Bayesian Object Matching

Stan Z. Li
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore 639798

Abstract

A Bayesian approach for object matching is presented. An object and a scene are each represented by its features, such as critical points, line segments and surface patches, constrained by unary properties and contextual relations. The matching is posed as a labeling problem where each feature in the scene is assigned (associated with) a feature of the known model objects. The prior distribution of a scene labeling is modeled as a Markov random field (MRF), which encodes the between-object constraints. The conditional distribution of the observed features given the labeling is assumed to be Gaussian, which encodes the within-object constraints. An optimal solution is defined as a maximum *a posteriori* (MAP) estimate. Relationships with previous work are discussed. Experimental results are shown.

1 Introduction

Object matching and recognition is a high level vision task. For non-simple objects, it is usually performed on object features, such as critical points, line segments and surface patches, rather than directly on image pixels as in low level vision. The features are constrained by their (unary) properties and mutual (contextual) relations. The features and the constraints can be compactly represented by a notion known in the computer vision community as "relational structure". Matching is to establish feature correspondences between the scene and the model objects to satisfy the constraints, or best satisfy the constraints when the matching is inexact. Object recognition is to identify the categories of objects in the scene, which is resolved immediately once the correspondences are found. This paper deals mainly with the matching problem.

We consider the model-based object matching in the following situations:

- There are two or more different objects in the model-base.
- Which ones of the modeled objects appear in the scene is unknown *a priori*.
- The objects appearing in the scene are transformed (rotated, translated and possibly scaled) versions of the corresponding models with unknown transformation parameters.
- The objects in the scene can be partially occluded and so some object features are hidden.
- The extracted quantitative feature properties and relations are subject to noise and processing errors.
- Spurious features may emerge due to mistakes made in feature extraction; they should be identified in the matching.
- Because the objects in the scene can occlude one another, features due to more than one object can be un-separated until matching is done, and the matching is performed with un-separated features.

Due to various uncertainties, exact and perfect solutions hardly exist; we usually look for some solution which optimally satisfies the constraints. In terms of statistics, we define the optimal solution to be the most probable one. *Maximum a posteriori* (MAP) probability is one of the most popular statistical criteria in many applications and in fact, has been the most popular choice in statistical image analysis. According to the Bayesian formulation, given the data, the posterior probability is proportional to the product of the prior and conditional probabilities.

Markov random field (MRF) theory provides a convenient and consistent way for modeling context dependent features. In image analysis, it has been successfully used for modeling interactions between image pixels and spatially correlated features. MRFs and the MAP criterion together give rise to the MAP-MRF framework used in this paper as well as in most other MRF-based applications. This framework, advocated by Geman and Geman (1984) and others, enables us to develop algorithms for a variety of vision problems systematically using rational principles rather than relying on *ad hoc* heuristics (Chellappa and Jain 1993; Mardia and Kanji 1993; Li 1995).

In this paper, we pose object matching as a labeling problem and formulate the problem in the MAP-MRF framework. The sites in an MRF correspond to features in the scene and the labels do to the features in a modeled object. We use MRFs to define the prior probability of the labels, and this characterizes the constraints on the features between different objects. When two features are due to the same object, the corresponding pair-site clique potential is zero; otherwise, it takes on a positive value (Li 1992a). This is a generalization of the line process (Geman and Geman 1984) used in low level vision to high level vision. The probability of the observed feature properties and relations, conditional on the MRF labels, are modeled as an independent Gaussian distribution whose (conditional) means are the feature properties and relations of the corresponding model object; this encodes the particular structure of the modeled object, *i.e.* the within-object constraints. The MAP solution is then defined based on the prior and the conditional probabilities.

Our matching method is based on a prerequisite that features have been extracted in a previous stage. This deviates from most papers in (Mardia and Kanji 1993) where recognition was performed directly on image data. Also, we assume that invariants are available for representing the features under the concerned group of transformations. This avoids the need to estimate object poses during matching.

Another related work is the Bayesian approach for object detection presented in (Baddeley and van Lieshout 1992). An interesting point in that work is the use of Markov object processes (Ripley and Kelley 1977; Baddeley and Moller 1989) for modeling the prior between-object constraints (within-object constraints are not used). Matching and recognition do not present much problem because only one type of simple object is considered, and the main purpose for using the prior constraints is to avoid multiple appearances of identical objects rather than to help matching and recognition.

