Chapter 4

High-Level MRF Models

High-level vision tasks, such as object matching and recognition and pose estimation, are performed on features extracted from images. The arrangements of such features are usually irregular, and hence the problems fall into categories LP3 and LP4. In this chapter, we present MAP-MRF formulations for solving these problems.

We begin with a study on the problem of object matching and recognition under contextual constraints. An MAP-MRF model is then formulated following the systematic approach summarized in Section 1.3.4. The labeling of a scene in terms of a model[1] object is considered as an MRF. The optimal labeling of the MRF is obtained by using the MAP principle. The matching of different types of features and multiple objects is discussed. A related issue, MRF parameter estimation for object matching and recognition, will be studied in Chapter 7.

We then derive two MRF models for pose computation, pose meaning the geometric transformation from one coordinate system to another. In visual matching, the transformation is from the scene (image) to the model object considered (or vice versa). In derived models, the transformation is from a set of object features to a set of image features. They minimize posterior energies derived for the MAP pose estimation, possibly together with an MRF for matching.

4.1 Matching under Relational Constraints

In high-level image analysis, we are dealing with image features, such as critical points, lines, and surface patches, that are more abstract than image pixels. Such features in a scene are not only attributed by (unary) properties about the features themselves but also related to each other by relations

[1] In this chapter, the word "model" is used to refer to both mathematical vision models and object models.

S.Z. Li, *Markov Random Field Modeling in Image Analysis*,
Advances in Pattern Recognition, DOI: 10.1007/978-1-84800-279-1_4,

between them. In other words, an object or a scene is represented by features constrained by the properties and relations. It is the bilateral or higher-order relations that convey the contextual constraints. They play a crucial role in visual pattern matching.

4.1.1 Relational Structure Representation

The features, properties and relations can be denoted compactly as a *relational structure* (RS) (Fischler and Elschlager 1973; Ambler et al. 1973; Cheng and Huang 1984; Radig 1984; Li 1992c; Li 1992a). An RS describes a scene or (part of) a model object. The problem of object recognition is reduced to that of RS matching.

Let us start with a scene RS. Assume there are m features in the scene. These features are indexed by a set $\mathcal{S} = \{1, \ldots, m\}$ of sites. The sites constitute the nodes of the RS. Each node $i \in \mathcal{S}$ has associated with it a vector $d_1(i)$ composed of a number of K_1 *unary properties* or *unary relations*, $d_1(i) = [d_1^{(1)}(i), \ldots, d_1^{(K_1)}(i)]^T$. A unary property could be, for example, the color of a region, the size of an area, or the length of a line. Each pair of nodes $(i, i' \in \mathcal{S}, i' \neq i)$ are related to each other by a vector $d_2(i, i')$ composed of a number of K_2 *binary (bilateral) relations*, $d_2(i, i') = [d_2^{(1)}(i, i'), \ldots, d_2^{(K_2)}(i, i')]^T$. A binary relation could be, for example, the distance between two points or the angle between two lines. More generally, among n features $i_1, \ldots, i_n \in \mathcal{S}$, there may be a vector $d_n(i_1, \ldots, i_n)$ of K_n n-ary relations. This is illustrated in Fig. 4.1. An n-ary relation is also called a relation, or constraint, of order n. The scope of relational dependencies can be determined by a neighborhood system $\mathcal{N}$ on $\mathcal{S}$. Now, the RS for the scene is defined by a triple

$$\mathcal{G} = (\mathcal{S}, \mathcal{N}, d) \tag{4.1}$$

where $d = \{d_1, d_2, \ldots, d_H\}$ and H is the highest-order. For $H = 2$, the RS is also called a *relational graph (RG)*. The highest-order H cannot be lower than 2 when contextual constraints must be considered.

The RS for a model object is similarly defined as

$$\mathcal{G}' = (\mathcal{L}, \mathcal{N}', D) \tag{4.2}$$

where $D = \{D_1, D_2, \ldots, D_H\}$, $D_1(I) = [D_1^{(1)}(I), \ldots, D_1^{(K_1)}(I)]^T$, $D_2(I, I') = [D_2^{(1)}(I, I'), \ldots, D_2^{(K_2)}(I, I')]^T$, and so on. In this case, the set of labels $\mathcal{L}$ replaces the set of sites. Each element in $\mathcal{L}$ indexes one of the M model features. In addition, the "neighborhood system" for $\mathcal{L}$ is defined to consist of all the other elements, that is,

$$\mathcal{N}'_I = \{I' \mid \forall I' \in \mathcal{L}, I' \neq I\} \tag{4.3}$$

This means each model feature is related to all the other model features. The highest-order considered, H, in $\mathcal{G}'$ is equal to that in $\mathcal{G}$. For particular n and

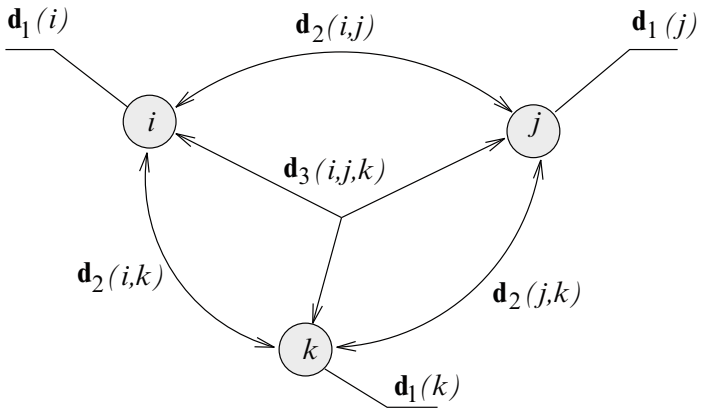

Figure 4.1: Three nodes and their unary, binary, and triple relations in the RS representation. From (Li 1992a) with permission; ©1992 Elsevier.

k ($1 \le k \le K_n$; $1 \le n \le H$), $D_n^{(k)}$ represents the same type of constraint as $d_n^{(k)}$; for example, both represent the angle between two line segments.

Relations of various orders impose unary, binary, ... , H-ary constraints on the features. Intercontextual constraints are represented by the second- or higher-order relations. Due to these constraints, a scene, an object, or a view of an object is seen as an integrated part rather than as individual features. The higher the order of relations is, the more powerful the constraints are but the higher the complication and expenses are in the computation.

There can be multiple model RSs in a model base. A model RS describes a whole model object or a part of it. When an RS is used to describe a part (for example, a view) of an object, the whole object may be described by several RSs and these RSs may be related by some inter-RS constraints.

Now introduce a virtual model composed of a single node $\mathcal{L}_0 = \{0\}$. It is called the NULL model. This special model represents everything not modeled by $\mathcal{G}'$, such as features due to all the other model objects and the noise. So the actual label set in matching the scene to the model plus the NULL contains $M + 1$ labels. It is denoted by

$$\mathcal{L}^+ = \{0, 1, \ldots, M\} \tag{4.4}$$

After the introduction of the NULL, the mapping from $\mathcal{S}$ to $\mathcal{L}$ is illustrated in Fig. 4.2.

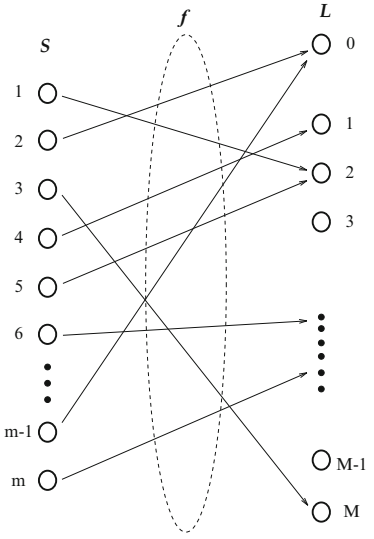

Figure 4.2: Discrete mapping involving the NULL label (numbered 0). All scene nodes (sites) not modeled by the object considered should be matched to this special label.

Figures 4.3 and 4.4 demonstrate two cup images and their segmentations based on the H-K surface curvatures (Besl and Jain 1985). Suppose we are matching the RG in Fig. 4.4(b) to that in Fig. 4.4(a). Based on the constraints from the unary properties of surface curvature and region area and the binary relations of distance between regions, the correct matching from Fig. 4.4(b) to Fig. 4.4(a) is $5 \rightarrow 1$, $1 \rightarrow 2$, and the rest to NULL.

Model-based matching can be considered as finding the optimal mapping from the image RS to the model RS (or vice versa). Such a mapping from one RS to another is called a *morphism*, written as

$$f : \mathcal{G}(\mathcal{S}, \mathcal{N}, d) \rightarrow \mathcal{G}'(\mathcal{L}, \mathcal{N}', D) \tag{4.5}$$

which maps each node in $\mathcal{S}$ to a node in $\mathcal{L}$

$$f : \mathcal{S} \rightarrow \mathcal{L} \tag{4.6}$$

and thus maps relations d_n to relations D_n

$$f : d_n \rightarrow D_n \tag{4.7}$$

A morphism is called an *isomorphism* if it is one-to-one and onto. It is called a *monomorphism* if it is one-to-one but not onto. It is called a *homomorphism* if it is many-to-one. We do not allow one-to-many mappings because

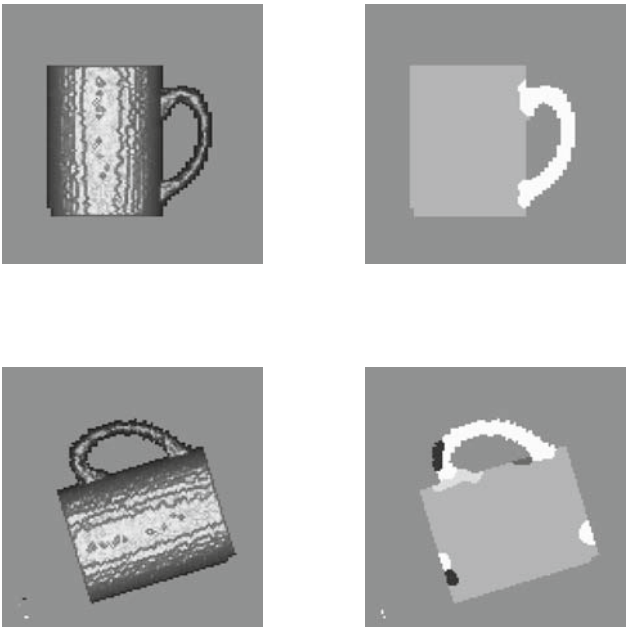

Figure 4.3: Cup images and segmentation based on H-K surface curvatures. Top: A cup image and its H-K map. Bottom: A transformed version of the cup image and the H-K map (note that some noise is introduced after the transformation due to quantization). From (Li 1992c) with permission; ©1992 Elsevier.

they contradict the definition of functions and, more crucially, increase the difficulties in finding the optimal solutions.

When the properties and relations in the RSs considered include numerical values, the mappings are numerical morphisms, which are more difficult to resolve than symbolic morphisms. Since such morphisms do not preserve relations in the exact, symbolic sense, they may be called *weak morphisms* – a term extended from the weak constraint models (Hinton 1978; Blake 1983; Blake and Zisserman 1987).

The goodness of a numerical morphism is usually judged by an objective function such as an energy. It is not very difficult to find a correct one-to-one mapping (isomorphism) between two *identical* RSs. The requirement that the unary properties of two nodes and the binary relations of two links must be exactly the same in order to be matched to each other provides a strong constraint for resolving the ambiguities. For the case where the two RSs have different numbers of nodes and the matched properties and relations are not

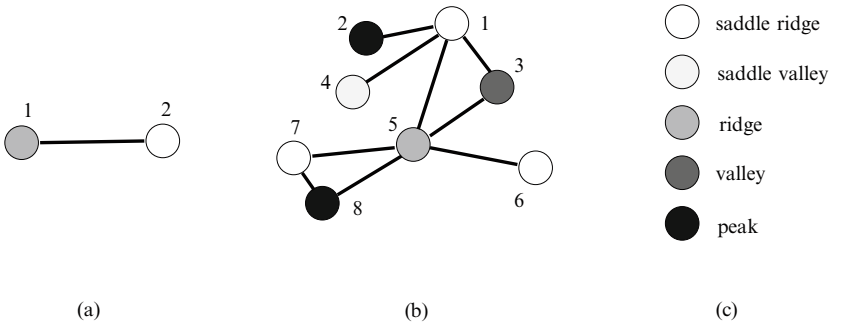

Figure 4.4: Relational graphs built from the cup H-K segmentation maps. The textures of the nodes denote different surface types, and the links represent the adjacency relations. (a) The RG for the original upright cup image. Node 1 corresponds to the body of the cup and node 2 to the handle. (b) The RG for the transformed cup. Node 5 corresponds to the body, node 1 to the handle, and the rest to NULL. (c) Legend for the correspondences between texture types and H-K surface types. From (Li 1992c) with permission; ©1992 Elsevier.

exactly the same, the matching is more difficult because it cannot exploit the advantage of the strong constraint of the exact equalities. Only "weak constraints" are available. The subsequent sections use MRF's to establish the objective function for weak homomorphisms for inexact, partial matching.

4.1.2 Work in Relational Matching

Two important computational issues in object matching are how to use contextual constraints and how to deal with uncertainties. Contextual constraints in object recognition are often represented using the notion of relational graphs. An object is represented by a set of features, their properties, and relations. Relative values between two image features embed context in matching (Fischler and Elschlager 1973). Object matching is then reduced to matching between two relational graphs. On the other hand, noise is inevitably introduced in the process of feature extraction and relation measurement.

