(f_i \mid f_{\mathcal{N}_i}) = \binom{M-1}{f_i} q^{f_i} (1-q)^{M-1-f_i} \tag{41}$$

where

$$q = \frac{\mathrm{e}^{\alpha_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_{i'}}}{1 + \mathrm{e}^{\alpha_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_{i'}}} \tag{42}$$

The corresponding energy takes the following form

$$U(f) = -\sum_{\{i\} \in \mathcal{C}_1} \ln \binom{M-1}{f_i} - \sum_{\{i\} \in \mathcal{C}_1} \alpha_i f_i - \sum_{\{i,i'\} \in \mathcal{C}_2} \beta_{i,i'} f_i f_{i'} \tag{43}$$

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 [21] , if the label set $\mathcal{L}$ is the real line and the joint distribution is multivariate normal. Its conditional p.d.f. is

$$p(f_i \mid f_{\mathcal{N}_i}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}[f_i - \mu_i - \sum_{i' \in \mathcal{N}_i} \beta_{i,i'}(f_{i'} - \mu_{i'})]^2} \tag{44}$$

It is the normal distribution with conditional mean

$$E(f_i \mid f_{\mathcal{N}_i}) = \mu_i - \sum_{i' \in \mathcal{N}_i} \beta_{i,i'}(f_{i'} - \mu_{i'}) \tag{45}$$

and conditional variance

$$var(f_i \mid f_{\mathcal{N}_i}) = \sigma^2 \tag{46}$$

The joint probability is a Gibbs distribution

$$p(f) = \frac{\sqrt{\det(B)}}{\sqrt{(2\pi\sigma^2)^m}} \, \mathrm{e}^{\frac{(f-\mu)^{\mathrm{T}} B (f-\mu)}{2\sigma^2}} \tag{47}$$

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 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

$$V_1(f_i) = (f_i - \mu_i)^2 / 2\sigma^2 \tag{48}$$

and

$$V_2(f_i, f_{i'}) = \beta_{i,i'}(f_i - \mu_i)(f_{i'} - \mu_{i'})/2\sigma^2 \tag{49}$$

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 (47) to be a valid p.d.f. is that B is symmetric and positive definite.

A related but different model is the simultaneous auto-regression (SAR) model. [136] 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

$$f_i = \mu_i + \sum \beta_{i,i'}(f_{i'} - \mu_{i'}) + e_i \tag{50}$$

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. as

$$p(f) = \frac{\det(B)}{\sqrt{(2\pi\sigma^2)^m}}\, \mathrm{e}^{\frac{(f-\mu)^{\mathrm{T}} B^{\mathrm{T}} B(f-\mu)}{2\sigma^2}} \tag{51}$$

where B is defined as before. Any SAR model is an auto-normal model with the B matrix in (47) being $B = B_2 + B_2^{\mathrm{T}} - B_2^{\mathrm{T}} B_2$ where $B_2 = B_{\mathrm{auto-regressive}}$. The reverse can also be done, though in a rather unnatural way via Cholesky decomposition [111]. Therefore, both models can have their p.d.f.'s in the form (47). However, for (51) to be a valid p.d.f., it requires only that $B_{\mathrm{auto-regressive}}$ be non-singular.

4.2 Multi-Level Logistic Model

The auto-logistic model can be generalized to *multi-level logistic* (MLL) model [40,34,35], also called Strauss process [124] and generalized Ising model [46]. There are M (> 2) discrete labels in the label set, $\mathcal{L} = \{1, \ldots, M\}$. In this type of models, 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

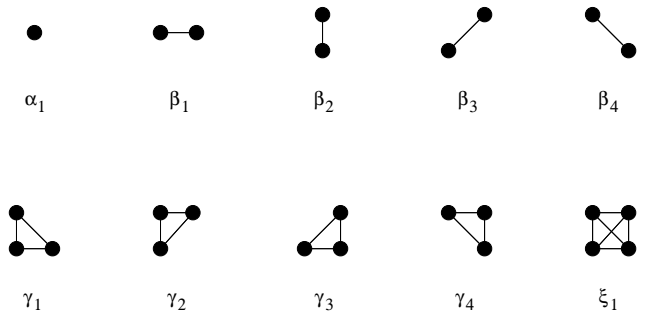

Fig. 4. Clique types and associated potential parameters for the second order neighborhood system. Sites are shown in dots and neighboring relationships in joining lines.

are defined by

$$V_c(f) = \begin{cases} \zeta_c & \text{if all sites on } c \text{ have the same label} \\ -\zeta_c & \text{otherwise} \end{cases} \tag{52}$$

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

$$V_c(f) = V_c(f_i) = \alpha_I \quad \text{if } f_i = I \in \mathcal{L}_d \tag{53}$$

where α_I is the potential for label value I. Fig.4 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 (37), so that only α (for single-site cliques) and β (for pair-site cliques) parameters are non-zero. The potential function for pair-wise cliques is written as

$$V_2(f_i, f_{i'}) = \begin{cases} \beta_c & \text{if sites on } \{i, i'\} = c \in \mathcal{C}_2 \text{ have the same label} \\ -\beta_c & \text{otherwise} \end{cases} \tag{54}$$

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 pair-wise cliques (*cf.* Fig.4) and so there can be four different β_c's. When the model is isotropic all the four take the same value. Owing to its simplicity, the pair-wise MLL model (54) has been widely used for modeling regions and textures [40,46,34,35,105,80,135].

When the MLL model is isotropic, it depicts blob-like regions. In this case, the conditional probability can be expressed as follows [124]

$$P(f_i = I \mid f_{\mathcal{N}_i}) = \frac{\mathrm{e}^{-\alpha_I - \beta n_i(I)}}{\sum_{I=1}^{M} \mathrm{e}^{-\alpha_I - \beta n_i(I)}} \tag{55}$$

where $n_i(I)$ is the number of sites in $\mathcal{N}_i$ which are labeled I. It reduces to (40) when there are only two labels, 0 and 1. In contrast, an anisotropic model tends to generate texture-like patterns.