The rest of the paper is organized as follows: Section 2 formulates the MAP-MRF matching problem. It includes three main parts: (i) the notion of labeling for the matching problem, (ii) the use of relational structure as a compact object representation embedding various constraints, and (iii) the definitions of probability functions for Bayesian matching. Relationships between this high level vision model and low level MAP-MRF models are discussed. Experimental results are presented in Section 3. This is followed by conclusions in Section 4.

2 MAP-MRF Object Matching

2.1 MAP-MRF Labeling

A natural representation in the study of MRFs is labeling (Besag 1974). Object matching, like many other image analysis problems, can be posed as a labeling problem. A labeling problem is specified in terms of a set of sites and a set of labels. Let

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

be a set of m discrete sites representing the features in the scene, and

$$\mathcal{L} = \{1, \ldots, M\} \tag{2}$$

be a set of M discrete labels representing the features in a model. Object matching is to assign a label from $\mathcal{L}$ to each site in $\mathcal{S}$.

Let $F = \{F_1, \ldots, F_m\}$ be a family of random variables, or a random field, defined on $\mathcal{S}$, in which each random variable F_i assumes a value in $\mathcal{L}$. A realization of F is called a configuration or a labeling, and is denoted by $f = \{f_1, \ldots, f_m\}$. In connection to the image matching problem, $f_i \in \mathcal{L}$ denotes the model feature matched to the image feature i. If all the m sites have the same label set $\mathcal{L}$, the admissible configuration space for f is the product of m $\mathcal{L}$'s

$$\mathbb{F} = \mathcal{L}^m \tag{3}$$

If it this is not true, special treatment is needed (see discussions in Section 2.5).

An MRF is always defined with respect to a neighborhood system. A neighborhood system for $\mathcal{S}$ is given by

$$\mathcal{N} = \{\mathcal{N}_i \mid \forall i \in \mathcal{S}\} \tag{4}$$

where $\mathcal{N}_i$ is the collection of sites neighboring i. Denote $f_{\mathcal{S}-\{i\}} = \{f_{i'} \mid i' \in \mathcal{S} - \{i\}\}$ and $f_{\mathcal{N}_i} = \{f_{i'} \mid i' \in \mathcal{N}_i\}$. A random field F is said to be an MRF on $\mathcal{S}$ with respect to $\mathcal{N}$ if the Markov condition

$$P(f_i \mid f_{\mathcal{S}-\{i\}}) = P(f_i \mid f_{\mathcal{N}_i}) \tag{5}$$

and the positivity condition

$$P(f) > 0 \qquad \forall f \in \mathbb{F} \tag{6}$$

are satisfied.

In the above, an MRF is defined according to the local conditional distributions. However, the global joint distribution $P(f)$ is often required for applications such as the MAP-MRF matching in this paper. Yet, defining the joint distribution from the conditional distributions is a difficult problem (Besag 1974). Fortunately, there is a MRF-GD (Gibbs distribution) theorem (Hammersley and Clifford 1971; Besag 1974; Geman and Geman 1984) which can be used for conveniently obtaining the joint distribution. The theorem states that F is an MRF on $\mathcal{S}$ with respect to $\mathcal{N}$ if and only if its joint distribution $P(F = f)$ is a Gibbs distribution with respect to $\mathcal{N}$. In other words, if F is locally an MRF, then globally its joint distribution is of Gibbs; and vice versa.

A GD of F with respect to $\mathcal{N}$ is of the following form

$$P(f) = Z^{-1} \times e^{-U(f)/T} \tag{7}$$

In the above, Z is a normalizing constant, T is a global control parameter called the temperature and $U(f)$ is the energy of the form

$$U(f) = \sum_{c \in \mathcal{C}} V_c(f) \tag{8}$$

where $\mathcal{C}$ is the set of "cliques" and $V_c(f)$ are the clique potential functions. In essence, a GD is featured by two things: it belongs to the exponential family; the energy is defined on cliques. A clique c for $(\mathcal{S}, \mathcal{N})$ is a subset of $\mathcal{S}$ consisting of either a single site or several neighboring sites. When cliques containing up to two sites are considered, the energy has the following form

$$U(f) = \sum_{\{i\} \in \mathcal{C}_1} V_1(f_i) + \sum_{\{i,i'\} \in \mathcal{C}_2} V_2(f_i, f_{i'}) \tag{9}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are the set of single-site and the set of pair-site cliques, respectively, and V_1 and V_2 are single- and pair-site potentials, respectively.