In the maximum clique method (Ambler et al. 1973; Ghahraman et al. 1980) for relational graph matching, an associated graph is formed from the two relational graphs, one for the scene and the other for a model object

(Ambler et al. 1973). Each node of the associated graph represents a possible match. The optimal matching is given by the maximal cliques.

A class of constraint satisfaction problems for matching was studied by Shapiro and Haralick (1981). In their work, the criterion is the weighted number of mismatched relations. Most such work involves definiting some criteria to describe the "goodness", or conversely the cost, of matching with relational constraints. Mistakes in symbolic relations are mapped into a number, and this number is then used as a criterion to decide whether a homomorphism is acceptable or not. The inexact matching problem is viewed as finding the best homomorphism (i.e., the one with minimum number of errors w.r.t. a given attribute value threshold, a missing part threshold, and a relation threshold).

Relaxation labeling (RL) (Rosenfeld et al. 1976) has been a useful method for solving the matching problem. The constraints are propagated via a compatibility function, and the ambiguity of labeling is reduced by using an iterative RL algorithm. In our view, the most important part in RL-based recognition is the definition of the compatibility function. Various RL schemes should be considered as algorithms for finding solutions. This will be further examined in Section 9.3.2.

Typical early works on relational matching using RL include (Davis 1979) and (Bhanu and Faugeras 1984; Bhanu 1984). In (Davis 1979), the objective function consists of four terms. Each term either encodes a particular constraint or penalizes unmatched features, the idea dating back to work by Fischler and Elschlager (1973). An association graph (Ambler et al. 1973) is used for matching relational structures. In the search for optimal matching, incompatible nodes for which some evaluation function is below a threshold are deleted from the graph. This generates a sequence of association graphs until a fixed point is reached. In (Bhanu and Faugeras 1984; Bhanu 1984), matching is posed as an optimization problem, and the optimization is performed by using an RL algorithm presented in (Faugeras and Berthod 1981).

A feature common to most of the matching works above is that thresholds are used to determine whether two matches are compatible. This effectively converts the weighted-graph matching into symbolic matching. While greatly reducing search space, this may rule out the potential matches. Because of the noise, the observation of the objects, which represents feature properties and relations extracted from the scene, can be considered as a set of random variables. Furthermore, some object features may be missing, and spurious features may emerge due to noise and unmodeled objects. The matching strategy has to deal with these uncertainties. It is hard to judge that in the presence of uncertainties a difference of 1.000001 is impossible while 0.999999 is possible.

In the weak constraint satisfaction paradigm, "hard" constraints are allowed to be violated without causing the failure of constraint satisfaction. However, each such violation is penalized by adding an amount to a cost function that measures the imperfection of the matching. This is usually

implemented by using the line process in low-Level vision (Geman and Geman 1984; Marroquin 1985). In a weak notion of graph matching, Bienenstock (1988) proposed a scheme for an approximation of graph isomorphism in which relation-preserving characteristics of isomorphism can be violated but each violation incurs a small penalty. This is a transplant of the idea of the line process at the lower-level to the higher-level perception. Nevertheless, at a higher-level where more abstract representations are used, the weak constraint satisfaction problem becomes more complicated.

Li makes use of contextual constraints not only on the prior configuration of labelings but also on the observed data into the labeling process (Li 1992c; Li 1992a; Li 1992b; Li et al. 1993). He proposes an energy function, on a heuristic basis, that combines contextual constraints from both the prior knowledge and the observation. Kittler et al. (1993) later derive from probabilistic viewpoint the same compatibility used in Li's energy function.

MRF's provide a formal basis for matching and recognition under contextual constraints. Modestino and Zhang (1989) describe an MRF model for image interpretation. They consider an interpretation of a scene as an MRF and define the optimal matching as the MAP estimate of the MRF. Unfortunately, the posterior probability therein is derived not by using the laws of probability but designed directly by using some heuristic rules. This contradicts the promises of MAP-MRF modeling. Cooper (1990) describes a coupled network for simultaneous object recognition and segmentation. MRF is used to encode prior qualitative and possibly quantitative knowledge in the nonhomogeneous and anisotropic situations. The network is applied to recognize Tinkertoy objects. An interesting development are Markov processes of objects proposed by Baddeley and van Lieshout (1993). Other works in MRF-based recognition can be found in (Grenander et al. 1991; Baddeley and van Lieshout 1992; Friedland and Rosenfeld 1992; Kim and Yang 1992; Cooper et al. 1993). The MRF model described in the next section is based on (Li 1994a).

4.2 Feature-Based Matching

The labeling of a scene in terms of a model object is denoted by $f = \{f_i \in \mathcal{L}^+ \mid i \in \mathcal{S}\}$, where elements in $\mathcal{S}$ index image features and those in $\mathcal{L}$ model object features plus the NULL. It is also interpreted as a relational mapping from $\mathcal{G}(\mathcal{S}, \mathcal{N}, d)$ to $\mathcal{G}'(\mathcal{L}, \mathcal{N}', D)$; see (4.5). We assume that f is a realization of an MRF w.r.t. $\mathcal{N}$. Below, we derive its posterior probability using the MAP-MRF approach, in which contextual constraints not only on the prior configuration but also the observation are considered.

4.2.1 Posterior Probability and Energy

The prior distribution of f is defined using MRF's. In RS matching, the neighborhood covers all other related sites (scene features). One may restrict the scope of interaction by defining $\mathcal{N}_i$ as the set of other features that are within a distance r from i (see (2.3))

$$\mathcal{N}_i = \{i' \in \mathcal{S} \mid [\text{dist}(\text{feature}_{i'}, \text{feature}_i)]^2 \leq r,\ i' \neq i\} \tag{4.8}$$

where the function dist is a suitably defined distance function for features. The distance threshold r may be reasonably related to the size of the model object. The set of first-order cliques is

$$\mathcal{C}_1 = \{\{i\} \mid i \in \mathcal{S}\} \tag{4.9}$$

The set of second-order cliques is

$$\mathcal{C}_2 = \{\{i, i'\} \mid i' \in \mathcal{N}_i, i \in \mathcal{S}\} \tag{4.10}$$

Here, only cliques of up to order two are considered.

The single-site potential is defined as

$$V_1(f_i) = \begin{cases} v_{10} & \text{if } f_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

where v_{10} is a constant. If f_i is the NULL label, it incurs a penalty of v_{10}; otherwise the nil penalty is imposed . The pair-site potential is defined as

$$V_2(f_i, f_{i'}) = \begin{cases} v_{20} & \text{if } f_i = 0 \text{ or } \ f_{i'} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

where v_{20} is a constant. If either f_i or $f_{i'}$ is the NULL, it incurs a penalty of v_{20} or the nil penalty otherwise. The above clique potentials define the prior energy $U(f)$. The prior energy is then

$$U(f) = \sum_{i \in \mathcal{S}} V_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \tag{4.13}$$

The definitions of the above prior potentials are a generalization of that penalizing line process variables (Geman and Geman 1984; Marroquin 1985). The potentials may also be defined in terms of stochastic geometry (Baddeley and van Lieshout 1992).

The conditional p.d.f., $p(d \mid f)$, of the observed data d, also called the *likelihood function* when viewed as a function of f given d fixed, has the following characteristics:

1. It is conditioned on pure nonNULL matches $f_i \neq 0$.
2. It is independent of the neighborhood system $\mathcal{N}$.

3. It depends on how the model object is observed in the scene, which in turn depends on the underlying transformations and noise.

Assume (1) that D and d are composed of types of features which are invariant under the class of transformations considered[2]; and (2) that they are related via the observation model

$$d_1(i) = D_1(f_i) + e_1(i), \qquad d_1(i,i') = D_2(f_i, f_{i'}) + e_2(i,i') \tag{4.14}$$

where e is additive, independent, zero-mean Gaussian noise. The assumptions of the independent and Gaussian noise may not be accurate but offer an approximation when an accurate observation model is not available.

Then the likelihood function is a Gibbs distribution with the energy

$$U(d \mid f) = \sum_{i \in \mathcal{S}, f_i \neq 0} V_1(d_1(i) \mid f_i) + \sum_{i \in \mathcal{S}, f_i \neq 0} \sum_{i' \in \mathcal{S} - \{i\}, f_{i'} \neq 0} V_2(d_2(i,i') \mid f_i, f_{i'}) \tag{4.15}$$

where the constraints, $f_i \neq 0$ and $f_{i'} \neq 0$, restrict the summations to take over the nonNULL matches. The likelihood potentials are

$$V_1(d_1(i) \mid f_i) = \begin{cases} \sum_{k=1}^{K_1} [d_1^{(k)}(i) - D_1^{(k)}(f_i)]^2 / \{2[\sigma_1^{(k)}]^2\} & \text{if } f_i \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

and

$$V_2(d_2(i,i') \mid f_i, f_{i'}) = \begin{cases} \sum_{k=1}^{K_2} [d_2^{(k)}(i,i') - D_2^{(k)}(f_i, f_{i'})]^2 / \{2[\sigma_2^{(k)}]^2\} \\ \qquad \text{if } i' \neq i \text{ and } f_i \neq 0 \text{ and } f_{i'} \neq 0 \\ 0 \qquad \text{otherwise} \end{cases} \tag{4.17}$$

where $[\sigma_n^{(k)}]^2$ ($k = 1, \ldots, K_n$ and $n = 1, 2$) are the variances of the corresponding noise components. The vectors $D_1(f_i)$ and $D_2(f_i, f_{i'})$ are the "mean vectors", conditioned on f_i and $f_{i'}$, for the random vectors $d_1(i)$ and $d_2(i,i')$, respectively.

Using $U(f \mid d) = U(f) + U(d \mid f)$, we obtain the posterior energy

$$\begin{aligned} U(f \mid d) = \; & \textstyle\sum_{i \in \mathcal{S}} V_1(f_i) + \\ & \textstyle\sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) + \\ & \textstyle\sum_{i \in \mathcal{S}: f_i \neq 0} V_1(d_1(i) \mid f_i) + \\ & \textstyle\sum_{i \in \mathcal{S}: f_i \neq 0} \sum_{i' \in \mathcal{S} - \{i\}: f_{i'} \neq 0} V_2(d_2(i,i') \mid f_i, f_{i'}) \end{aligned} \tag{4.18}$$

There are several parameters involved in the posterior energy: the noise variances $[\sigma_n^{(k)}]^2$ and the prior penalties v_{n0}. Only the relative, not absolute, values of $[\sigma_n^{(k)}]^2$ and v_{n0} are important because the solution f^* remains the

[2]The discovery and computation of visual invariants is an active area of research; see (Mundy and Zisserman 1992).

same after the energy E is multiplied by a factor. The v_{n0} in the MRF prior potential functions can be specified to achieve the desired system behavior. The higher the prior penalties v_{n0}, the fewer features in the scene will be matched to the NULL for the minimal energy solution.

Normally, symbolic relations are represented internally by a number. The variances $[\sigma_n^{(k)}]^2$ for those relations are zero. One may set corresponding $[\sigma_n^{(k)}]^2$ to 0^+ (a very small positive number), which is consistent with the concept of discrete distributions. Setting $[\sigma_n^{(k)}]^2 = 0^+$ causes the corresponding distance to be infinitely large when the symbolic relations compared are not the same. This inhibits symbolically incompatible matches, if an optimal solution is sought and thus imposes the desired symbolic constraint. A method for learning $[\sigma_n^{(k)}]^2$ parameters from examples will be presented in Chapter 8.

4.2.2 Matching to Multiple Objects

The MAP configuration f^* derived in the above is the optimal mapping from the scene to the model object under consideration. In other words, it is the optimal labeling of the scene in terms of the model object.

Suppose there are L potential model objects. Then L MAP solutions, $f^{(1)}, \ldots, f^{(L)}$, can be obtained after matching the scene to each of the models in turn.[3] However, any feature in the scene can have only one match of model feature. To resolve this, we use the following method of cost minimization.

Rewrite $E(f) = U(f \mid d)$ in (4.18) in the form

$$E(f) = \sum_{i \in \mathcal{S}} E_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{S}, i' \neq i} E_2(f_i, f_{i'}) \triangleq \sum_{i \in \mathcal{S}} E(f_i \mid f_{\mathcal{N}_i}) \tag{4.19}$$

where

$$E_1(f_i) = \begin{cases} v_{10} & \text{if } f_i = 0 \\ V_1(d_1(i) \mid f_i) & \text{otherwise} \end{cases} \tag{4.20}$$

and

$$E_2(f_i, f_{i'}) = \begin{cases} v_{20} & \text{if } i' \in \mathcal{N}_i \ \&\ (f_i = 0 \text{ or } \ f_{i'} = 0) \\ V_2(d_2(i, i') \mid f_i, f_{i'}) & \text{if } i' \neq i \text{ and } f_i \neq 0 \text{ and } f_{i'} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.21}$$

are local posterior energies of orders one and two, respectively. $E(f_i \mid f_{\mathcal{N}_i}) = E_1(f_i) + \sum_{i' \in \mathcal{S}} E_2(f_i, f_{i'})$ is the cost incurred by the local match $i \rightarrow f_i$ given the rest of the matches. It will be used as the basis for selecting the best-matched objects for i in matching to multiple model objects. The image

[3] Fast indexing of model objects (Lamdan and Wolfson 1988) is a topic not studied in this work.