A hierarchical two-level Gibbs model has been proposed to represent both noise-contaminated and textured images [34,35]. 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 segmentation of noisy and textured image [34,35,80,66,135], for example, the higher level determines the prior of f for the region process while the lower level Gibbs contributes to the conditional probability of the data given f. Note that different levels of MRFs in the hierarchy can have different neighborhood systems.

4.3 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 with respect to certain properties. Since its early applications [53,64,68] aimed to impose 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 problems. It has been developed into a general framework, called regularization [109,11], for a variety of low level 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.

The equations (52) and (54) of the MLL model with negative ζ and β coef-

ficients provide a method for constructing smoothness terms for un-ordered, 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 which prefers uniform labels.

For spatially (and also temporally in image sequence analysis) continuous MRFs, the smoothness prior often involves derivatives. This is the case with the analytical regularization. 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 involved cliques; for example, $[f'(x)]^2$ where $n = 1$ corresponds to a pair-site smoothness potential. Different orders implies different class of smoothness.

Let us take continuous restoration or reconstruction of non-texture 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 the surface f we expect to reconstruct. Assume that the surface is flat – *a priori*. A flat surface which has equation $f(x) = a_0$ should have zero first-order derivative, $f'(x) = 0$. Therefore, we may choose the prior energy as

$$U(f) = \int [f'(x)]^2 \mathrm{d}x \tag{56}$$

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 (56) alone has a constant height (grey 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 the above energy becomes

$$U(f) = \sum_i [f_i - f_{i-1}]^2 \tag{57}$$

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

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) = \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \tag{58}$$

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

$$V_c(f) = V_2(f_i, f_{i'}) = \frac{1}{2}(f_i - f_{i'})^2 \tag{59}$$

Its 2D equivalent is

$$\int\int\{[f_x(x,y)]^2+[f_y(x,y)]^2\}\mathrm{d}x\mathrm{d}y \tag{60}$$

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, the following may be chosen

$$U(f)=\int[f''(x)]^2\mathrm{d}x \tag{61}$$

which is called a *rod*. The surface which minimizes (61) alone has a constant gradient. In the discrete case, we use the second-order difference to approximate the second-order derivative and the above energy becomes

$$U(f)=\sum_i[f_{i+1}-2f_i+f_{i-1}]^2 \tag{62}$$

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

$$U(f)=\int[f'''(x)]^2\mathrm{d}x \tag{63}$$

The surface which minimizes the above energy alone has a constant curvature. In the discrete case, we use the third-order difference to approximate the second-order derivative and the above energy becomes

$$U(f)=\sum_i[f_{i+1}-3f_i+3f_{i-1}-f_{i-2}]^2 \tag{64}$$

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

$$\int\int\{[f_{xx}(x,y)]^2+2[f_{xy}(x,y)]^2+[f_{yy}(x,y)]^2\}\mathrm{d}x\mathrm{d}y \tag{65}$$

and the squared Laplacian

$$\int\int\{f_{xx}(x,y)+f_{yy}(x,y)\}^2\mathrm{d}x\mathrm{d}y \tag{66}$$

The surface which minimizes one of the smoothness prior energy alone has either a constant grey 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.

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 non-textured. *Anisotropic* priors have to be used for texture patterns. This can be done, for example, by choosing (37) with direction-dependent V_2's.

4.4 Hierarchical GRF Model

A hierarchical two-level Gibbs model has been proposed to represent both noise-contaminated and textured images [34,35]. 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 segmentation of noisy and textured image [34,35,80,66,135], for example, the higher level determines the prior of f for the region process while the lower level Gibbs contributes to the conditional probability of the data given f. Note that different levels of MRFs 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-in's, respectively. For example, each region may be filled in by an auto-normal texture [120,135] or an auto-binomial texture [66]; the MLL for the region formation may be substituted by another appropriate MRF.

A drawback of the hierarchical model is that the conditional probability $P(d_i \mid f_i = I)$ for regions given by $\{i \in \mathcal{S} \mid f_i = I\}$ can not 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 [13,14] (a proof of the consistency of the pseudo-likelihood estimate is given in [47]) or by using the eigen-analysis method [137].

5 Optimization-Based Approach

Optimization has been playing an essential and important role in image analysis. A problem can be formulated as optimizing a criteria, explicitly or implicitly. The extensive use of optimization principles is due to various uncertainties in imaging processes. Noise and other degradation factors, such as caused by disturbances and quantization in sensing and signal processing, are sources of uncertainties. Different appearances and poses of objects, their mutual and self occlusion and possible shape deformation also cause ambiguities in visual interpretation. Under such circumstances, we can hardly obtain exact or perfect solutions and have to resort to inexact yet optimal solutions.

In a pioneer vision system [114], object identification and pose estimation are performed using the simplest least squares (LS) fitting. Nowadays, optimization is pervasive in all aspects of image analysis, including image restoration and reconstruction [53,127,46,82,67,93], shape from shading [68], stereo, motion and optical flow [132,64,63,105,7], texture [61,73,30], edge detection [131,126], image segmentation [119,85], perceptual grouping [94,101,62], interpretation of line drawings [83], object matching and recognition [41,32,117,16,10,100,106,134,42,86,87], and pose estimation [58].

In all of the above cited examples, the solution is explicitly defined as an optimum of an objective function by which the goodness, or otherwise cost, of the solution is measured. Optimization may also be performed implicitly: the solution may optimize an objective function but in an implicit way which may or may not be realized. Hough transform [65,39,6,69] is a well-known technique for detecting lines and curves by looking at peaks of an accumulation function. It is later found to be equivalent to template matching [122] and can be reformulated as a maximizer of some probabilities such as the likelihood [59]. Edge detection was performed using some simple operators like derivatives of Gaussian [116]. The operators can be derived by using regularization principles in which an energy function is explicitly minimized [110].