In the MAP-MRF framework, the optimal solution is the one which maximizes the posterior probability. The posterior probability is usually also a GD

$$P(F = f \mid d) = Z_E^{-1} \times e^{-E(f)} \tag{10}$$

where d is the observation and, $E(f) = \frac{1}{T}U(f) + U(d \mid f)$, called the posterior energy, is a sum of the prior energy $U(f)$ (the energy in the prior probability) and the likelihood energy $U(d \mid f)$ (the energy in the conditional probability). The MAP solution is equivalently found by minimizing the posterior energy

$$f^* = \arg\min_{f \in \mathbb{F}} E(f) \tag{11}$$

When $\mathcal{L}$ consists of (un-ordered) discrete labels as in our case (as opposed *e.g.* to continuous labels in a real interval), the minimization of the posterior energy is a combinatorial optimization problem. The interested reader is referred to (Li 1995) for a detailed exposure to labeling problems in image analysis, the MAP-MRF framework and minimization algorithms for MRFs.

2.2 Relational Structure Representation

As mentioned in the introduction, an object is represented by features constrained by feature properties and relations. The features, properties and relations can be denoted compactly as a *relational structure* (RS). An RS describes a scene or (part of) a model object. When both the scene and modeled objects are represented by RS's, the problem of object recognition is reduced to that of RS matching (Li 1992a; Li 1992b).

Let us start with a scene RS. Assume there are m features in the scene, indexed by a set $\mathcal{S} = \{1, \ldots, m\}$ of sites. The sites constitute the nodes of the RS. Each node $i \in \mathcal{S}$ is 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. Note that a property or a relation can be numerical (*e.g.* size of a region and distance between two points) or symbolic (*e.g.* qualitative color of a region and adjacency between two regions). More generally, there may be a vector $d_n(i_1, \ldots, i_n)$ of K_n n-ary relations among n features $i_1, \ldots, i_n \in \mathcal{S}$. When H is the highest order, all the relations are represented by $d = \{d_1, d_2, \ldots, d_H\}$. For $H = 2$, the RS is reduced to a *relational graph (RG)*.

Second or higher order relations are required to convey contextual or inter- constraints. Under these constraints, features due to an object are 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 complexity and computational cost.

The RS for the scene is defined by a triple

$$\mathcal{G} = (\mathcal{S}, \mathcal{N}, d) \tag{12}$$

The neighborhood system $\mathcal{N}$ for $\mathcal{S}$ determines the scope of the relational dependence. Only the sites belonging to the same clique can have direct influence to each other, where the degree of the influence is also subject to the clique potential function. However, indirect influence can be propagated to other sites.

The RS for a model object is similarly defined as

$$\mathcal{G}' = (\mathcal{L}, \mathcal{N}', D) \tag{13}$$

where $D = \{D_1, D_2, \ldots, D_H\}$ has the equivalent meaning to that of $d = \{d_1, d_2, \ldots, d_H\}$. In this case, the relations in D is for the labels in $\mathcal{L}$, *i.e.* for the model features. In addition, the "neighborhood system" for $\mathcal{L}$ is defined as consisting of all the other labels, that is,

$$\mathcal{N}'_I = \{I' \mid \forall I' \in \mathcal{L}, I' \neq I\} \tag{14}$$

This means each model feature is related to all the other model features. The highest considered order, H, in $\mathcal{G}'$ is equal to that in $\mathcal{G}$. For particular n and k ($1 \leq k \leq K_n$; $1 \leq n \leq H$), $D_n^{(k)}$ represents the same type of constraint as $d_n^{(k)}$; for example, both represent the angle between line segments. Note that only the same type of constraints are comparable (*e.g.* an angle in the scene compared to an angle in a model but not to a distance).