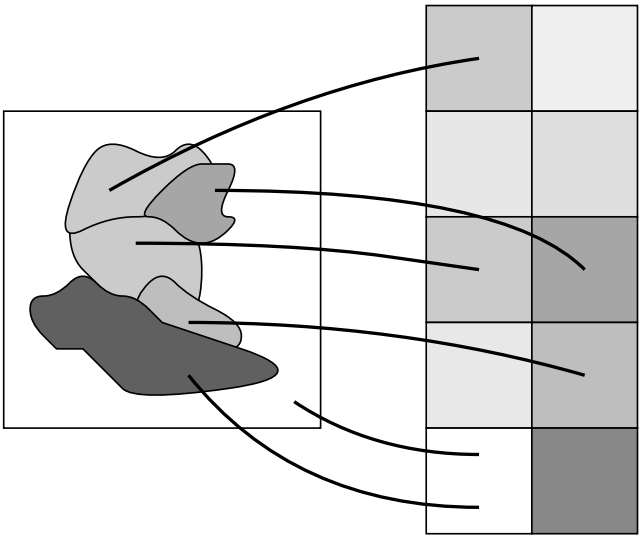

Figure 4.5: Mapping from the scene to multiple model objects. Different textures represent different structures. Bold lines represent submappings. Note that the background and nonmodel structure are mapped to the NULL structure (the blank square). From (Li 1992b) with permission; ©1992 Elsevier.

feature i is considered to come from object ℓ_i if the decision incurs the least cost

$$\ell_i = \arg \min_{\ell \in \{1,\ldots,L\}} E(f_i^{(\ell)} \mid f_{\mathcal{N}_i}^{(\ell)}) \tag{4.22}$$

The final label for i is feature number $f_i^{(\ell_i)}$ of object ℓ_i. Note, however, that the MAP principle is applied to matching to a single model object, not to multiple objects; the simple rule (4.22) does not maximize the posterior since at least the partition functions are different for matching to different objects.

Applying (4.22) to every i yields an overall mapping from the scene to the models, composed of several submappings, as illustrated in Fig. 4.5. On the right, the squares with different textures represent the candidate model structures to which the scene is to be matched. Among them, the blank square represents the NULL model. On the left is the scene structure. The regions, each corresponding to a subpart of a model structure, are overlapping (not separated). The recognition of the overlapping scene is (1) to partition the scene structure into parts such that each part is due to a single object and (2) to find correspondences between features in each part and those in the corresponding model object. The parts of the scene corresponding to the

background and unmodeled objects should be mapped to the NULL, or in other words assigned the NULL label.

4.2.3 Extensions

The model above may be extended in a number of ways. With the assumption that different constraints are independent of each other, embedding a higher constraint can be achieved by adding a new energy term. Matching with different types of object features (e.g., points and lines) can be treated as coupled MRF's.

Incorporating Higher Constraints

In matching schemes based on invariants, the features chosen to represent the object modeled and the scene must be invariant to the expected transformations from the object to the observation for the process of cognition to be accomplished. The more complicated the transformations are, the higher the order of features needed for the invariant object representation (Li 1992a); the order needed may be higher than two.

In previous subsections, only constraints of up to second-order were considered. Incorporation of higher-order constraints can be achieved by adding higher-order energy terms. A clique of order $n > 2$ is an n-tuple $\{i_1, \ldots, i_n\}$ in which i_r and i_s $(r \neq s)$ are neighbors to each other. The incorporation is done as follows. First, the following n-th order a priori clique potentials are added to the prior energy $U(f)$:

$$V_{n0}(f_{i_1}, \ldots, f_{i_n}) = \begin{cases} v_{n0} & \text{if } f_{i_k} = 0 \quad \exists i_k \in \{i_1, \ldots, i_n\} \\ 0 & \text{otherwise} \end{cases} \tag{4.23}$$

where v_{n0} is a constant of prior penalty. Second, the likelihood energy for the nth order observation has the likelihood potentials

$$V_n(d_n(i_1, \ldots, i_n) \mid f_{i_1}, \ldots, f_{i_n}) = \sum_{k=1}^{K_n} [d_n^{(k)}(i_1, \ldots, i_n) - D_n^{(k)}(f_{i_1}, \ldots, f_{i_n})]^2 / \{2[\sigma_n^{(k)}]^2\} \tag{4.24}$$

The corresponding posterior can be obtained using the Bayes rule, resulting in the nth order energy

$$E_n(f_{i_1}, \ldots, f_{i_n}) = \begin{cases} v_{n0}, & \text{if } f_{i_k} = 0 \ \exists i_k \in \{i_1, \ldots, i_n\} \\ V_n(d_n(i_1, \ldots, i_n) \mid f_{i_1}, \ldots, f_{i_n}), & \text{otherwise} \end{cases} \tag{4.25}$$

Adding together all the energy terms yields

$$E(f) = \sum_{i \in \mathcal{S}} E_1(f_i) + \sum_{i,i' \in \mathcal{S}} E_2(f_i, f_{i'}) + \cdots + \sum_{i_1, \ldots, i_H \in \mathcal{S}} E_H(f_{i_1}, \ldots, f_{i_H}) \tag{4.26}$$

where H is the highest-order.

Coupled MRF's for Matching with Different Features

Let us consider the situation where an object consists of different types of features, such as points and lines. Obviously, a point in the scene should not be matched to a line in an object model. This is a symbolic constraint. In this case, the positivity condition of MRF in (2.8) does not hold any more if the configuration space $\mathbb{F}$ is still defined as the simple product as in (4.4) for a single MRF.

To overcome this limitation, we partition the whole set $\mathcal{L}$ of labels to a few admissible sets for different types of sites. This results in a few coupled MRF's. These MRF's are coupled to each other via inter-relations d_n ($n \geq 2$). For example, the distance between a point and a line can constrain the two different types of features. Furthermore, they are also coupled via the label NULL which is a "wildcard" compatible with all types of features.

If there are two different types of features, then $\mathcal{L}$ can be partitioned into two admissible sets, with each set consisting of indices to features of the same type. In the most general case, each of the m sites has its own set of labels $\mathcal{L}_i \subseteq \mathcal{L}$ ($i = 1, \ldots, m$), each $\mathcal{L}_i$ being determined using the symbolic unary constraints; and the label for site i assumes a value $f_i \in \mathcal{L}_i^+$, where $\mathcal{L}_i^+ = \{0\} \cup \mathcal{L}_i$. Then, the configuration space is defined as

$$\mathbb{F} = \mathcal{L}_1^+ \times \mathcal{L}_2^+ \times \cdots \times \mathcal{L}_m^+ \tag{4.27}$$

In this situation, the energy $E(f)$ has the same form as usual and the solution is still found by $f^* = \arg\min_{f \in \mathbb{F}} E(f)$. The only difference is in the definition of the configuration space $\mathbb{F}$ in which the solution is searched for.

Relationships with Low Level MRF Models

Let us compare the present model with low-Level vision MRF models prototyped by Geman and Geman (1984). The present model is similar to the MRF models for piecewise constant image restoration, edge detection, and texture segmentation in that the labels are discrete. Of course, their prior distributions must be different to cope with different tasks.

In surface reconstruction involving discontinuities (Marroquin 1985; Blake and Zisserman 1987; Chou and Brown 1990; Szeliski 1989; Geiger and Girosi 1991), there are commonly two coupled MRF's: a surface field and a line process field. The former field is defined on $\mathcal{S}_1$, the domain of an image grid. It assumes configurations in the space $\mathcal{L}_1^{\mathcal{S}_1}$ where $\mathcal{L}_1$ is a real interval. The latter is defined on $\mathcal{S}_2$, the dual of $\mathcal{S}_1$. It assumes configurations in the space $\mathcal{L}_2^{\mathcal{S}_2}$, where $\mathcal{L}_2$ is the set of labels such as {edge, nonedge}. These fields are coupled to each other by the interaction between the line process variable and the neighboring pixels.

The concept of discontinuity in the high-level is the relational bond in the scene RS. For example, when no f_i or $f_{i'}$ assumes the NULL value, i and i' are relationally constrained; otherwise, when $f_i = 0$ or $f_{i'} = 0$, the relational bond between i and i' is broken. This corresponds to the line process.

The main difference between this high-level model and those low-Level models is in the encoding of higher-order relational constraints. Low-level models use unary observations only, such as pixel intensity; although intensity difference between neighboring pixels is also used, it is derived directly from the intensity. The present model uses relational measurements of any order. This is important for high-level problems in which contextual constraints play a more important role. Moreover, in the present model, the neighborhood system is nonhomogeneous and anisotropic, which also differs from the image case.

The matching method we have presented is based on a prerequisite that invariants are available for object representation under the group of transformations concerned. If geometric variants are also used as sources of constraints, object poses have to be resolved during the computation of matching; see the next section.

4.2.4 Experiments

The following presents some experiments. Given a number of model objects, a scene is generated. Properties and relations in model objects and the scene are measured using the same program. Only the second-order energy E_2 is taken into consideration. The energy is minimized by using a relaxation labeling algorithm (see Section 9.3.2). In the computation of the minimal solution, interactions or compatibilities are represented by integers of only 8 bits and good results are achieved; this demonstrates the error-tolerant aspect of the model. The optimal matching result is displayed by aligning the matched object features to the scene while the unmatched are not displayed. The alignment is performed between the matched pairs by using the least squares fitting method (Umeyama 1991). The parameter $v_{20} = 0.7$ is fixed for all the experiments. Parameters $[\sigma_2^{(k)}]^2$ vary for different applications.

Matching Objects of Point Patterns

There are three model objects as shown in Fig. 4.6(a)–Fig. 4.6(c). Each of the objects consists of three types of point features, shown in different sizes. The scene in (d) is generated from the model objects as follows: (1) Take a subset of features from each of the three objects, (2) do a transformation (rotation and translation) on each of the subsets, (3) mix the subsets together after that, (4) randomly deviate the locations of the points using either Gaussian or uniform noise, and (5) add spurious point features.

In this case of point matching, there is only one unary property (i.e., the point size) denoted $d_1(i)$. Each $d_1(i)$ takes a value $\{1, 2, 3\}$. There is only a

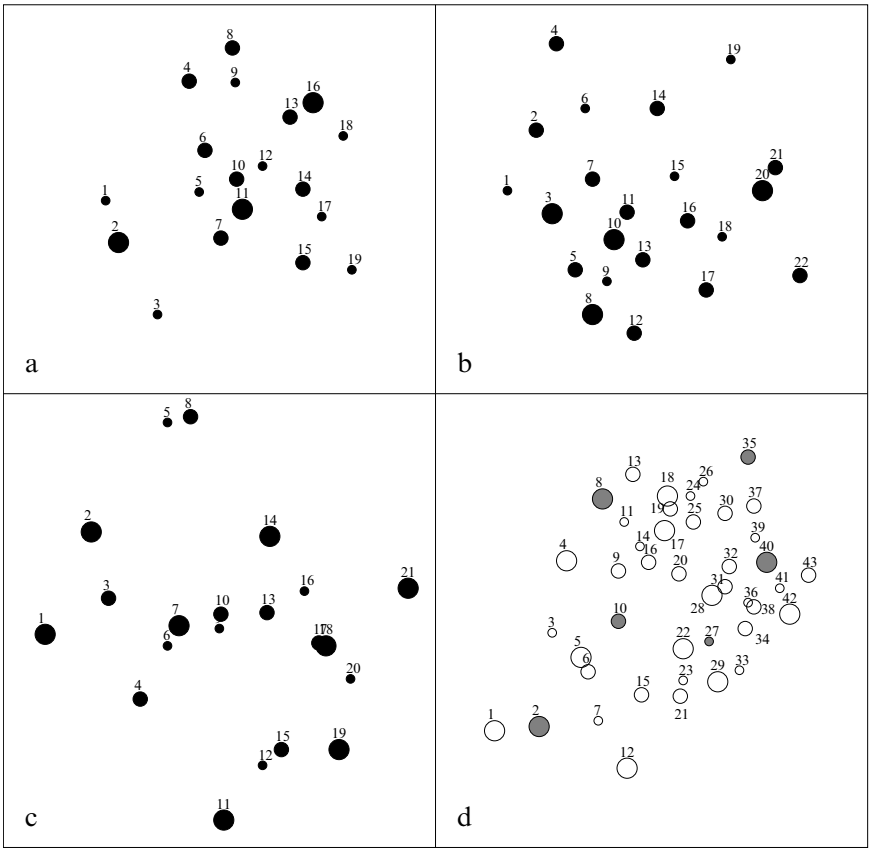

Figure 4.6: Matching objects of points. (a–c) The model objects. (d) The scene.

binary relation; i.e., the Euclidean distance between two points $d_2(i, i') = \text{dist}(i, i')$. The points, their sizes, and their distances constitute an RS. The unary property is symbolic, and this restricts the set of admissible labels for each $i \in \mathcal{S}$ as $\mathcal{L}_i = \{I \mid D_1(I) = d_1(i), \forall I \in \mathcal{L}\}$. The parameter is chosen as $[\sigma_2^{(1)}]^2 = 0.1$.