We find it important to study image analysis problems from the viewpoint of optimization and to develop methodologies for optimization-based modeling. The following presents some discussions on the optimization-based approach.

5.1 Research Issues

There are three basic issues in the optimization-based approach to image analysis: problem representation, objective function and optimization algorithms. There are two aspects of a representation: descriptive and computational. The former concerns how to represent image features and object shapes, which re-

lates to photometry and geometry [79,104,72] and is not an emphasis of this article. The latter concerns how to represent the solution, which relates to the choice of sites and label set for a labeling problem. For example, in image segmentation, we may use a chain of boundary locations to represent the solution; we may alternatively use a region map to do the same job. Comparatively speaking, however, the region map is a more natural representation for MRFs.

The second issue is how to formulate an objective function for the optimization. The objective function maps a solution to a real number measuring the quality of the solution in terms of some goodness or cost. The formulation determines how various constraints, which may be pixel properties like intensity and color and/or context like relations between pixels or object features, are encoded into the function. Because the optimal solution which is the optimum of the objective function, the formulation defines the optimal solution.

The third is how to optimize the objective, *i.e.* how to search for the optimal solution in the admissible space. Two major concerns are (1) the problem of local minima existing in non-convex functions and (2) the efficiency of algorithms in space and time. They are somewhat contradictory and currently there is no algorithms which guarantee the global solution with good efficiency.

These three issues are related to one another. In the first place, the scheme of representation influences the formulation of the objective function and the design of the search algorithm. On the other hand, the formulation of an objective function affects the search. For example, suppose two objective functions have the same point as the unique global optimum but one of them is convex whereas the other is not; obviously the convex one is much more desired because it provides convenience for the search.

In the following presentation, we will be mainly dealing with minimization problems. An objective function is in the form of an energy function and is to be minimized.

5.2 Role of Energy Functions

The role of an energy function is twofold: (1) as the quantitative measure of the global quality of the solution and (2) as a guide to the search for a minimal solution. As the quantitative cost measure, an energy function defines the minimal solution as its minimum, usually a global one. In this regard, it is important to formulate an energy function so that the "correct solution" is embedded as the minimum. We call this the correctness of the formulation.

To understand an optimization approach, one should not mix problems in for-

mulation and those in search. Differentiating the two different kinds of problems helps debug the modeling. For example, if the output of an optimization procedure (assuming the implementation is correct) is not what is expected, there are two possible reasons: (1) the formulation of the objective function is not a correct one for modeling the reality and (2) the output is a low quality local minimum. Due to which one is the problem should be identified before the modeling can be improved.

The role of an energy function as a guide to the search may or may not be fully played. In real minimization, for example, when the energy function is smooth and convex w.r.t. its variables, global minimization is equivalent to local minimization and the gradient of the energy function provides sufficient information about where to search for the global solution. In this case, the role of guiding the search can be fully played. However, when the problem is non-convex, there is no general method which can efficiently utilize the energy function to guide the search. In this case, the role as the search-guide is underplayed.

In certain cases, it may be advantageous to consider the formulation of an energy function and the search simultaneously. This is to formulate the function appropriately to facilitate the search. The work of graduated non-convexity (GNC) [17] is an example in this regard. There, the energy function is deformed gradually from a convex form to its target form in the process of approximating the global solution using a gradient-based algorithm.

Local minimization in real spaces is the most mature area in optimization and many formal approaches exist for solving it. This is not so for combinatorial and global minimization. In the latter cases, heuristics become an important and perhaps necessary element in practice. In the heuristic treatment of global minimization, rather restrictive assumptions are made. An example is the bounded model [5,19]. It assumes that a measurement error is upper-bounded by a certain threshold (within the threshold, the error may be assumed to be evenly distributed). Whether the assumption is valid depends on the threshold. It is absolutely true when the threshold is infinitely large. But in practice, the threshold is almost always set to a value which is less than that required to entirely validate the bounded-error assumption. The lower the value, the higher the efficiency is, but the less general the algorithm becomes.

In the hypothesis-verification approach, efficient algorithms are used to generate hypothetical solutions, such as Hough transform [65,39], interpretation tree search [54] and geometric hashing [81]. The efficiency comes from the fast elimination of infeasible solutions, or pruning of the solution space, by taking advantage of heuristics. In this way, a relatively small number of solution candidates are picked up relatively quickly and are then verified or evaluated thoroughly, for example, by using an energy function. In this strategy, the

energy function is used for the evaluation only, not as a guide to the search.

Note that the advantage of formal approaches is in the evaluation and the advantage of heuristic approaches is in the search. A good strategy for the overall design of a specialized system may be the following: use a heuristic algorithm to quickly find a small number of solution candidates and then evaluate the found candidates using an energy function derived formally to give the best solution.

5.3 Formulation of Objective Functions

In pattern recognition, there are two basic approachesto formulating an energy function: parametric and nonparametric. In the parametric approach, the types of underlying distributions are known and the distributions are parameterized by a few parameters. Therefore, the functional form of the energy can be obtained and the energy function is completely defined when the parameters are specified.

In the nonparametric approach, sometimes called distribution free approach, no assumptions about the distributions are made. There, a distribution is either estimated from the data or approximated by a pre-specified basis functions with several unknown parameters in it to be estimated. In the latter case, the pre-specified basis functions will determine the functional form of the energy.

Despite the terms parametric and nonparametric, both approaches are somewhat parametric in nature. This is because in any case, there are always parameters that must be determined to define the energy function.

The two most important aspects of an energy function are its form and the involved parameters. The form and parameters together define the energy function which in turn defines the minimal solution. The form depends on assumptions about the solution f and the observed data d. We express this using the notation $E(f \mid d)$. Denote the set of involved parameters by θ. With θ, the energy is expressed further as $E(f \mid d, \theta)$. In general, given the functional form for E, a different d or θ defines a different energy function, $E(f \mid d, \theta)$, w.r.t. f and hence a (possibly) different minimal solution f^*.