There can be more than one model RS 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 RS's. In this case, the RS's may be related by some inter-RS constraints but this is not further discussed in this paper.

To deal with spurious features, we introduce a virtual model composed of a single node $\mathcal{L}_0 = \{0\}$ and called the NULL model. This special model represents everything not modeled by $\mathcal{G}'$, this including features due to all other model objects as well as noise. So the actual label set for the matching contain $M + 1$ labels

$$\mathcal{L}^+ = \{0, 1, \ldots, M\} \tag{15}$$

Without confusion, we still use the notation $\mathcal{L}$ to denote the above augmented label set unless there is a necessity to differentiate.

Model-based matching is then finding a mapping from the image RS to the model RS (or vice versa). Such a mapping from one RS to another is called a *morphism* (Li 1992b), written as

$$f : \mathcal{G}(\mathcal{S}, \mathcal{N}, d) \rightarrow \mathcal{G}'(\mathcal{L}, \mathcal{N}', D) \tag{16}$$

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

$$f : \mathcal{S} \rightarrow \mathcal{L} \tag{17}$$

Note that only the same type of features can be matched to each other (*e.g.* a point matched to a point not a line; see Section 2.5). Thus the morphism maps relations d_n to relations D_n

$$f : d_n \rightarrow D_n \tag{18}$$

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

When some of the properties and relations are quantitative, 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 (Blake and Zisserman 1987).

The goodness of fit for 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* RS's because 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 RS's have different numbers of nodes and numerical differences are allowed for the matched properties and relations, the matching is more difficult because it cannot exploit the advantage of the strong constraint of the exact equalities. In the following, we derive the posterior distribution as the objective function for weak homomorphisms for inexact, partial matching under weak constraints.

2.3 Defining Probability Distributions

The prior distribution $P(f)$ is defined based on MRFs. In object matching, one may restrict the scope of interaction by defining $\mathcal{N}_i$ as the set of the other features which are within a distance r

from i

$$\mathcal{N}_i = \{i' \in \mathcal{S} \mid [\text{dist}(\text{feature}_{i'}, \text{feature}_i)]^2 \leq r,\ i' \neq i\} \tag{19}$$

In the above, the function "dist" is a suitably defined function for distances between features. It is tricky as how to define a distance between non-point features such as lines. In our experiments, we simply choose the Euclidean distance between midpoints as the distance between lines. The distance threshold r may be reasonably related to the size of the model object currently under consideration. In matching to more than one model object, each object is considered in turn and hence r may differ from object to object. The set of first order cliques is

$$\mathcal{C}_1 = \{\{i\} \mid i \in \mathcal{S}\} \tag{20}$$

The set of second order cliques is

$$\mathcal{C}_2 = \{\{i, i'\} \mid i' \in \mathcal{N}_i, i \in \mathcal{S}\} \tag{21}$$

In our formulation, only cliques of up to order two are considered.

The single-site potential is defined as

$$V_1(f_i) = \begin{cases} v_{10} & \text{if } f_i = 0 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where $v_{10} > 0$ is a constant. If f_i is the NULL label, it incurs a penalty of v_{10}, or nil penalty otherwise. The pair-site potential is defined as

$$V_2(f_i, f_{i'}) = \begin{cases} v_{20} & \text{if } f_i = 0 \text{ or } \ f_{i'} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where $v_{20} > 0$ is a constant. If either f_i or $f_{i'}$ is the NULL , it incurs a penalty of v_{20}, or nil penalty otherwise; this characterizes interaction between NULL and non-NULL labels. One may set $V_2(f_i, f_{i'})$ to a negative value, for example $-v_{20}$, when both f_i and $f_{i'}$ are nonzero, encouraging the concurrence of features from the same object. However, setting it to zero does not affect the energy minimum because only relative values are important in the minimization. The prior energy is then defined as

$$U(f) = \sum_{i \in \mathcal{S}} V_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \tag{24}$$

where $\sum_{i \in \mathcal{S}}$ is another form for $\sum_{c \in \mathcal{C}_1}$ and $\sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i}$ another form for $\sum_{c \in \mathcal{C}_2}$ (where $\{i, j\}$ and $\{j, i\}$ are considered as distinct pair-site cliques).