Figure 4.7 shows the matching results in which the matched object points are aligned with the scene in Fig. 4.7(d). The black points in Fig. 4.7(a) correspond to points 5, 6, 7, 11, 12, 14, 15, 17, and 19 of the object in Fig. 4.6(a). Those in Fig. 4.7(b) correspond to points 3, 5, 8, 9, 10, 11, 12, 13, 16, 17, and 18 of the object in Fig. 4.6(b). Those in Fig. 4.7(c) are points 4, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, and 21 of the object in Fig. 4.6(c). In (d) is shown the union of all the individual results. The spurious points 2, 8, 10, 27, 35, and 40 in Fig. 4.6(d) have found no counterparts in

the three objects. They are correctly classified as the NULL. There is one mismatch: Point 19 in Fig. 4.6(d) is matched to the NULL while its correct home should be point 10 of Fig. 4.6(a).

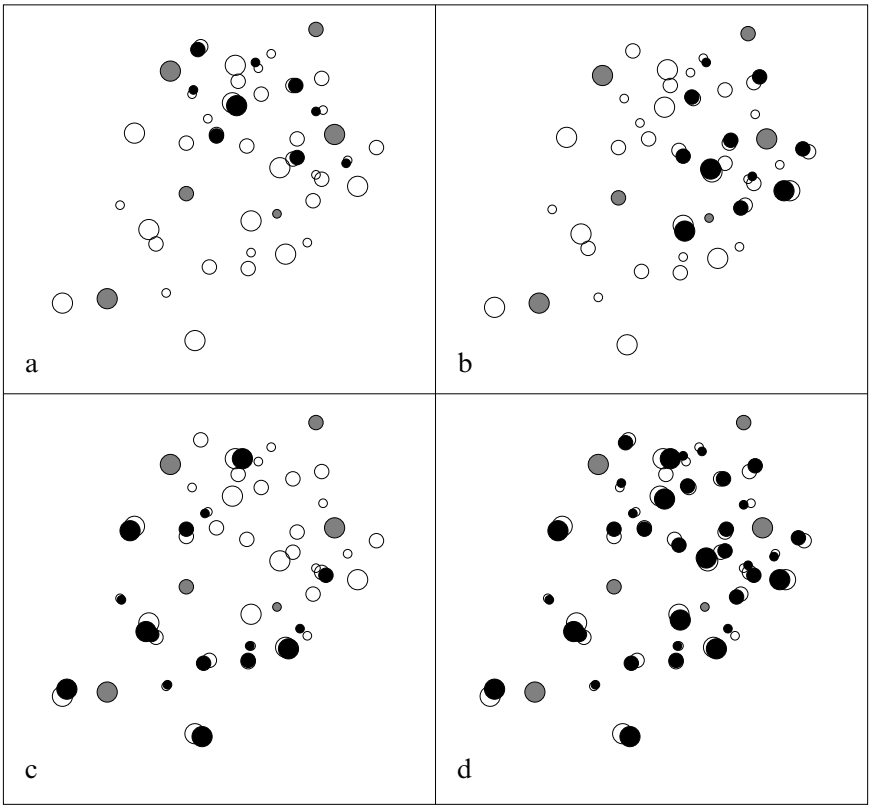

Figure 4.7: Results of matching objects of points. (a–c) Matched points (in black) from the respective models aligned with the scene. (d) All matched points aligned with the scene.

Matching Objects of Line Patterns

There are five objects made of lines, as shown in Fig. 4.8(a)–Fig. 4.8(e). The scene in Fig. 4.8(f) consists of a subset of deviated lines taken from the first three objects Fig. 4.8(a)–Fig. 4.8(c), shown as dotted lines, and spurious line features shown as dashed lines. Four types of binary relations are measured:

(1) $d_2^{(1)}(i, i')$: the angle between lines i and i';

(2) $d_2^{(2)}(i, i')$: the distance between the midpoints of the lines;

(3) $d_2^{(3)}(i, i')$: the minimum distance between the endpoints of the lines; and

(4) $d_2^{(4)}(i, i')$: the maximum distance between the endpoints of the lines.

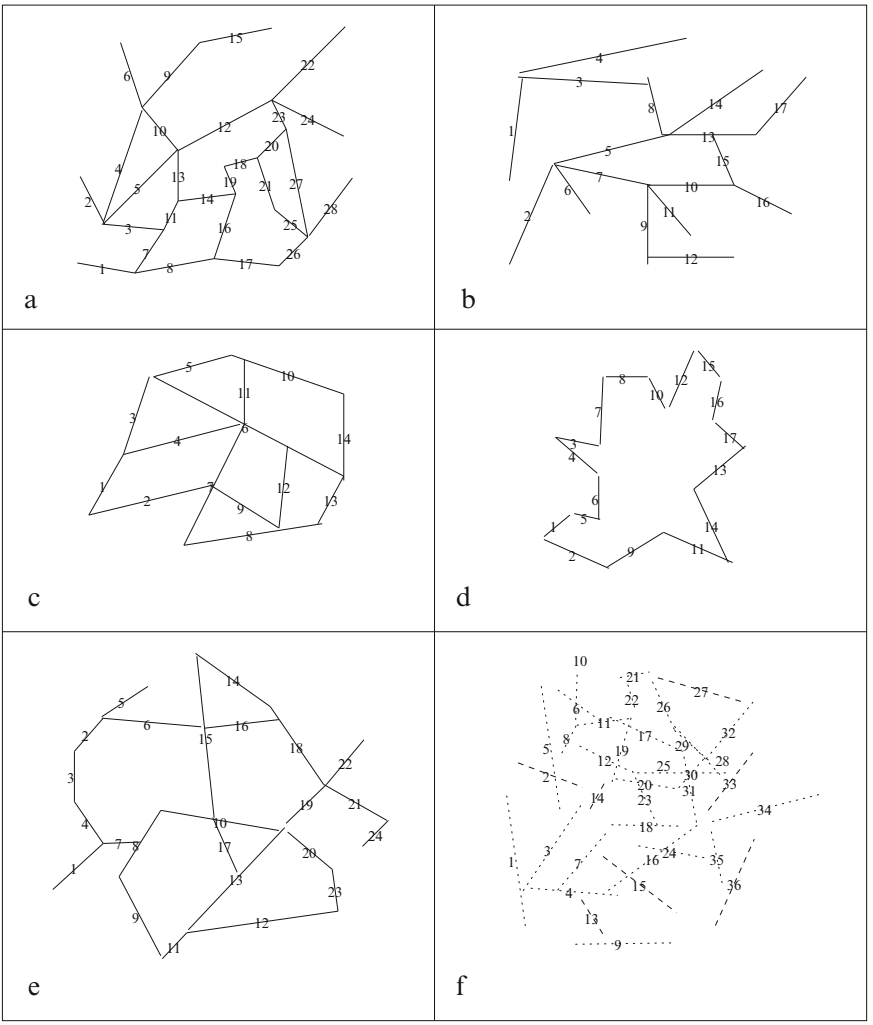

Figure 4.8: Matching objects of lines. (a–e) The five model objects. (f) The scene.

No unary relations are used. The value for the prior clique potential is fixed at $v_{20} = 0.70000$. The values for weighting the binary measurements are $1/[\sigma_2^{(1)}]^2 = 0.36422$, $1/[\sigma_2^{(2)}]^2 = 0.20910$, $1/[\sigma_2^{(3)}]^2 = 0.44354$, and $1/[\sigma_2^{(4)}]^2 = 0.74789$, which are estimated using a supervised learning procedure to be presented in Chapter 8.

Figure 4.9 shows the matching results in which the matched object lines from the first three objects are aligned with the scene in Fig. 4.9(d). The solid lines in Fig. 4.9(a) correspond to lines 11, 13, 14, 16, 17, 18, 19, 21, 25, 26, and 28 of the object in Fig. 4.8(a). Those in Fig. 4.9(b) correspond to lines 5, 8, 10, 11, 13, 14, 15, 16, and 17 of the object in Fig. 4.8(b). Those in Fig. 4.9(c) correspond to points 1, 2, 3, 5, 8, 9, 11, 12, and 14 of the object in Fig. 4.8(c). Objects in Fig. 4.9(d) and Fig. 4.9(e) do not have matches. In Fig. 4.9(d) is shown the union of all individual results. The spurious lines 2, 13, 14, 15, 27, 33, and 36 in Fig. 4.8(f) have found no counterparts in the object models. They are correctly classified as the NULL.

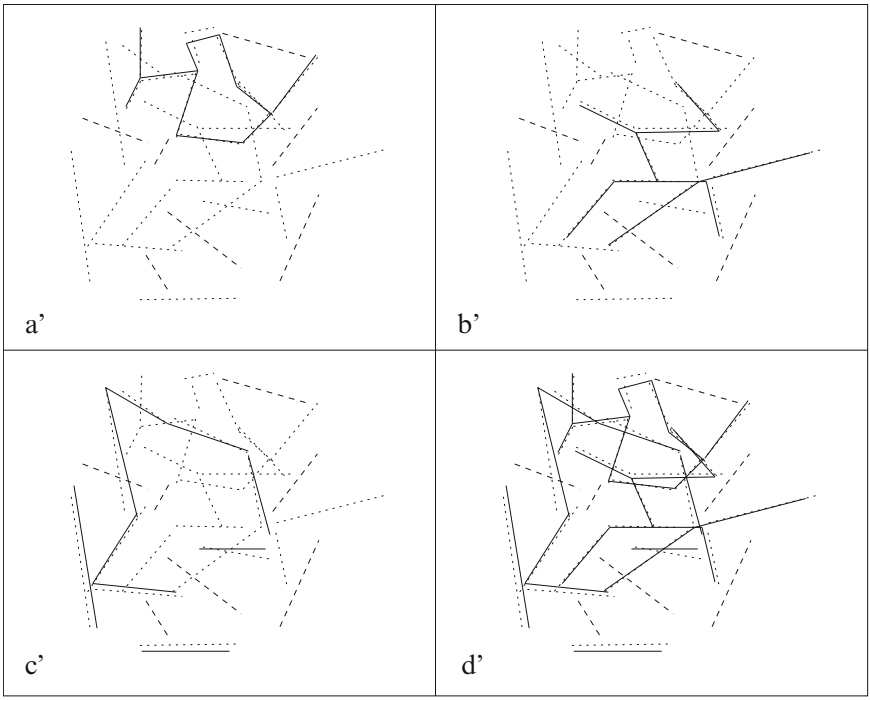

Figure 4.9: Results of matching objects of lines. (a–c) Matched lines (solid) from the respective models aligned with the scene. (d) All matched lines aligned with the scene.

Matching Curved Objects under Similarity Transformations

In the experiment shown in Fig. 4.10. There are eight model jigsaw objects. The scene contains rotated, translated, and scaled parts of the model objects, some of which are considerably occluded. Boundaries are computed from the image using the Canny detector followed by hysteresis and edge linking. After that, corners of the boundaries are located as $p_1, \ldots, p_m$.

The sites correspond to the corners on the scene curve and the labels correspond to the feature points on the model curve considered. The neighbors of a site are defined as the five forward points and the five backward points. Invariant relations are derived from the boundaries as well as the corners based on a similarity-invariant representation of curves (Li 1993). No unary properties are used ($K_1 = 0$). Only binary relations are used, which are of the following five types ($K_2 = 5$):

(1) $d_2^{(1)}(i, i')$: ratio of curve arc length $\widehat{p_i p_{i'}}$ and chord length $\overline{p_i p_{i'}}$,

(2) $d_2^{(2)}(i, i')$: ratio of curvature at p_i and $p_{i'}$,

(3) $d_2^{(3)}(i, i')$: invariant coordinates vector,

(4) $d_2^{(4)}(i, i')$: invariant radius vector, and

(5) $d_2^{(5)}(i, i')$: invariant angle vector

which are derived from the boundaries and the corners using a similarity-invariant representation of curves (Li 1993).

The parameters involved are supplied by an automated optimal estimation procedure (see Chapter 8) with the values $v_{20} = 0.7$, $1/\sigma_2^{(1)} = 0.00025$, $1/\sigma_2^{(2)} = 0$, $1/\sigma_2^{(3)} = 0.04429$, $1/\sigma_2^{(4)} = 0.02240$, and $1/\sigma_2^{(5)} = 0.21060$. The minimal labeling f^* is found by using a deterministic relaxation labeling algorithm (Hummel and Zucker 1983). The final result of recognition is shown in Fig. 4.11, in which model objects are aligned with the corresponding (sub)parts in the scene. Note that most of the model objects share common structures of round extrusion and intrusion. This means extensive ambiguities exist, which has to be resolved by using context.