Since the parameters are part of the definition of the energy function $E(f \mid d, \theta)$, the minimal solution $f^* = \arg\min_f E(f \mid d)$ is not completely defined if the parameters are not specified even if the functional form is known. These parameters must be specified or estimated by some means. This is an important area of study in the MRF modeling.

In formal models, as opposed to heuristic ones, an energy function is formulated based on an established criterion. Because of inevitable uncertainties in imaging and vision processes, principles from statistics, probability and information theory are often used as the formal basis. When the knowledge about the data distribution is avaliable but not about the prior information, the *maximum likelihood* (ML) criterion may be used, $f^* = \arg\max P(d \mid f)$. On the other hand, if only the prior information is available, the *maximum entropy* criterion may be chosen, $f^* = \arg\max - \sum_{i=1}^{m} P(f_i) \ \ln P(f_i)$. The maximum entropy criterion is simply taking this fact into account: Configurations with higher entropy are more likely because nature can generate them in more ways [71].

When both the prior and likelihood distributions are known, the best result is achieved by that maximizes a Bayes criterion according to Bayes statistics [129]. Bayes statistics is a theory of fundamental importance in estimation and decision making. Although there have been philosophical and scientific controversies about their appropriateness in inference and decision making (see [26] for a short review), Bayes criteria, the MAP principle in particular, are the most popular ones in image analysis and in fact, MAP is the most popular criterion in optimization-based MRF modeling. The equivalence theorem of between Markov random fields and Gibbs distribution established in Section 3.4 provides a convenient way for specifying the joint prior probability, solving a difficult issue in MAP-MRF labeling.

In the principle of *minimum description length* (MDL) [112,113], the optimal solution to a problem is that needs the smallest set of vocabulary in a given language for explaining the input data. The MDL has close relationships to the statistical methods such as the ML and MAP [113]. For example, if $P(f)$ is related to the description length and $P(d \mid f)$ related to the description error, then MDL is equivalent to MAP. However, it is a more natural and intuitive when prior probabilities are not well defined. The MDL has been used for image analysis problems at different levels such as segmentation [82,108,31,33,75] and object recognition [20].

6 Bayes Labeling of MRFs

Bayes statistics is a theory of fundamental importance in estimation and decision making. According to this theory, when both the prior distribution and the likelihood function of a pattern are known, the best that can be estimated from these sources of knowledge is the Bayes labeling. The maximum *a pos-*

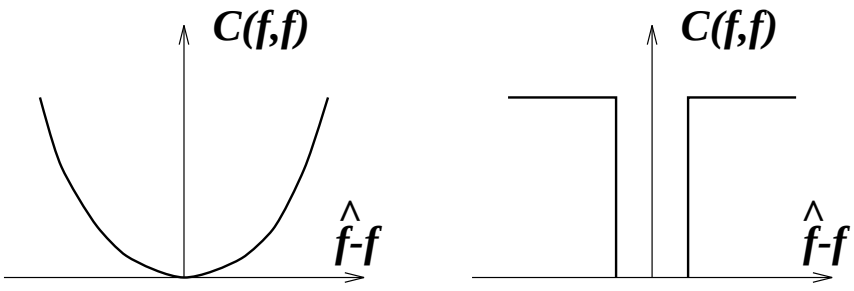

Fig. 5. Two choices of cost functions.

terior (MAP) solution, as a special case in the Bayes framework, is sought in many image analysis algorithms.

The MAP-MRF framework is advocated by Geman and Geman (1984) and others [48,35,47,38,15,125,45]. Since the paper of [46], numerous problems have been formulated in this framework. This section reviews related concepts and derives involved probabilistic distributions and energies in MAP-MRF labeling. For more detailed materials on Bayes theory, the reader is referred to books like [129].

6.1 Bayes Estimation

In Bayes estimation, a risk is minimized to obtain the optimal estimate. The Bayes risk of estimate f^* is defined as

$$R(f^*) = \int_{f \in \mathbb{F}} C(f^*, f) P(f \mid d) df \tag{67}$$

where d is the observation, $C(f^*, f)$ is a cost function and $P(f \mid d)$ is the posterior distribution. First of all, we need to compute the posterior distribution from the prior and the likelihood. According to the Bayes rule, the posterior probability can be computed by using the following formulation

$$P(f \mid d) = \frac{p(d \mid f) P(f)}{p(d)} \tag{68}$$

where $P(f)$ is the prior probability of labelings f, $p(d \mid f)$ is the conditional p.d.f. of the observations d, also called the likelihood function of f for d fixed, and $p(d)$ is the density of d which is a constant when d is given.

The cost function $C(f^*, f)$ determines the cost of estimate f when the truth is f^*. It is defined according to our preference. Two popular choices are the

quadratic cost function

$$C(f^*, f) = \|f^* - f\|^2 \tag{69}$$

where $\|a - b\|$ is a distance between a and b, and the δ (0-1) cost function

$$C(f^*, f) = \begin{cases} 0 \text{ if } \|f^* - f\| \leq \delta \\ 1 \text{ otherwise} \end{cases} \tag{70}$$

where $\delta > 0$ is any small constant. A plot of the two cost functions are shown in Fig.5.

The Bayes risk under the quadratic cost function measures the variance of the estimate

$$R(f^*) = \int_{f \in \mathbb{F}} \|f^* - f\|^2 P(f \mid d) df \tag{71}$$

Letting $\frac{\partial R(f^*)}{\partial f^*} = 0$, we obtain the minimal variance estimate

$$f^* = \int_{f \in \mathbb{F}} f P(f \mid d) df \tag{72}$$

The above is the mean of the posterior probability.