A few remarks on the above pair-site clique potential function $V_2(f_i, f_{i'})$ follow:

1. It characterizes the interaction between NULL and non-NULL labels and this obviously deals with spurious features.

2. It also characterizes the between-object interactions because features not due to the currently considered model object are all labeled as the NULL . The between-object interactions are modeled via modeling the interactions between features belonging to different objects. This contrasts to Markov object processes based works for object detection (Ripley 1981; Baddeley and van Lieshout 1992) where between-object interactions are modeled directly between objects rather than via features.

3. It also imposes within-object constraints in that it associates label pairs belonging to the same object a lower cost than label pairs belonging to different objects, But this is imposes the constraints in a weak way. Within-object constraints are imposed more strongly via the conditional distribution to be defined below.

The conditional p.d.f. of the observed data d, $p(d \mid f)$, 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 non-NULL 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 (i) that D and d are composed of types of features which are invariant under the considered class of transformations; (ii) that they are related via the observation model

$$d_1(i) = D_1(f_i) + e_1(i) \tag{25}$$

$$d_2(i,i') = D_2(f_i, f_{i'}) + e_2(i,i') \tag{26}$$

where e is additive independent zero mean Gaussian noise. The assumptions of the independent Gaussian noise may not be accurate but offers an approximation when an accurate observation model is not available. Then the likelihood function is a Gibbs distribution with the energy

$$U(d \mid f) = \sum_{i \in \mathcal{S}, f_i \neq 0} V_1(d_1(i) \mid f_i) + \sum_{i \in \mathcal{S}, f_i \neq 0} \; \sum_{i' \in \mathcal{S} - \{i\}, f_{i'} \neq 0} V_2(d_2(i,i') \mid f_i, f_{i'}) \tag{27}$$

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

$$V_1(d_1(i) \mid f_i) = \begin{cases} \sum_{k=1}^{K_1} [d_1^{(k)}(i) - D_1^{(k)}(f_i)]^2 / \{2[\sigma_1^{(k)}]^2\} & \text{if } f_i \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

and

$$V_2(d_2(i,i') \mid f_i, f_{i'}) = \begin{cases} \sum_{k=1}^{K_2} [d_2^{(k)}(i,i') - D_2^{(k)}(f_i, f_{i'})]^2 / \{2[\sigma_2^{(k)}]^2\} & \\ & \text{if } i' \neq i \& f_i \neq 0 \ \& \ f_{i'} \neq 0 \\ & \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

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

$$\begin{aligned} U(f \mid d) = \quad & \textstyle\sum_{i \in \mathcal{S}} V_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) + \\ & \textstyle\sum_{i \in \mathcal{S}: f_i \neq 0} V_1(d_1(i) \mid f_i) + \sum_{i \in \mathcal{S}: f_i \neq 0} \ \sum_{i' \in \mathcal{S} - \{i\}: f_{i'} \neq 0} V_2(d_2(i, i') \mid f_i, f_{i'}) \end{aligned} \tag{30}$$

As such, the formulation makes use of contextual constraints on the labeling *a priori* and on the observed data.

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 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. A method for learning $[\sigma_n^{(k)}]^2$ parameters from examples is presented in (Li 1997).

Normally, symbolic or qualitative relations are represented internally by a number. One may set the corresponding variances $[\sigma_n^{(k)}]^2$ 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 compared symbolic relations are not the same. This inhibits symbolically incompatible matches, and thus imposes the desired symbolic constraint.

2.4 Matching to Multiple Types of Objects

The MAP configuration f^* derived in the above is the optimal mapping from the scene to the considered model object. In other words, it is the optimal labeling of the scene in terms of that particular model object. Suppose that 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 where

$$f^{(\ell)} = \{f_1^{(\ell)}, \ldots, f_m^{(\ell)}\} \tag{31}$$

is the MAP labeling for matching the scene to model object ℓ. In these labelings, a feature in the scene may have been matched to more than one model feature in different objects. However, any feature in the scene should have only one match of model feature. In the following, a scheme is presented which uses cost minimization to resolve this ambiguity.