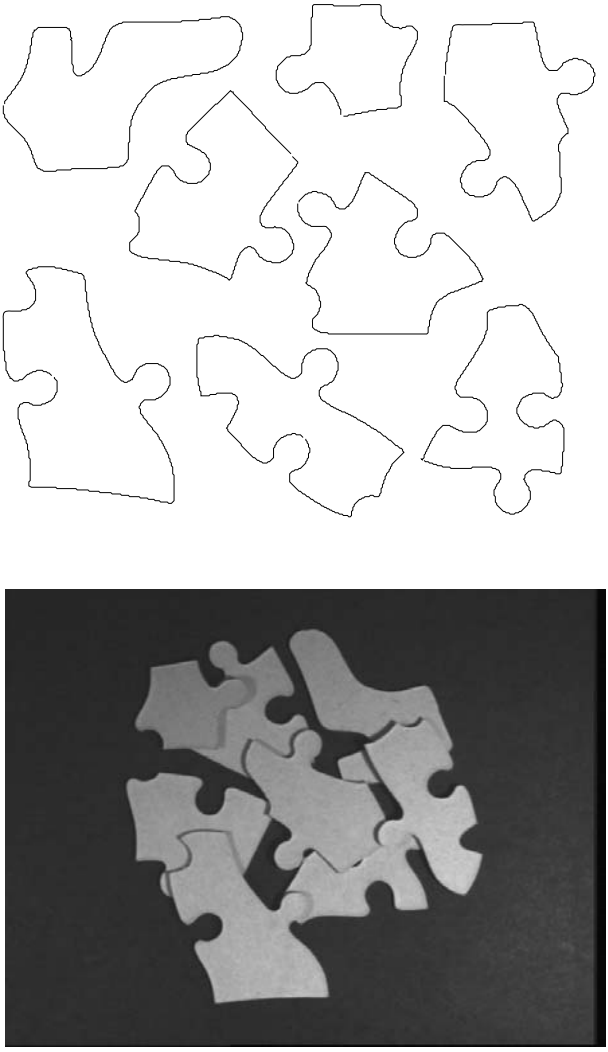

Figure 4.10: (Top) The eight model jigsaw objects. From (Li 1997b) with permission; ©1997 Kluwer. (Bottom) An overlapping scene.

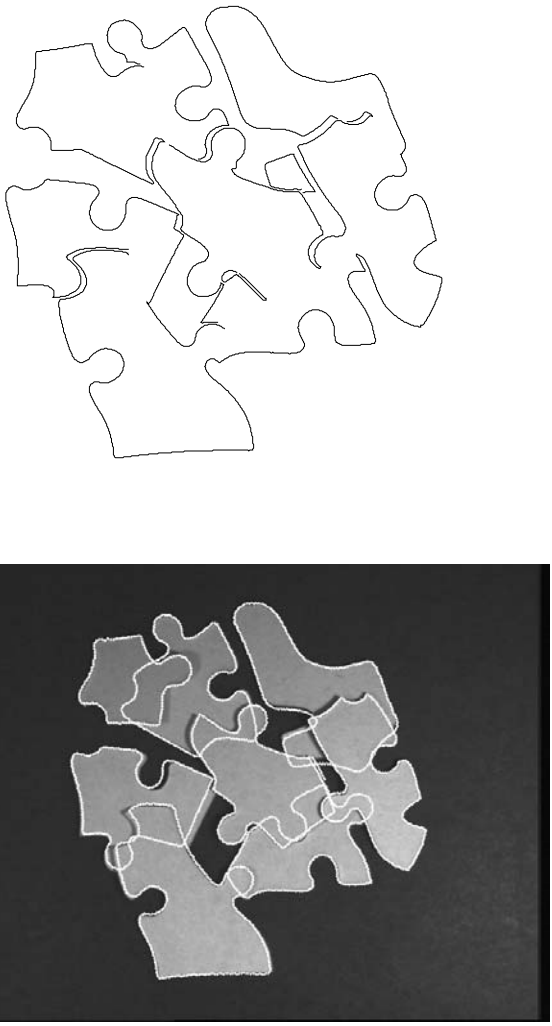

Figure 4.11: (Top) Boundaries detected from the scene. (Bottom) Matched objects are aligned with the scene. From (Li 1994a) with permission; ©1994 IEEE.

4.3 Optimal Matching to Multiple Overlapping Objects

Previously, matching between the scene and model objects was performed by considering one model object at a time. Once a part of the scene has been matched to a model object, it is excluded from subsequent matching; or better still, multiple matching results are obtained, each being optimal w.r.t. a single model object, as was done in Section 4.2 (see also (Li 1998a)). Inconsistencies may exist among the individual results because the matching to one model is done independently of the other models. Post-processing may be applied to obtain an overall consistent solution.

When a scene contains multiple mutually occluded objects, two sources of contextual constraints are required: *between-object constraints* (BOCs) and *within-object constraints* (WOCs). WOCs of an object, which describe the particular structure of the object itself, are used to identify instances of that object. BOCs, which are the constraints on features belonging to different objects, are used to discriminate between objects and unmodeled features.

Here, a statistically optimal, MAP-MRF-based formulation for recognition of multiple, partially occluded objects is presented. The MAP solution is defined w.r.t. all model objects, not just individual ones. Such a solution is optimal overall, and consistent by itself. A two-stage MAP estimation approach is proposed to reduce the computational cost. The first stage finds feature correspondence between the scene and *each* model object. The second stage solves the original, target MAP-MRF estimation problem in a much reduced space constructed from the stage 1 solutions. The energy functions for the two MAP estimation problems are formulated. BOCs are encoded in the prior distribution modeled as a Markov random field (MRF). WOCs are encoded in the likelihood distribution modeled as a Gaussian. This way, both BOCs and WOCs are incorporated into the posterior distribution. Experimental results are incorporated into the presentation to illustrate the theoretical formulation.

4.3.1 Formulation of MAP-MRF Estimation

Figure 4.12 illustrates an example of object and scene representation. Let $\mathcal{O}^{(all)} = \{\mathcal{O}^{(1)}, \ldots, \mathcal{O}^{(L)}\}$ be a number of L model objects ($L = 8$ on the left in Fig. 4.12). The objects in the scene are rotated, translated, scaled, and partially occluded versions of the model objects (Fig. 4.12, middle). The objects and the scene are represented by features (e.g., corners on boundary curves on the right in Fig. 4.12) and constraints on the feature such as unary properties and binary relations between features (the interested reader is referred to (Li 1997a) for an invariant representation of curved objects). The task is to recognize (separate and identify) the objects in the scene, optimally in the MAP sense, w.r.t. the L given objects. This can be done via feature

matching, which is aimed at establishing feature correspondence between the scene and the model objects based on partial observation of the objects.

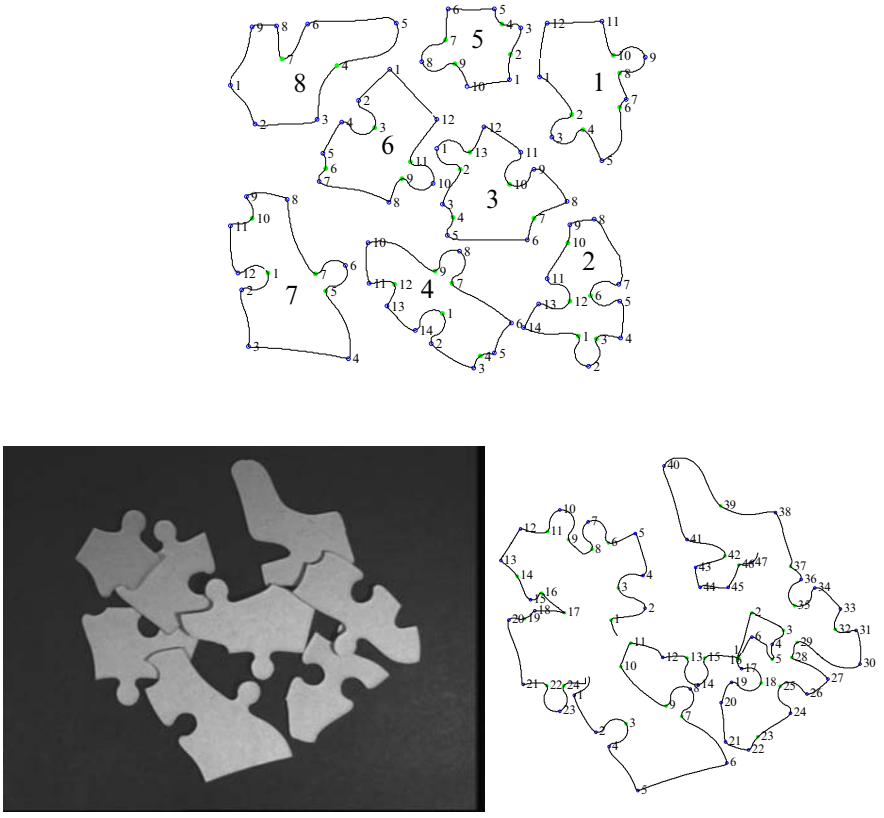

Figure 4.12: Top: Eight model jigsaw objects. Lower left: A scene. Lower right: Three boundary curves and the corner features extracted from the scene.

In this problem, we have $\mathcal{S} = \{1, \ldots, m\}$ corresponds to the set of m image features. On the right in Fig. 4.12, there are three sequences of corner features in the scene, represented by three $\mathcal{S}$ sets with $m = 24$, $m = 47$, and $m = 6$.

Let $\mathcal{L}^{(\alpha)} = \{1, \ldots, M^{(\alpha)}\}$ be a set of $M^{(\alpha)}$ labels corresponding to the set of $M^{(\alpha)}$ features for model object α ($M^{(1)} = 12$ for model object 1 on the left in Fig. 4.12). A virtual label, called the NULL and numbered 0, is added to represent everything not belonging to $\mathcal{L}^{(\alpha)}$ (such as features due to other model objects and those due to background and noise). This augments $\mathcal{L}^{(\alpha)}$ into ${\mathcal{L}^{(\alpha)}}^{+} = \{0, 1, \ldots, M^{(\alpha)}\}$. Without confusion, the notation $\mathcal{L}$ is still used to denote the augmented set $\mathcal{L}^{+}$ unless there is a need to elaborate. The set

of all model features plus the NULL is $\mathcal{L}^{(all)} = \mathcal{L}^{(1)} \cup \mathcal{L}^{(2)} \cdots \mathcal{L}^{(L)}$. It consists of $\#\mathcal{L}^{(all)}$ elements where $\#\mathcal{L}^{(all)} = 1 + \sum_{\alpha=1}^{L} M^{(\alpha)}$.

The overall matching from $\mathcal{S}$ to the L model objects is represented by a label configuration $f = \{f_1^{(\alpha_1)}, \ldots, f_m^{(\alpha_m)}\}$. It is a mapping from the set of the sites to the set of the labels, $f : \mathcal{S} \rightarrow \mathcal{L}^{(all)}$. Three things are told by a label $f_i^{(\alpha_i)} \in \mathcal{L}^{(\alpha_i)}$. (i) It separates image features belonging to a model object from those not belonging to any in the following way: If $f_i^{(\alpha_i)} \neq 0$ (a nonNULL label), then image feature i belongs to an object; otherwise, if $f_i^{(\alpha_i)} = 0$, it belongs to the background, noise, or an unmodeled object. (ii) If $f_i^{(\alpha_i)} \neq 0$, α_i indicates that image feature i belongs to model object α_i. (iii) If $f_i^{(\alpha_i)} \neq 0$, $f_i^{(\alpha_i)}$ indexes the corresponding feature of object α_i, to which image feature i is matched.

The MAP solution for matching the scene to all the objects is defined by

$$f^* = \arg \max_{f \in \mathbb{F}^{(all)}} P(f \mid d, \mathcal{O}^{(all)}) \tag{4.28}$$

where $P(f \mid d, \mathcal{O}^{(all)})$ is the posterior probability of the labeling f given the observation d and the L object models, and $\mathbb{F}^{(all)}$ is the space of all admissible configurations (solutions). When all the labels are admissible for all the sites, $\mathbb{F}^{(all)}$ is the Cartesian product of the m $\mathcal{L}^{(all)}$'s; that is, $\mathbb{F}^{(all)} = \prod_{i=1}^{m} \mathcal{L}^{(all)}$.

Assuming that f, which is a realization of a family of m random variables, is a Markov random field (MRF), then its posterior is a Gibbs distribution $P(f \mid d, \mathcal{O}^{(all)}) \propto \mathrm{e}^{-E^{(all)}(f)}$ where

$$E^{(all)}(f) \stackrel{\triangle}{=} U(f|\mathcal{O}^{(all)}) + U(d \mid f, \mathcal{O}^{(all)}) \tag{4.29}$$

is the posterior energy consisting of the prior energy $U(f \mid \mathcal{O}^{(all)})$ and the likelihood energy $U(d \mid f, \mathcal{O}^{(all)})$. The solution to problem (4.28) equivalently minimizes the posterior energy: $f^* = \arg\min_{f \in \mathbb{F}^{(all)}} E^{(all)}(f)$.

The objective of (4.28), finding the minimum f^* in $\mathbb{F}^{(all)}$, is a formidable job since the configuration space $\mathbb{F}^{(all)}$ consists of a huge number of $\#\mathbb{F}^{(all)} = (1 + \sum_{\alpha=1}^{L} M^{(\alpha)})^m$ elements when all the model features (labels) are admissible. In the following, a two-stage MAP-MRF estimation approach is proposed to tackle this problem.