For the δ cost function, the Bayes risk is

$$R(f^*) = \int_{f : \|f^* - f\| > \delta} P(f \mid d) df = 1 - \int_{f : \|f^* - f\| \leq \delta} P(f \mid d) df \tag{73}$$

When $\delta \to 0$, the above is approximated by

$$R(f^*) = 1 - \kappa P(f \mid d) \tag{74}$$

where κ is the volume of the space containing all points f for which $\|f^* - f\| \leq \delta$. Minimizing the above is equivalent to maximizing the posterior probability. Therefore, the minimal risk estimate is

$$f^* = \arg\max_{f \in \mathbb{F}} P(f \mid d) \tag{75}$$

which is known as the MAP estimate. Because $p(d)$ in (68) is a constant for a fixed d, $P(f \mid d)$ is proportional to the joint distribution

$$P(f \mid d) \propto P(f, d) = p(d \mid f)P(f) \tag{76}$$

Then the MAP estimate is equivalently found by

$$f^* = \arg\max_{f \in \mathbb{F}} \{p(d \mid f)P(f)\} \tag{77}$$

Obviously, when the prior distribution, $P(f)$, is flat, the MAP is equivalent to the maximum likelihood.

6.2 MAP-MRF Labeling

In the MAP-MRF labeling, $P(f \mid d)$ is the posterior distribution of an MRF. An important step in Bayes labeling of MRFs is to derive this distribution. Here we use a simple example to illustrate the formulation of a MAP-MRF labeling problem. The problem is to restore images from noisy data. Assuming that the image surfaces are flat, then the joint prior distribution of f is

$$P(f) = \frac{1}{Z} \mathrm{e}^{-U(f)} \tag{78}$$

where $U(f) = \sum_i \sum_{i' \in \{i-1, i+1\}} (f_i - f_{i'})^2$ is the *prior energy* for the type of surfaces. Assuming that the observation is the true surface height plus the independent Gaussian noise, $d_i = f_i + e_i$, where $e_i \sim N(\mu, \sigma^2)$, then the likelihood distribution is

$$p(d \mid f) = \frac{1}{\Pi_{i=1}^{m} \sqrt{2\pi\sigma^2}} \mathrm{e}^{-U(d \mid f)} \tag{79}$$

where

$$U(d \mid f) = \sum_{i=1}^{m} (f_i - d_i)^2 / 2\sigma^2 \tag{80}$$

is the *likelihood energy*. Now the posterior probability is

$$P(f \mid d) \propto \mathrm{e}^{-U(f \mid d)} \tag{81}$$

where

$$U(f \mid d) = U(d \mid f) + U(f) = \sum_{i=1}^{m} (f_i - d_i)^2 / 2\sigma_i^2 + \sum_{i=1}^{m} (f_i - f_{i-1})^2 \quad (82)$$

is the *posterior energy*. The MAP estimate is equivalently found by minimizing the posterior energy function

$$f^* = \arg\min_f U(f \mid d) \quad (83)$$

There is only one parameter in this simple example, σ_i. When it is determined, $U(f \mid d)$ is fully specified and the MAP-MRF solution is completely defined.

* *

This text has described the most important results in MRF theory pertinent to image analysis, and introduced general approaches for solving problems therein. The interested reader is referred [88] for more specific topics, techniques and algorithms; including the use of MRFs in various applications ranging from image restoration, segmentation texture modeling at lower level to object matching and recognition at higher level, and related issues such as how to cope with discontinuities, relations with robust estimators, modeling parameter estimation for image and texture analysis and for object recognition, and local and global optimization techniques and algorithms.

Acknowledgement This work was supported by NTU-AcRF RG 43/95 and RG 51/97.

References

[1] K. Abend, T. J. Harley, and L. N. Kanal. "Classification of binary random patterns". *IEEE Transactions on Information Theory*, 11(4):538–544, 1965.

[2] J. Aloimonos and D. Shulman. *Integration of Visual Modules.* Academic Press, London, UK, 1989.

[3] A. Amini, S. Tehrani, and T. Weymouth. "Using dynamic programming for minimizing the energy of active contours in the presence of hard constraints". In *Proceedings of IEEE International Conference on Computer Vision*, pages 95–99, 1988.

[4] A. J. Baddeley and M. N. M. van Lieshout. "Object recognition using Markov spatial processes". In *Proceedings of International Conference Pattern Recognition*, volume B, pages 136–139, 1992.

[5] H. S. Baird. *Model-based image matching using location.* MIT Press, Cambridge, Mass, 1985.

[6] D. H. Ballard. "Generalizing the Hough transform to detect arbitary shapes". *Pattern Recognition*, 13(2):111–122, 1981.

[7] S. T. Barnard. "Stereo matching by hierarchical, microcanonical annealing". In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 832–835, 1987.

[8] H. G. Barrow and J. M. Tenenbaum. "Computational vision". *Proceedings of the IEEE*, 69(5):572–595, May 1981.

[9] H. G. Barrow and J. M. Tenenbaum. "Interpreting line drawings as three dimensional surfaces". *Artificial Intelligence*, 17:75–117, 1981.

[10] J. Ben-Arie and A. Z. Meiri. 3d objects recognition by optimal matching search of multinary relations graphs. *Computer Vision, Graphics and Image Processing*, 37:345–361, 1987.

[11] M. Bertero, T. A. Poggio, and V. Torre. "Ill-posed problems in early vision". *Proceedings of the IEEE*, 76(8):869–889, August 1988.

[12] J. Besag. "Spatial interaction and the statistical analysis of lattice systems" (with discussions). *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[13] J. Besag. "Statistical analysis of non-lattice data". *The Statistician*, 24(3):179–195, 1975.

[14] J. Besag. "Efficiency of pseudo-likelihood estimation for simple Gaussian fields". *Biometrika*, 64:616–618, 1977.

[15] J. Besag. "Towards Bayesian image analysis". *Journal of Applied Statistics*, 16(3):395–406, 1989.

[16] B. Bhanu and O. D. Faugeras. "Shape matching of two-dimensional objects". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):137–155, March 1984.

[17] A. Blake and A. Zisserman. *Visual Reconstruction.* MIT Press, Cambridge, MA, 1987.