Rewrite $E(f) = U(f \mid d)$ in (30) into the following form

$$E(f) = \sum_{i \in \mathcal{S}} E_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{S}, i' \neq i} E_2(f_i, f_{i'}) \stackrel{\triangle}{=} \sum_{i \in \mathcal{S}} E(f_i \mid f_{\mathcal{N}_i}) \tag{32}$$

where

$$E_1(f_i) = \begin{cases} v_{10} & \text{if } f_i = 0 \\ V_1(d_1(i) \mid f_i) & \text{otherwise} \end{cases} \tag{33}$$

and

$$E_2(f_i, f_{i'}) = \begin{cases} v_{20} & \text{if } i' \in \mathcal{N}_i \ \& \ (f_i = 0 \text{ or } \ f_{i'} = 0) \\ V_2(d_2(i, i') \mid f_i, f_{i'}) & \text{if } i' \neq i \ \& \ f_i \neq 0 \ \& \ f_{i'} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

are local posterior energy of order one and two, respectively, and

$$E(f_i \mid f_{\mathcal{N}_i}) = E_1(f_i) + \sum_{i' \in \mathcal{S}, i' \neq i} E_2(f_i, f_{i'}) \tag{35}$$

is the cost of the local match $i \to f_i$ given the rest of the matches. $E(f_i^{(\ell)} \mid f_{\mathcal{N}_i}^{(\ell)})$ $(\ell = 1, \ldots, L)$ will be used as the basis for selecting the best matched objects for i in matching to multiple model objects.

The rule is the following: The image feature i is considered to come from object ℓ_i if the decision incurs the least local cost

$$\ell_i = \arg \min_{\ell \in \{1, \ldots, L\}} E(f_i^{(\ell)} \mid f_{\mathcal{N}_i}^{(\ell)}) \tag{36}$$

As the result, the final label for i is feature number $f_i^{(\ell_i)}$ of object ℓ_i. Applying (36) to every $i \in \mathcal{S}$ yields an overall mapping from the scene to the models. The overall mapping is composed of several sub-mappings where sub-mapping maps from a part of the scene features to some model features in each matched object model.

A few remarks follow: First, the simple rule (36) is not one in the MAP sense. Second, although (36) has an appearance similar to that of the iterative conditional mode (ICM), it is not the ICM, because ℓ is the model number rather than a label (model feature) and the selection is performed once rather than iteratively.

We could perform the optimal matching to all the L objects in one stage (Li 1992c), instead of matching to each object one by one and then resolving the inconsistency between individual matching results. A possible scheme is to let the label set $\mathcal{L}$ comprise all the features of all the objects, and then treat the labeling problem as assigning a label f_i from such a label set $\mathcal{L}$ to each $i \in \mathcal{S}$. In this scheme, the matching and recognition are performed simultaneously. A major concern is the size of the configuration space which is exponentially large. Assuming that every object feature is admissible for every image feature, then the overall label set is

$$\mathcal{L} = \mathcal{L}^{(1)} \bigcup \mathcal{L}^{(2)} \bigcup \cdots \bigcup \mathcal{L}^{(L)} \tag{37}$$

where $L^{(\ell)}$ is the label set for object ℓ. Assuming that there are $M^{(\alpha)}$ features in object α, then the size of $\mathcal{L}$ is $\#\mathcal{L} = 1 + M^{(1)} + M^{(2)} + \cdots + M^{(L)}$ (the NULL label is included). The corresponding configuration space is $\mathbb{F} = \mathcal{L}^m$. Its size is $(1 + M^{(1)} + M^{(2)} + \cdots + M^{(L)})^m$ which is much larger tha the szie $(1+M^{(1)})^m + (1+M^{(2)})^m + \cdots + (1+M^{(L)})^m$ for the one by one scheme. This would require much more computational effort and be more prone to the local minimum problem in finding the MAP matching. Therefore, the one by one scheme is adopted in almost all existing object recognition works, though it usually requires a resolution of the inconsistency as discussed in the above.