Formulation of Energy Functions

The prior distribution of $f^{(\alpha)} = \{f_1^{(\alpha)}, \ldots, f_m^{(\alpha)}\}$ for matching w.r.t. object α is assumed to be an MRF and hence is a Gibbs distribution $P(f^{(\alpha)} \mid \mathcal{O}^{(\alpha)}) \propto \mathrm{e}^{-U(f^{(\alpha)} \mid \mathcal{O}^{(\alpha)})}$. The prior energy takes the form

$$U(f^{(\alpha)} \mid \mathcal{O}^{(\alpha)}) = \sum_{i \in \mathcal{S}} V_1(f_i^{(\alpha)}) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)}) \tag{4.30}$$

where $\mathcal{N}_i$ is the set of neighbors for i, and $V_1(f_i^{(\alpha)})$ and $V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)})$ are single- and pair-site clique prior potential functions, respectively for $f^{(\alpha)}$. The clique potentials are defined based on (4.11(and (4.12) as

$$V_1(f_i^{(\alpha)}) = \begin{cases} 0 & \text{if } f_i \neq 0 \\ v_{10} & \text{if } f_i = 0 \end{cases} \tag{4.31}$$

$$V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)}) = \begin{cases} 0 & \text{if } f_i^{(\alpha)} \neq 0 \text{ and } f_{i'}^{(\alpha)} \neq 0 \\ v_{20} & \text{if } f_i^{(\alpha)} = 0 \text{ or } f_{i'}^{(\alpha)} = 0 \end{cases} \tag{4.32}$$

where $v_{10} > 0$ and $v_{20} > 0$ are penalty constants for NULL labels. These definitions encode BOCs, that is, constraints between different objects and between an object and the background. In a way, this is similar to the line process model (Geman and Geman 1984) for differentiating edge and nonedge elements.

The likelihood distribution $p(d \mid f^{(\alpha)}, \mathcal{O}^{(\alpha)})$ describes the statistical properties of the features seen in the scene and is therefore conditioned on pure nonNULL matches ($f_i^{(\alpha)} \neq 0$) only. The likelihood is a Gibbs distribution with the energy function, which is based on (4.15)

$$\begin{aligned} U(d \mid f^{(\alpha)}, \mathcal{O}^{(\alpha)}) \quad = \quad & \sum_{i \in \mathcal{S}, f_i^{(\alpha)} \neq 0} V_1(d_1(i) \mid f_i^{(\alpha)}) \\ & \sum_{i \in \mathcal{S}, f_i^{(\alpha)} \neq 0} \; \sum_{i' \in \mathcal{S} \setminus i, f_{i'}^{(\alpha)} \neq 0} V_2(d_2(i, i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha)}) \end{aligned} \tag{4.33}$$

where $d_1(i)$ is the set of unary properties of image feature i, $d_2(i, i')$ is the set of binary relations between i and i', and $V_1(d_1(i) \mid f_i^{(\alpha)})$ and $V_2(d_2(i, i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha)})$ are the potentials in the likelihood distributions (where the distributions may be assumed to be Gaussian). The likelihood potentials encode WOCs, that is, constraints on image features belonging to object α only. Both BOCs and WOCs are incorporated into the posterior distribution with the posterior energy $E^{(\alpha)}(f^{(\alpha)}) = U(f^{(\alpha)} | \mathcal{O}^{(\alpha)}) + U(d \mid f^{(\alpha)}, \mathcal{O}^{(\alpha)})$.

The posterior in stage 2 is the target posterior distribution $P(f \mid d, \mathcal{O}^{(all)})$ of the original problem (1), with the posterior energy $U(f \mid d, \mathcal{O}^{(all)}) = U(f \mid \mathcal{O}^{(all)}) + U(d \mid f, \mathcal{O}^{(all)})$. In this stage, if f_i is non-NULL, then it is associated with a model α, so it should be read as $f_i^{(\alpha)}$ (and $f_{i'}$ as $f_{i'}^{(\alpha')}$). In the following derivation, it is assumed that model objects α and α' are independent of each other when $\alpha \neq \alpha'$.

The prior energy is

$$U(f \mid \mathcal{O}^{(all)}) = \sum_{i \in \mathcal{S}} V_1(f_i \mid \mathcal{O}^{(all)}) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'} \mid \mathcal{O}^{(all)}) \tag{4.34}$$

The single-site prior potentials are defined as $V_1(f_i \mid \mathcal{O}^{(all)}) = V_1(f_i^{(\alpha)})$, which is the same as that in (4.11) for matching to a single model object α. The pair-site potential $V_2(f_i, f_{i'} \mid \mathcal{O}^{(all)}) = V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha')} \mid \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')})$, where

$$V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha')} \mid \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')}) = \begin{cases} V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)}) & \text{if } \alpha = \alpha' \\ v_{20} & \text{otherwise} \end{cases} \tag{4.35}$$

where $V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)})$ are as defined in (4.12) for matching to object α. Substituting (4.12) into (4.35) yields

$$V_2(f_i, f_{i'} \mid \mathcal{O}^{(all)}) = \begin{cases} 0 & \text{if } (\alpha = \alpha') \text{ and } (f_i \neq 0) \text{ and } (f_{i'} \neq 0) \\ v_{20} & \text{otherwise} \end{cases} \tag{4.36}$$

where f_i is associated with object α and $f_{i'}$ with object α'. The definitions above are an extension of (4.12) for dealing with multiple objects. In (4.12), features due to other objects (not belonging to object α) are all labeled as NULL ; (4.36) simply takes this into consideration.

The likelihood energy is

$$\begin{aligned} U(d \mid f, \mathcal{O}^{(all)}) \quad = \quad & \sum_{i \in \mathcal{S}, f_i \neq 0} V_1(d_1(i) \mid f_i, \mathcal{O}^{(all)}) + \\ & \sum_{i \in \mathcal{S}, f_i \neq 0} \sum_{i' \in \mathcal{S}_{\setminus i}, f_{i'} \neq 0} V_2(d_2(i, i') \mid f_i, f_{i'}, \mathcal{O}^{(all)}) \end{aligned} \tag{4.37}$$

The single-site likelihood potentials are $V_1(d_1(i) \mid f_i, \mathcal{O}^{(all)}) = V_1(d_1(i) \mid f_i^{(\alpha)})$, defined in the same way as for (4.15). The pair-site likelihood potentials are $V_2(d_2(i, i') \mid f_i, f_{i'}, \mathcal{O}^{(all)}) = V_2(d_2(i, i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha')}, \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')})$, where

$$V_2(d_2(i, i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha')}, \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')}) = \begin{cases} V_2(d_2(i, i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha)}) & \text{if } \alpha = \alpha' \\ 0 & \text{otherwise} \end{cases} \tag{4.38}$$

where $V_2(d_2(i, i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha)})$ are defined in the same way as for (4.15) when matching to object α.

Some parameters have to be determined, such as v_{10} and v_{20} in the MRF prior, in order to define the MAP solutions completely. They may be estimated by using a supervised learning algorithm (see Chapter 8).

4.3.2 Computational Issues

Finding Solution in Two Stages

Stage 1 solves L subproblems, $f^{(\alpha)} = \arg\max_{f \in \mathbb{F}^{(\alpha)}} P(f \mid d, \mathcal{O}^{(\alpha)})$ for $\alpha = 1, \ldots, L$, resulting in L MAP solutions $\{f^{(1)}, \ldots, f^{(L)}\}$. Then, a reduced configuration space is constructed from the L solutions. In stage 2, the solution of (4.28) w.r.t. all the L objects is found in the reduced space.

Stage 1 matches the scene to each of the L objects individually (which can be done in parallel for all model objects). Let $P(f \mid d, \mathcal{O}^{(\alpha)}) \propto \mathrm{e}^{-E^{(\alpha)}(f)}$ be the posterior distribution of f for matching the scene to object α $(1 \le \alpha \le L)$. The MAP-MRF solution for this is $f^{(\alpha)} = \{f_1^{(\alpha)}, \ldots, f_m^{(\alpha)}\} = \arg\min_{f \in \mathbb{F}^{(\alpha)}} E^{(\alpha)}(f)$ where $f_i^{(\alpha)}$ denotes the corresponding model feature in object α. The configuration space $\mathbb{F}^{(\alpha)}$ for object α consists of only $\#\mathbb{F}^{(\alpha)} = (1 + M^{(\alpha)})^m$ elements. For the L objects, the total size is $\sum_1^L (1 + M^{(\alpha)})^m$, much smaller than $\#\mathbb{F}^{(all)}$ which is $(1 + \sum_{\alpha=1}^{L} M^{(\alpha)})^m$.

Two things are told in $f^{(\alpha)}$: First, it separates image features belonging to object α from the other image features in the following way: If image feature i belongs to object α, then $f_i^{(\alpha)} \neq 0$ (a nonNULL label); otherwise, $f_i^{(\alpha)} = 0$. Second, if $f_i^{(\alpha)} \neq 0$, $f_i^{(\alpha)}$ is the model feature to which image feature i is matched.

$f^{(\alpha)}$ is optimal w.r.t. model object α but not to another one, so inconsistencies may exist among the L MAP solutions. A feature $i \in \mathcal{S}$ in the scene may have been matched to more than one model feature belonging to different objects; that is, there may exist more than one $\alpha \in \{1, \ldots, L\}$ for which $f_i^{(\alpha)} \neq 0$. For example, in Fig. 4.13, model instances found by $f^{(2)}$, $f^{(3)}$, and $f^{(7)}$ compete for a common part of the scene, which is mostly due to the common structures, such as the round extrusions and intrusions, of the matched objects. (Figure 4.13 also shows that a MAP solution allows multiple instances of a model object: e.g., $f^{(7)}$ contains two instances of model object 7)

Stage 2 solves the original MAP problem of (4.28) w.r.t. all the L objects in a reduced solution space constructed from the stage 1 solutions $\{f^{(1)}, \ldots, f^{(L)}\}$ (see the next subsection for the construction of the reduced space). This also resolves possible inconsistencies among the L MAP solutions because only one label f_i is assigned to i in the overall solution f. Figure 4.14 shows the overall optimal result, which is the output of stage 2, for the MAP recognition of the scene w.r.t. all the models. It is consistent by itself.

Reduced Solution Space

Consider an illustration in Table 4.1, where a scene with $m = 12$ features is matched to $L = 5$ model objects, resulting in five MAP solutions $f^{(\alpha)}$ $(\alpha = 1, \ldots, 5)$. Let $\mathcal{S}' \subset \mathcal{S}$ be the set of sites that according to the stage 1 solution have been matched to more than one nonNULL label, $\mathcal{S}' = \{i \in \mathcal{S} \mid f_i^{(\alpha)} \neq 0 \text{ for more than one } \alpha\}$. For example, $\mathcal{S}' = \{8, 9, 10, 11\}$ for the $f^{(\alpha)}$'s in Table 4.1. We can derive the reduced set of admissible labels for i, denoted by $\mathcal{L}_i^{(all)}$, from the $f^{(\alpha)}$'s as follows.

- For $i \in \mathcal{S}'$, $\mathcal{L}_i^{(all)}$ consists of all the nonNULL labels assigned to i by the $f^{(\alpha)}$'s, plus the NULL label; that is, $\mathcal{L}_i^{(all)} = 0 \cup \{f_i^{(\alpha)} \neq$

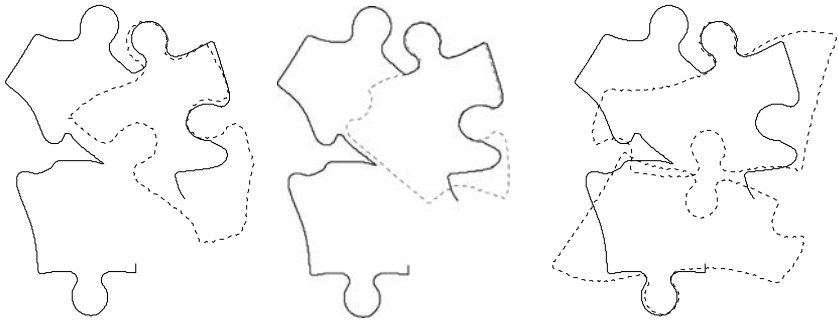

Figure 4.13: $f^{(2)}$, $f^{(3)}$, and $f^{(7)}$ compete for a common part of the scene.

Figure 4.14: The overall matching and recognition result.

$0 \mid \alpha = 1, \ldots, L\}$. For the case in Table 4.1, for example, $\mathcal{L}_8^{(all)} = \{0, 10^{(2)}, 1^{(4)}, 7^{(5)}\}$, $\mathcal{L}_9^{(all)} = \{0, 9^{(2)}, 6^{(5)}\}$, $\mathcal{L}_{10}^{(all)} = \{0, 7^{(2)}, 3^{(3)}, 5^{(5)}\}$, and $\mathcal{L}_9^{(all)} = \{0, 4^{(3)}, 4^{(5)}\}$.

- For $i \notin \mathcal{S}'$, $\mathcal{L}_i^{(all)}$ consists in a unique label; e.g., $\mathcal{L}_6^{(all)} = \{3^{(4)}\}$, and $\mathcal{L}_3^{(all)} = \{0\}$.

Object α contributes one or no label to $\mathcal{L}_i^{(all)}$, as opposed to $M^{(\alpha)}$ labels before the reduction. For $i \in \mathcal{S}'$, the size of $\mathcal{L}_i^{(all)}$ is at most $L + 1$; for

Table 4.1: Matching and recognition of an image containing $m = 12$ features to $L = 5$ objects.