[18] A. Bowyer. "Computing Dirichlet tessellations". *Computer Journal*, 24:162–166, 1981.

[19] T. M. Breuel. "Fast recognition using adaptive subdivision of transformation space". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 445–451, 1992.

[20] T. M. Breuel. "Higher-order statistics in visual object recognition". Memo #93-02, IDIAP, Martigny, Switzerland, June 1993.

[21] R. Chellappa. "Two-dimensional discrete gaussian Markov random field models for image processing". In L. N. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition 2*, pages 79–112, 1985.

[22] R. Chellappa and A. Jain, editors. *Markov Random Fields: Theory and Applications.* Academic Press, 1993.

[23] R. Chellappa and R. L. Kashyap. "Digital image restoration using spatial interaction models". *IEEE Transactions on Acoustic, Speech and Signal Processing*, 30:461–472, 1982.

[24] P. B. Chou and C. M. Brown. "The theory and practice of Bayesian image labeling". *International Journal of Computer Vision*, 4:185–210, 1990.

[25] C. K. Chow. "A recognition method using neighbor dependence". *IRE Transactions on Electronic Computer"*, 11:683–690, 1962.

[26] J. J. Clark and A. L. Yuille. *Data Fusion for Sensory Information Processing Systems.* Kluwer Academic Publishers, Norwell, MA, 1990.

[27] F. S. Cohen and D. B. Cooper. "Simple parallel hierarchical and relaxation algorithms for segmenting noncasual Markovian random fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(2):195–218, March 1987.

[28] D. B. Cooper, J. Subrahmonia, Y. P. Hung, and B. Cernuschi-Frias. "The use of Markov random fields in estimating amd recognizing object in 3D space". In R. Chellappa and A. Jain, editors, *Marokov Random Fields: Theory and Applications*, pages 335–367, Boston, 1993. Academic Press.

[29] P. R. Cooper. "Parallel structure recognition with uncertainty: Coupled segmentation and matching". In *Proceedings of IEEE International Conference on Computer Vision*, pages 287–290, 1990.

[30] G. C. Cross and A. K. Jain. "Markov random field texture models". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(1):25–39, January 1983.

[31] T. Darrell, S. Sclaroff, and A. Pentland. "Segmentation by minimal description". In *Proceedings of IEEE International Conference on Computer Vision*, pages 112–116, 1990.

[32] L. S. Davis. "Shape matching using relaxation techniques". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):60–72, January 1979.

[33] J. Dengler. "Estimation of discontinuous displacement vector fields with the minimum description length criterion". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 276–282, 1991.

[34] H. Derin and W. S. Cole. "Segmentation of textured images using using Gibbs random fields". *Computer Vision, Graphics and Image Processing*, 35:72–98, 1986.

[35] H. Derin and H. Elliott. "Modeling and segmentation of noisy and textured images using Gibbs random fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):39–55, January 1987.

[36] H. Derin, H. Elliott, R. Cristi, and D. Geman. "Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):707–720, November 1984.

[37] H. Derin and P. A. Kelly. "Discrete-index Markov-type random fields". *Proceedings of the IEEE*, 77(10):1485–1510, October 1989.

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

[39] R. O. Duda and P. E. Hart. "Use of Hough transform to detect lines and curves in picture". *Communications of the ACM*, 15(1):11–15, 1972.

[40] H. Elliott, H. Derin, R. Cristi, and D. Geman. "Application of the Gibbs distribution to image segmentation". In *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, pages 32.5.1–32.5.4, San Diego, March 1984.

[41] M. Fischler and R. Elschlager. "The representation and matching of pictorial structures". *IEEE Transactions on Computers*, C-22:67–92, 1973.

[42] N. S. Friedland and A. Rosenfeld. "Compact object recognition using energy-function based optimization". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:770–777, 1992.

[43] D. Geiger and F. Girosi. "Parallel and deterministic algorithms from MRF's: surface reconstruction". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):401–412, May 1991.

[44] D. Geman, S. Geman, C. Graffigne, and P. Dong. "Boundary detection by constrained optimization". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):609–628, July 1990.

[45] D. Geman and B. Gidas. *"Image analysis and computer vision"*, chapter 2, pages 9–36. National Academy Press, 1991.

[46] S. Geman and D. Geman. "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

[47] S. Geman and C. Graffigne. "Markov random field image models and their applications to computer vision". In A. M. Gleason, editor, *Proceedings of the International Congress of Mathematicians: Berkeley, August 3-11, 1986*, pages 1496–1517, 1987.

[48] S. Geman and D. McClure. "Bayesian image analysis: An application to single photon emission tomography". In *Proceedings of the Statistical Computing Section*, pages 12–18, Washington, DC, 1985.

[49] B. Gidas. "A renormalization group approach to image processing problems". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:164–180, 1989.

[50] U. Grenander. *Tutorials in Pattern synthesis.* Brown University, Division of Applied Mathematics, 1983.

[51] U. Grenander and D. M. K. Y. Chow. *Hands : a pattern theoretic study of biological shapes.* Springer-Verlag, New York, 1991.

[52] D. Griffeath. Introduction to random fields. In J. G. Kemeny, J. L. Snell, and A. W. Knapp, editors, *Denumerable Markov Chains*, chapter 12, pages 425–458. Springer-Verlag, New York, 2nd edition, 1976.

[53] W. E. L. Grimson. *From Images to Surfaces: A Computational Study of the Human Early Visual System.* MIT Press, Cambridge, MA, 1981.

[54] W. E. L. Grimson and T. Lozano-Prez. "Localizing overlapping parts by searching the interpretation tree". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):469–482, April 1987.

[55] J. M. Hammersley and P. Clifford. "Markov field on finite graphs and lattices". unpublished, 1971.

[56] F. R. Hansen and H. Elliott. "Image segmentation using simple Markov random field models". *Computer Graphics Image Processing*, 20:101–132, 1982.

[57] R. M. Haralick. "Decision making in context". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(4):417–428, July 1983.