$i =$	1	2	3	4	5	6	7	8	9	10	11	12
$f^{(1)}$	0	0	0	0	0	0	0	0	0	0	0	0
$f^{(2)}$	0	0	0	0	0	0	0	$10^{(2)}$	$9^{(2)}$	$7^{(2)}$	0	0
$f^{(3)}$	0	0	0	0	0	0	0	0	0	$3^{(3)}$	$4^{(3)}$	0
$f^{(4)}$	0	0	0	$5^{(4)}$	$4^{(4)}$	$3^{(4)}$	$2^{(4)}$	$1^{(4)}$	0	0	0	0
$f^{(5)}$	0	0	0	0	0	0	0	$7^{(5)}$	$6^{(5)}$	$5^{(5)}$	$4^{(5)}$	$3^{(5)}$
f^*	0	0	0	$5^{(4)}$	$4^{(4)}$	$3^{(4)}$	$2^{(4)}$	$7^{(5)}$	$6^{(5)}$	$5^{(5)}$	$4^{(5)}$	$3^{(5)}$

$i \notin \mathcal{S}'$, the size of $\mathcal{L}_i^{(all)}$ is one, whereas before the reduction, the size was $1 + \sum_{\alpha=1}^{L} M^{(\alpha)}$ for every i.

The reduced space is constructed as $\mathbb{F}_{reduced}^{(all)} = \mathcal{L}_1^{(all)} \times \mathcal{L}_2^{(all)} \times \cdots \times \mathcal{L}_m^{(all)}$, where $\times$ is the Cartesian product of sets. Its size is much reduced. For the case in Table 4.1, the previous size of the raw solution space $\#\mathbb{F}^{(all)} = (\sum_{\alpha=1}^{5} M^{(\alpha)} + 1)^{12}$ configurations (e.g., 31384283770 for $M^{(\alpha)} = 10$) is reduced to $4 \times 3 \times 4 \times 3 = 144$. It is so small that an exhaustive search is affordable.

Stage 2 performs the target minimization in $\mathbb{F}_{reduced}^{(all)}$. In an iterative search algorithm, only those labels on the sites $i \in \mathcal{S}'$ are subject to changes, whereas those not in $\mathcal{S}'$ are fixed. This is equivalent to maximizing the conditional posterior $f_{\mathcal{S}'}^* = \arg\max_{f_{\mathcal{S}'} \in \mathbb{F}_{\mathcal{S}'}^{(all)}} P(f_{\mathcal{S}'} \mid d, f_{\mathcal{S}-\mathcal{S}'}, \mathcal{O}^{(all)})$, where $f_{\mathcal{S}'} = \{f_i \mid i \in \mathcal{S}'\}$ is the set of labels to be updated, $f_{\mathcal{S}-\mathcal{S}'} = \{f_i \mid i \in \mathcal{S} - \mathcal{S}'\}$ is the set of labels that are fixed during the maximization, and $\mathbb{F}_{\mathcal{S}'}^{(all)} = \prod_{i \in \mathcal{S}'} \mathcal{L}_i^{(all)}$.

A crucial question for the validity of the two-stage approach is whether the solution of (4.28) is contained in the reduced solution space $\mathbb{F}_{reduced}^{(all)}$. The necessary and sufficient condition is that $f^{(\alpha)}$ contains correct matches for object α (it is also allowed to contain spurious matches). Now that the global solution can be found in $\mathbb{F}_{reduced}^{(all)}$ (e.g., by an exhaustive search), this means that the two-stage strategy can find the global solution of (4.28) if $\mathcal{L}_i^{(\alpha)}$ derived from $f^{(\alpha)}$ contains the correct matching components for the original problem.

The optimization in MAP matching and recognition is combinatorial. While an optimum is sought in a global sense, many optimization algorithms are based on local optimization. The Hummel-Zucker relaxation labeling algorithm (Hummel and Zucker 1983) is preferable in terms of the minimized energy value and computational costs and is used in the implementation. It converges after dozens of iterations. The computational time is dominated

by relaxation labeling in the first stage and is roughly the complexity of the whole system.

4.4 Pose Computation

Pose computation aims to estimate the transformation needed to map an object model from the model coordinate system into the sensory data (Ayache and Faugeras 1986; Faugeras and Hebert 1986; Bolles and Horaud 1986; Stockman 1987; Haralick et al. 1989; Grimson 1990; Umeyama 1991). In this section, we derive two MRF models for pose estimation. The first is a model for pose clustering from corresponding point data containing multiple poses and outliers. The second model attempts to solve 3D matching and pose simultaneously from a 2D image without using view invariants.

4.4.1 Pose Clustering and Estimation

The problem of pose clustering is stated as follows. Let a set of corresponding points be given as the data, $d = \{(p_i, P_i) \mid i \in \mathcal{S}\}$, where p_i's are the *model* features, P_i are the *scene* features[4] and $\mathcal{S} = \{1, \ldots, m\}$ indexes the set of the matched pairs. Let f_i be the geometric transformation from p_i to P_i, and consider the set $f = \{f_1, \ldots, f_m\}$ as a configuration of the "pose field". In the case of noiseless, perfect correspondences, the following m equations, each transforming a model feature to a scene feature, should hold simultaneously:

$$P_i = f_i(p_i) \qquad i \in \mathcal{S} \tag{4.39}$$

We want to find the optimal pose configuration in the MAP sense; i.e., $f^* = \arg\min_f U(f \mid d)$.

Assume that each f_i is confined to a certain class $\mathcal{L}$ of pose transformations such that the admissible pose space is $\mathbb{F} = \mathcal{L}^m$. This imposes constraints on the parameters governing f_i. The number of transformation parameters (degree of freedom) needed depends on the class of transformation and the representation adopted for the pose transformation. In the case of the 3D–3D Euclidean transformation, for example, it can consists of an orthogonal rotation O_i followed by a translation T_i (i.e., $f_i = (O_i, T_i)$); the relation between the corresponding points is $P_i = f_i(p_i) = O_i p_i + T_i$. The simple matrix representation needs 12 parameters: nine elements in the rotation matrix O_i plus three elements in the translation vector T_i. The rotation angle representation needs six parameters: three for the three rotation angles and three for the translation. Quaternions provide still another choice. A single pair (p_i, P_i) alone is usually insufficient to determine a pose transformation f_i; more are needed for the pose to be fully determined.

[4]Note that, in this section, the uppercase notations are for models and the lowercase notations for the scene.

If all the pairs in the data, d, are inliers and are due to a single transformation, then all f_i, $i \in \mathcal{S}$, which are points in the pose space, must be close to each other; and the errors $\|P_i - f_i(p_i)\|$, where $\|\cdot\|$ is the Euclidean distance, must all be small. Complications increase when there are multiple pose clusters and outlier pairs. When there are multiple poses, f_i's should form distinct clusters. In this case, the set f is divided into subsets, each giving a consistent pose transformation from a partition of $\{p_i\}$ to a partition of $\{P_i\}$. Figure 4.15 illustrates a case in which there are two pose clusters and some outliers. Outlier pairs, if contained in the data, should be excluded from the pose estimation because they can cause large errors. Multiple pose identification with outlier detection has a close affinity to the prototypical problem of image restoration involving discontinuities (Geman and Geman 1984) and to that of matching overlapping objects using data containing spurious features (Li 1994a).

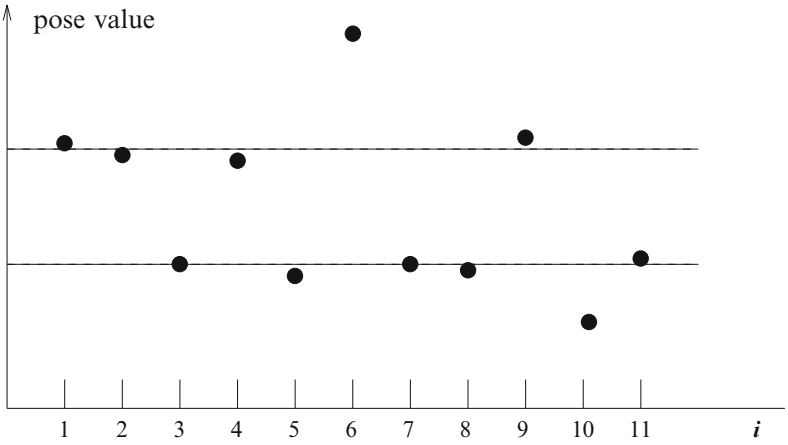

Figure 4.15: Pose clusters in one-dimensional parameter space. Poses f_1, f_2, f_4, and f_9, due to pairs 1, 2, 4, and 9, agree to one transformation and poses f_3, f_5, f_7, f_8, and f_{11} agree to another. Poses f_6 and f_{10} form isolated points so that pair 6 and pair 10 are outliers.

Now we derive the MAP-MRF formulation. The neighborhood system is defined by

$$\mathcal{N}_i = \{i' \in \mathcal{S} \mid [\text{dist}(p_i, p_{i'})]^2 \leq r,\ i' \neq i\} \tag{4.40}$$

where $\text{dist}(A, B)$ is some suitably defined measure of distance between model features and the scope r may be reasonably related to the size of the largest model object. We consider cliques of up to order two, and so the clique set $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$, where $\mathcal{C}_1 = \{\{i\} \mid i \in \mathcal{S}\}$ is the set of single-site (first-order) cliques and $\mathcal{C}_2 = \{\{i, i'\} \mid i' \in \mathcal{N}_i, i \in \mathcal{S}\}$ the pair-site (second-order) cliques.

Under a single pose transformation, nearby model features are likely to appear together in the scene, whereas model features distantly apart tend to be less likely. This is the coherence of spatial features. We characterize this using the Markovianity condition $p(f_i \mid f_{\mathcal{S}-\{i\}}) = P(f_i \mid f_{\mathcal{N}_i})$. The positivity condition $P(f) > 0$ also holds for all $f \in \mathbb{F}$, where $\mathbb{F}$ is the set of admissible transformations.

The MRF configuration f follows a Gibbs distribution. The two-site potentials determine interactions between the individual f_i's. They may be defined as

$$V_2(f_i, f_{i'}) = g(\|f_i - f_{i'}\|) \tag{4.41}$$

where $\|\cdot\|$ is a norm in the pose space and $g(\cdot)$ is some function. To be able to separate different pose clusters, the function $g(\cdot)$ should stop increasing as $\|f_i - f_{i'}\|$ becomes very large. A choice is

$$g_{\alpha,T}(\eta) = \min\{\eta^2, \alpha\} \tag{4.42}$$

where $\alpha > 0$ is a threshold parameter; this is the same as that used in the line process model for image restoration with discontinuities (Geman and Geman 1984; Marroquin et al. 1987; Blake and Zisserman 1987). It may be any APF defined in (5.28). Its value reflects the cost associated with the pair of pose labels f_i and $f_{i'}$ and will be large when f_i and $f_{i'}$ belong to different clusters. But it cannot be arbitrarily large since a large value such as might be given by a quadratic g tends to force the f_i and f_i' to stay in one cluster as the result of energy minimization, even when they should not. Using an APF imposes piecewise smoothness.

The single-site potentials $V_1(f_i)$ may be used to force f_i to stay in the admissible set $\mathcal{L}$ if such a force is needed. For example, assume $f_i = (O_i, T_i)$ is a 2D-2D Euclidean transformation. Then, the rotation matrix $O_i = [o_{i,r,s} \mid r, s = 1, 2]$ must be orthogonal. The unary potential for the orthogonality constraint can be expressed by $(o_{i,1,1}o_{i,2,1} + o_{i,1,2}o_{i,2,2})^2 + (o_{i,1,1}o_{i,1,2} + o_{i,2,1}o_{i,2,2})^2$. It has the value of zero only when O_i is orthogonal. If no scale change is allowed, then the scaling factor should be exactly one, and an additional term $[\det(O_i) - 1]^2$ can be added, where $\det(O_i)$ is the determinant. Adding these two gives the single-site potential as

$$\begin{aligned} V_1(f_i) \quad = \quad & a\left[(o_{i,1,1}o_{i,2,1} + o_{i,1,2}o_{i,2,2})^2 + (o_{i,1,1}o_{i,1,2} + o_{i,2,1}o_{i,2,2})^2\right] \\ & + b[\det(O_i) - 1]^2 \end{aligned} \tag{4.43}$$

where a and b are the weighting factors. In this case, V_1 imposes the orthogonality. It is also possible to define $V_1(f_i)$ for other classes of transformations. Summing all prior clique potentials yields the following prior energy

$$U(f) = \sum_{i \in \mathcal{S}} V_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \tag{4.44}$$

which defines the prior distribution $P(f)$.

The likelihood function is derived below. Assume that the features are point locations and that they are subject to the additive noise model, $P_i = f_i(p_i) + e_i$, where $e_i \sim N(0, \sigma^2)$ is a vector of i.i.d. Gaussian noise. Then the distribution of the data d conditional on the configuration f is

$$P(d \mid f) \propto \mathrm{e}^{-U(d \mid f)} \tag{4.45}$$

where the likelihood energy is

$$U(d \mid f) = \sum_{i \in \mathcal{S}} \|f_i(p_i) - P_i\|^2 / [2\sigma^2] \tag{4.46}$$

The location $f_i(p_i)$ is the conditional "mean" of the random variable P_i. The quantity $f_i(p_i) - P_i$ reflects the error between the location $f_i(p_i)$ predicted by f_i and the actual location P_i.

After that, the posterior energy follows immediately as $U(f \mid d) = U(f) + U(d \mid f)$. The optimal solution is $f^* = \arg\min_f U(f \mid d)$. As the result of energy minimization, inlier pairs undergoing the same pose transformation will form a cluster, whereas outlier pairs will form isolated points in the pose space, as illustrated in Fig. 4.15.