[58] R. M. Haralick, H. Joo, C. Lee, X. Zhuang, V. Vaidya, and M. Kim. " Pose estimation from corresponding point data". *IEEE Transactions on Systems, Man and Cybernetics*, 19:1426–1446, 1989.

[59] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision.* Addison-Wesley, Reading, MA, 1992.

[60] J. G. Harris, C. Koch, E. Staats, and J. Lou. "Analog hardware for detecting discontinuities in early vision". *International Journal of Computer Vision*, 4:211–223, 1990.

[61] M. Hassner and J. Slansky. "The use of Markov random field as models of texture". *Computer Graphics Image Processing*, 12:357–370, 1980.

[62] L. Herault and R. Horaud. "Figure-ground discrimination: A combinatorial optimization approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:899–914, 1993.

[63] E. C. Hildreth. *The Measurement of Visual Motion.* MIT Press, Cambridge, MA, 1984.

[64] B. K. P. Horn and B. G. Schunck. "Determining optical flow". *Artificial Intelligence*, 17:185–203, 1981.

[65] P. V. C. Hough. "A method and means for recognizing complex patterns". U.S. Patent No. 3,069,654, 1962.

[66] R. Hu and M. M. Fahmy. "Texture segmentation based on a hierarchical Markov random field model". *Signal Processing*, 26:285–385, 1987.

[67] Y. P. Hung, D. B. Cooper, and B. Cernuschi-Frias. "Asymtotic Baysian surface estimation using an image sequence". *International Journal of Computer Vision*, 6(2):105–132, 1991.

[68] K. Ikeuchi and B. K. P. Horn. "Numerical shape from shading and occluding boundaries". *Artificial Intelligence*, 17:141–184, 1981.

[69] J. Illingworth and J. Kittler. "A survey of Hough transform". *Computer Vision, Graphics and Image Processing*, 43:221–238, 1988.

[70] A. K. Jain, Y. Zhong, and S. Lakshmanan. "Object matching using deformable templates". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(3):267–278, March 1996.

[71] E. Jaynes. "On the rationale of maximum-entropy methods". *Proceedings of the IEEE*, 70(9):939–952, 1982.

[72] K. Kanatani. *Geometric computation for machine vision.* Oxford University Press, New York, 1993.

[73] R. L. Kashyap, R. Chellappa, and A. Khotanzad. "Texture classification using features derived from random process models". *Pattern Recognition Letters*, 1:43–50, 1982.

[74] M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active contour models". In *Proceedings of IEEE International Conference on Computer Vision*, pages 259–268, 1987.

[75] K. Keeler. "Map representations and coding based priors for segmentation". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 420–425, 1991.

[76] I. Y. Kim and H. S. Yang. "Efficient image understanding based on the Markov random field model and error backpropagation network". In *Proceedings of International Conference Pattern Recognition*, volume A, pages 441–444, 1992.

[77] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. American Mathematical Society, Providence, R.I., 1980.

[78] C. Koch. "Computing motion in the presence of discontinuities: algorithm and analog networks". In R. Eckmiller and C. c. d. Malsburg, editors, *Neural Computers*, volume F41 of *NATO ASI Series*, pages 101–110. Springer-Verlag, 1988.

[79] J. J. Koenderink. *Solid Shape*. MIT Press, 1990.

[80] S. Lakshmanan and H. Derin. "Simultaneous parameter estimation and segmentation of gibbs random fields using simulated annealing". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:799–813, 1989.

[81] Y. Lamdan and H. Wolfson. "Geometric hashing: a general and efficient model-based recognition scheme". In *ICCV88*, pages 238–249, 1988.

[82] Y. G. Leclerc. "Constructing simple stable descriptions for image partitioning". *International Journal of Computer Vision*, 3:73–102, 1989.

[83] Y. G. Leclerc and M. A. Fischler. "An optimization-based approach to the interpretation of single line drawings as 3D wire frames". *International Journal of Computer Vision*, 9:113–136, 1992.

[84] D. Lee and T. Pavlidis. "One dimensional regularization with discontinuities". In *Proc. 1st International Conference on Computer Vision*, pages 572–577, London, England, 1987.

[85] S. Z. Li. "Invariant surface segmentation through energy minimization with discontinuities". *International Journal of Computer Vision*, 5(2):161–194, 1990.

[86] S. Z. Li. " Towards 3D vision from range images: An optimization framework and parallel networks". *CVGIP: Image Understanding*, 55(3):231–260, May 1992.

[87] S. Z. Li. "A Markov random field model for object matching under contextual constraints". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 866–869, Seattle, Washington, June 1994.

[88] S. Z. Li. *Markov Random Field Modeling in Computer Vision*. Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1995. See `http://markov.eee.ntu.edu.sg:8000/~szli/MRF_Book/MRF_Book.html`. Second edition to appear in 2000.

[89] S. Z. Li. "On discontinuity-adaptive smoothness priors in computer vision". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(6):576–586, June 1995.

[90] S. Z. Li. "Parameter estimation for optimal object recognition: Theory and application". *International Journal of Computer Vision*, 21(3):207–222, 1997.

[91] S. Z. Li. "Bayesian object matching". *Journal of Applied Statistics*, 25(3):425–443, 1998.

[92] S. Z. Li. "Close-form solution and parameter selection for convex minimization based edge-preserving smoothing". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(9):916–932, September 1998.

[93] S. Z. Li. "MAP image restoration and segmentation by constrained optimization". *IEEE Transactions on Image Processing*, 7(12):1730–1735, 1998.

[94] D. G. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer, 1985.

[95] K. V. Mardia, T. J. Hainsworth, and J. F. Haddon. "Deformable templates in image sequences". In *Proceedings of International Conference Pattern Recognition*, volume B, pages 132–135, 1992.

[96] K. V. Mardia and G. K. Kanji, editors. *Statistics and Images: 1 & 2.* Advances in Applied Statistics. Carfax, 1993,1994.