4.4.2 Simultaneous Matching and Pose Estimation

In the previous pose estimation formulation, a set of matched pairs is assumed to be available. Here we assume the situation in which the matching has not been done and pose estimation has to be performed during the matching. Pose estimation during matching is practiced when invariants are unavailable or difficult to compute; e.g., because the class of transformations is not linear or involves projections. In the following, an MRF model for simultaneous 3D-from-2D matching and pose estimation is derived without using view invariants. Matching and pose estimation are jointly sought as in (Wells 1991). The formulation is an extension of that given in Section 4.2.

Let $\mathcal{S} = \{1, \ldots, m\}$ index a set of m points on a 3D *model* object, $\{p_i \mid i \in \mathcal{S}\}$. Let $\mathcal{L} = \{1, \ldots, M\}$ be the label set indexing a set of M *scene* points in 2D, $\{P_I \mid I \in \mathcal{L}\}$, and $\mathcal{L}^+ = \{0\} \bigcup \mathcal{L}$ be the augmented set with 0 representing the NULL label. Let $f = \{f_1, \ldots, f_m\}$, $f_i \in \mathcal{L}^+$, denote the matching from the $\{p_i\}$ to $\{P_I\} \bigcup$NULL. When i is assigned the virtual point 0, $f_i = 0$, it means that there is no corresponding point found in the physically existing point set $\mathcal{L}$. Let $\mathcal{T}$ be the *projective* pose transformation from the 3D model points p_i to the matched 2D image points P_{f_i} ($f_i \neq 0$). We have $P_{f_i} = \mathcal{T}(p_i)$, for all i for which $f_i \neq 0$, under an exact pose.

Now we derive the MAP-MRF formulation. The neighborhood system is defined by

$$\mathcal{N}_i = \{i' \in \mathcal{S} \mid \|p_i - p_{i'}\|^2 \leq r,\ i' \neq i\} \tag{4.47}$$

The single-site potential is an extension of (4.11) as

$$V_1(f_i, \mathcal{T}) = \begin{cases} v_{10} & \text{if } f_i = 0 \\ v(\mathcal{T}(p_i), P_{f_i}) & \text{otherwise} \end{cases} \tag{4.48}$$

where v_{10} is a constant. The function $v(\mathcal{T}(p_i), P_{f_i})$ encodes the prior knowledge about $\mathcal{T}$. It may include prior terms, such as V_1 in the previous subsection for the admissibility of pose transformations. If the p.d.f. of the pose is known (e.g., to be a normal distribution centered at a known mean pose, which is assumed in (Wells 1991)), then $v(\mathcal{T}(p_i), P_{f_i})$ is a multivariate Gaussian function. The pair-site potential is defined as in (4.12)

$$V_2(f_i, f_{i'}, \mathcal{T}) = \begin{cases} v_{20} & \text{if } f_i = 0 \text{ or } \ f_{i'} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.49}$$

where v_{20} is a constant.

The likelihood function characterizes the distribution of the errors and relates to the observation model and the noise in it. Given $f_i \neq 0$, the model point $p_i = (x_i, y_i, z_i)$ is projected to a point $\mathcal{T}(p_i) \triangleq \hat{P}_{f_i} = (\hat{X}_i, \hat{Y}_i)$ by the projective transformation $\mathcal{T}$. In the inexact situation, $\hat{P}_{f_i} \neq P_{f_i}$, where $P_{f_i} = (X_i, Y_i) \triangleq d_1(i)$ is the corresponding image point actually observed.

Assume the additive noise model

$$P_{f_i} = \mathcal{T}(p_i) + e_i = \hat{P}_{f_i} + e_i \tag{4.50}$$

where $e_i \sim N(0, \sigma^2)$ is a vector of i.i.d. Gaussian noise. Then the likelihood function is

$$p(d_1(i) \mid f_i, \mathcal{T}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 \mathrm{e}^{-V_1(d_1(i) \mid f_i, \mathcal{T})} \tag{4.51}$$

where

$$V_1(d_1(i) \mid f_i, \mathcal{T}) = [(X_i - \hat{X}_i)^2 + (Y_i - \hat{Y}_i)^2]/[2\sigma^2] \tag{4.52}$$

is the unary likelihood potential. The joint likelihood is then $p(d_1 \mid f, \mathcal{T}) = \prod_i p(d_1(i) \mid f_i, \mathcal{T})$, where d_1 denotes the set of unary properties.

We also make use of the distances as an additional binary constraint. The distance, $\|p_i - p_{i'}\|$, between the two model points in 3D is projected to the distance

$$d_2(i, i') = \|\hat{P}_{f_i} - \hat{P}_{f_{i'}}\| = \sqrt{(X_i - X_{i'})^2 + (Y_i - Y_{i'})^2} \tag{4.53}$$

in 2D. Its p.d.f. can be derived, based on the distribution of the projected points given in (4.50), in the following way. Let $X = (X_i, Y_i, X_{i'}, Y_{i'})$. These random variables are assumed independent, so their joint conditional p.d.f. is

$$p(X_i, Y_i, X_{i'}, Y_{i'} | f_i, f_{i'}, \mathcal{T}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^4 \mathrm{e}^{-[(X_i - \hat{X}_i)^2 + (Y_i - \hat{Y}_i)^2 + (X_{i'} - \hat{X}_{i'})^2 + (Y_{i'} - \hat{Y}_{i'})^2]/[2\sigma^2]} \tag{4.54}$$

Introduce new random variables, $Z(X) = (Z_1, Z_2, Z_3, Z_4)$, as

$$\begin{array}{rcl} Z_1 & = & \sqrt{(X_i - X_{i'})^2 + (Y_i - Y_{i'})^2} \\ Z_2 & = & Y_i \\ Z_3 & = & X_{i'} \\ Z_4 & = & Y_{i'} \end{array} \tag{4.55}$$

each of which is a function of the X variables. Note that we are deriving the p.d.f. of Z_1. The inverse of $Z(X)$, denoted by $X = X(Z)$, is determined by

$$\begin{array}{rcl} X_i & = & \sqrt{Z_1^2 - (Z_2 - Z_4)^2} + Z_3 \\ Y_i & = & Z_2 \\ X_{i'} & = & Z_3 \\ Y_{i'} & = & Z_4 \end{array} \tag{4.56}$$

The Jacobian of the inverse is defined to be the determinant

$$J = \det\left[\nabla Z(X)\right] = \frac{Z_1}{\sqrt{Z_1^2 - (Z_2 - Z_4)^2}} \tag{4.57}$$

which is a function of the Z variables. The joint conditional p.d.f. $p_Z(Z)$ for Z can be derived from the joint p.d.f. (4.54) using the relation (Grimmett 1982)

$$p_Z(Z \mid f_i, f_{i'}, \mathcal{T}) = p_X(X(Z) \mid f_i, f_{i'}, \mathcal{T}) \times |J| \tag{4.58}$$

The conditional distribution of $Z_1 = d_2(i, i')$ is then the conditional marginal

$$\begin{aligned} p(d_2(i,i') \quad \mid \quad & f_i, f_{i'}, \mathcal{T}) = p_{Z_1}(Z_1 \mid f_i, f_{i'}, \mathcal{T}) = \\ & \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p_Z(Z_1, Z_2, Z_3, Z_4 \mid f_i, f_{i'}, \mathcal{T}) \, \mathrm{d}Z_2 \, \mathrm{d}Z_3 \, \mathrm{d}Z_4 \end{aligned} \tag{4.59}$$

which is a function of $X_i, Y_i, X_{i'}, Y_{i'}$. This gives the binary likelihood potential $V(d_2(i,i') \mid f_i, f_{i'}, \mathcal{T})$. The joint p.d.f. of the set of binary features, d_2, is approximated by the "pseudo-likelihood"

$$p(d_2 \mid f, \mathcal{T}) = \prod_{i \in \mathcal{S}} \prod_{i' \in \mathcal{N}_i} p(d_2(i,i') \mid f_i, f_{i'}, \mathcal{T}) \tag{4.60}$$

The joint p.d.f. of $d = \{d_1, d_2\}$ is approximated by

$$p(d \mid f, \mathcal{T}) = p(d_1 \mid f, \mathcal{T}) \; p(d_2 \mid f, \mathcal{T}) \tag{4.61}$$

Now the posterior energy can be obtained as

$$\begin{array}{rl} U(f, \mathcal{T} \mid d) = & \sum_{i \in \mathcal{S}} V_1(f_i, \mathcal{T}) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}, \mathcal{T}) + \\ & \sum_{i \in \mathcal{S}: f_i \neq 0} V_1(d_1(i) \mid f_i, \mathcal{T}) + \\ & \sum_{i \in \mathcal{S}: f_i \neq 0} \sum_{i' \in \mathcal{S}: f_{i'} \neq 0} V_2(d_2(i,i') \mid f_i, f_{i'}, \mathcal{T}) \end{array} \tag{4.62}$$

The optimal solution is $(f^*, \mathcal{T}^*) = \arg\min_{(f,\mathcal{T})} U(f, \mathcal{T} \mid d)$. The nonNULL labels in f^* represents the matching from the model object considered to the scene and $\mathcal{T}^*$ determines the pose transformation therein. The model points that are assigned the NULL label are either spurious or due to other model objects. Another round of matching-pose operations may be formed on these remaining points in terms of another model object.

4.4.3 Discussion

Minimizing the energies derived in this section is difficult. The dimensionality of the search space is high. As a guide to the search, the energies are inefficient unless some strong prior constraints are available, such as the normal prior distribution of poses assumed in (Wells 1991). However, the derived models may be useful for verifying the matching and pose estimation results. Assume that pose candidates are found by using techniques such as the Hough transform or geometric indexing. The energies may be used as global cost measures for the matching and pose.

4.5 Face Detection and Recognition

Face detection finds the face areas (usually rectangles) in an image, giving the locations and sizes of the faces detected. Consider a subwindow (data d) of an image at each location in the image plane and each scale. The basic detection problem is to classify the subwindow as face or nonface. This dichotomy could be done based on computing the ratio of likelihood densities of face and nonface followed by comparing the ratio with a confidence threshold (Dass et al. 2002).

Let $\mathcal{S}$ be the set of m pixel locations in the subwindow to be classified and let the label set $\mathcal{L}$ consist of admissible pixel intensities. Dass, Jain, and Lu (2002) proposed the following auto-model to model the likelihood

$$p(d \mid f) = \frac{\exp\left\{\sum_i \alpha_i f_i + \sum_i \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}\right\}}{\sum_{f_1} \sum_{f_2} \cdots \sum_{f_N} \exp\left\{\sum_i \alpha_i f_i + \sum_i \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}\right\}} \tag{4.63}$$

where α_i and $\beta_{i,i'}$ are the auto-model parameters. For computational tractability, the pseudo-likelihood (see Section 7.1.2) approximation is used instead of (4.63)

$$PL = \prod_{i=1}^{m} \frac{\exp\left\{\alpha_i f_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}\right\}}{\sum_{f_i} \exp\left\{\alpha_i f_i + \sum_{i' \in \mathcal{N}_i} \beta_{i,i'} f_i f_{i'}\right\}} \tag{4.64}$$

Two such auto-models could be used. The first assumes homogeneous correlations for all the sites. This is described by two pairwise parameters: $\beta_{i,i'} = \beta_h$ when i' is a horizontal neighbor of i and $\beta_{i,i'} = \beta_v$ when i' is a vertical neighbor with constants β_h and β_v.

The second model assumes inhomogeneous parameters $\beta_{i,i'}$ across sites i but isometric for different directions; that is, $\beta_{i,i'} = \beta_i$ is dependent on i only. Thus, the pseudo-likelihood is

$$\mathrm{PL} = \prod_{i=1}^{m} \frac{\exp\left\{\alpha_i f_i + \beta_i \sum_{i' \in \mathcal{N}i} f_i f_{i'}\right\}}{\sum_{f_i} \exp\left\{\alpha_i f_i + \beta_i \sum_{i' \in \mathcal{N}i} f_i f_{i'}\right\}} \tag{4.65}$$

where m pairwise parameters are needed.

The parameters in the PL could be estimated using the maximum pseudo-likelihood (MPL) on a training set. The detection decision of classifying a subwindow as face or nonface is based on the pseudo-likelihoods of faces and nonfaces. Hence, two sets of parameters need to be estimated, one from a training set of faces and the other from a training set of nonface subwindows (Dass et al. 2002).

In the detection stage, a subwindow d is classified into a face or nonface based on the log pseudo-likelihood ratio (LPR)

$$\mathrm{LPR} = \log \frac{PL_{face}}{PL_{nonface}} \tag{4.66}$$

The LPR is compared with a confidence value, and thereby a decision is made.

Works on MRF modeling for face recognition in the MAP-MRF framework have been reported in several papers, e.g., (Huang et al. 2004; Park et al. 2005). In (Huang et al. 2004), a face image is divided into m blocks represented by sites $d = \{d_1, \ldots, d_m\}$. The label set $L = \{1, \ldots, M\}$ corresponds to the M ID's. Assuming added Gaussian noise, the data term $p(d \mid f)$ is a Gaussian function. A pairwise "smoothness" term is imposed on pairs of labels as $P(f_i, f_{i'}) = \delta(f_i, f_{i'})$. In (Park et al. 2005), straight lines, corresponding to sites, are extracted from a face image. By attaching properties and binary relations to the straight lines, a face is then represented as an ARG. A partial matching is used to match two ARGs and select the best match.