[97] K. V. Mardia, J. T. Kent, and A. N. Walder. "Statistical shape models in image analysis". In *Proceedings of 23rd Symposium Interface*, pages 550–575, 1991.

[98] J. L. Marroquin. "Probabilistic solution of inverse problems". *A. I. Lab. Tech. Report* No. 860, MIT, Cambridge, MA, 1985.

[99] J. L. Marroquin, S. Mitter, and T. Poggio. "Probabilistic solution of ill-posed problems in computational vision". *Journal of the American Statistical Association*, 82(397):76–89, March 1987.

[100] J. W. Modestino and J. Zhang. "A Markov random field model-based approach to image interpretation". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 458–465, 1989.

[101] R. Mohan and R. Nevatia. "Using perceptual organization to extract 3-d structures". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:1121–1139, 1989.

[102] J. Moussouris. "Gibbs and Markov systems with constraints". *Journal of statistical physics*, 10:11–33, 1974.

[103] D. Mumford and J. Shah. "Boundary detection by minimizing functionals: I". In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 22–26, San Francisco, CA, June 1985.

[104] J. L. Mundy and A. Zisserman, editors. *Geometric Invariants in Computer Vision.* MIT Press, Cambridge, MA, 1992.

[105] D. Murray and B. Buxton. "Scene segmentation from visual motion using global optimization". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:220–228, 1987.

[106] N. Nasrabadi, W. Li, and C. Y. Choo. "Object recognition by a Hopfield neural network". In *Proceedings of Third International Conference on Computer Vision*, pages 325–328, Osaka, Japan, December 1990.

[107] T. Pavlidis. "A critical survey of image analysis methods". In *ICPR*, pages 502–511, 1986.

[108] A. P. Pentland. "Automatic extraction of deformable part models". *International Journal of Computer Vision*, 4:107–126, 1990.

[109] T. Poggio, V. Torre, and C. Koch. "Computational vision and regularization theory". *Nature*, 317:314–319, 1985.

[110] T. Poggio, H. Voorhees, and A. Yuille. "Regularizing edge detection". *A. I. Lab. Memo* No. 773, MIT, Cambridge, MA, 1985.

[111] B. D. Ripley. *Spatial Statistics.* Wiley, New York, 1981.

[112] J. Rissanen. "Modeling by shortest data description". *Automatica*, 14:465–471, 1978.

[113] J. Rissanen. "A universal prior for integers and estimation by minimal discription length". *Annals of Statistics*, 11(2):416–431, 1983.

[114] L. G. Roberts. "Machine perception of three-dimensional solids". In e. a. J. T. Tippett, editor, *Optical and Electro-Optical Information Processing.* MIT Press, Cambridge, MA, 1965.

[115] A. Rosenfeld, R. Hummel, and S. Zucker. "Scene labeling by relaxation operations". *IEEE Transactions on Systems, Man and Cybernetics*, 6:420–433, June 1976.

[116] A. Rosenfeld and A. C. Kak. *Digital Image Processing.* Academic Press, New York, 1976.

[117] L. G. Shapiro and R. M. Haralick. "Structural description and inexact matching". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:504–519, September 1981.

[118] D. Shulman and J. Herve. "Regularization of discontinuous flow fields". In *Proc. Workshop on Visual Motion*, pages 81–86, 1989.

[119] J. F. Silverman and D. B. Cooper. "Bayesian clustering for unsupervised estimation of surface and texture models". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:482–495, 1988.

[120] T. Simchony and R. Chellappa. "Stochastic and deterministic algorithms for MAP texture segmentation". In *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, pages 1120–1123, 1988.

[121] M. A. Snyder. "On the mathematical foundations of smoothness constraints for the determination of optical flow and for surface reconstruction". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:1105–1114, 1991.

[122] G. C. Stockman and A. K. Agrawala. "Equivalence of Hough curve detection to template matching". *Communications of the ACM*, 20:820–822, 1977.

[123] G. Storvik. "A Bayesian approach to dynamic contours through stochastic sampling and simulated annealing". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(10):976–986, October 94.

[124] D. J. Strauss. "Clustering on colored lattice". *Journal of Applied Probability*, 14:135–143, 1977.

[125] R. Szeliski. *Bayesian modeling of uncertainty in low-level vision.* Kluwer, 1989.

[126] H. L. Tan, S. B. Gelfand, and E. Delp. "A cost minimization approach to edge detection using simulated annealing". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:3–18, 1992.

[127] D. T. Terzopoulos. "Multilevel computational process for visual surface reconstruction". *Computer Vision, Graphics and Image Processing*, 24:52–96, 1983.

[128] D. T. Terzopoulos. "Regularization of inverse visual problems involving discontinuities". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(4):413–424, July 1986.

[129] C. W. Therrien. *Decision, estimation, and classification : an introduction to pattern recognition and related topics.* Wiley, New York, 1989.

[130] A. N. Tikhonov and V. A. Arsenin. *Solutions of Ill-posed Problems.* Winston & Sons, Washington, 1977.

[131] V. Torre and T. Poggio. "On edge detection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(2):147–163, March 1986.

[132] S. Ullman. *The Interpolation of Visual Motion.* MIT Press, Cambridge, MA, 1979.

[133] D. F. Watson. "Computing the n-dimentional Delaunay tessellation with application to Voronoi polytopes". *Computer Journal*, 24:167–172, 1981.

[134] W. M. Wells III. "MAP model matching". In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 486–492, 1991.

[135] C. S. Won and H. Derin. "Unsupervised segmentation of noisy and textured images y using Markov random fields". *CVGIP: Graphics Model and Image Processing*, 54:308–328, 1992.

[136] J. W. Woods. "Two-dimensional discrete Markovian fields". *IEEE Transactions on Information Theory*, 18:232–240, 1972.

[137] Z. Wu and R. Leahy. "An approximation method of evaluating the joint likelihood for first-order GMRFs". *IEEE Transactions on Image Processing*, 2(4):520–523, 1993.