2.5 Coupled MRFs for Matching of Different Types of Features

Consider the situation where an object consists of different types of features such as points, lines and surface patches. Obviously, one type of features in the scene should be matched to the same type of features in the model, for example, a point matched to a point. In this case, the label set $\mathcal{L}$ for a site (image feature) $i \in \mathcal{S}$ should comprise only the same type of labels (model features); for example, if i is a line feature, the label set for i should comprise only the line features not points and patches. Recall that the configuration space $\mathbb{F}$ defined as the simple product of Eq.(15) was for a single MRF. From mathematical viewpoint, if one type of feature should be matched to only the same type of features, the positivity condition (6) will not hold for every $f \in \mathbb{F}$ when there are different types of features.

In general, each site $i \in \mathcal{S}$ can have its own set of labels $\mathcal{L}_i^+$ where $\mathcal{L}_i^+ \subseteq \mathcal{L}^+$ contains (i) all the model features which are of the same type as feature i and (ii) the "wildcard" NULL compatible to all types of features. Hence, the configuration space is redefined as

$$\mathbb{F} = \mathcal{L}_1^+ \times \mathcal{L}_2^+ \times \cdots \times \mathcal{L}_m^+ \tag{38}$$

Redefining $\mathbb{F}$ make the positivity condition (6) be satisfied when there are different types of features. Note that not all label subsets $\mathcal{L}_i$ are distinct to each other. If there are $K(\leq m)$ distinct $\mathcal{L}_i$, then there should be K distinct subsets. In this case, $\mathbb{F}$ is partitioned into K coupled MRFs.

The MRFs are coupled to each other via relations $d_n(i_1, \ldots, i_n)$ $(n \geq 2)$. For example, the distance relation between a point and a line can constrain the two different types of features. Furthermore, the MRFs are coupled also via the "wildcard" NULL .

Despite this treatment, 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.

2.6 Relationships with Low Level MRF Models

Let us compare the present model with low level vision MRF models prototyped by (Geman and Geman 1984). In image restoration and surface reconstruction involving discontinuities (Geman and Geman 1984; Marroquin 1985; Blake and Zisserman 1987), there are commonly two coupled MRFs, 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, non-edge}. 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 present model is the breakage of relational bond in the scene RS. For example, when none of f_i and $f_{i'}$ 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. Therefore, we say that our treatment of the interaction between NULL and non-NULL labels is similar to the line process model introduced by (Geman and Geman 1984).

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 observation only, such as pixel intensity; although intensity difference between neighboring pixels is also used, it is derived directly from the unary observation. The present model uses relational measurements of any orders. This is important for high level problems in which contextual constraints play a stronger role. Moreover, in the present model, the neighborhood system is non-homogeneous and anisotropic, as opposed to the image case where pixels are located in equi-spaced sites.

3 Experiment Results

In the following, experimental results are presented for matching objects of point patterns, line patterns and jigsaw objects. Given a number of model objects are given, a scene is composed of a combination of the modeled, and possibly some un-modeled, objects. Feature properties and relations for the model objects and for the scene are measured using the same program. In defining the posterior energy, only binary relations are considered while unary constraints are ignored. This equivalent to setting $v_{10} = 0$ and $V_1(d_1(i) \mid f_i) = 0$. The reason for setting v_{10} is that its influence on the result is not significant as opposed to that of v_{20}; the reason for setting $V_1(d_1(i) \mid f_i) = 0$ is that in our case, unary constraints we can think about are either variant, or not reliable enough (hence the corresponding σ_n^k are very large). Parameters $v_{20} = 0.7$ is fixed for all the experiments. Parameters $[\sigma_2^{(k)}]^2$ vary for different applications.

In the actual computation, the pair-site clique potentials in the posterior distribution, $V_2(f_i, f_{i'} \mid d_2(i,i')) = V_2(f_i, f_{i'}) + V_2(d_2(i,i') \mid f_i, f_{i'})$, are rescaled (multiplied by 255) and quantized into the integers in $[0, \ldots, 255]$ to reduce storage and good results are still 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 energy is minimized by using a continuous relaxation labeling algorithm (Hummel and Zucker 1983). In a separate study, we have shown that this algorithm yields good quality solutions (Li, Wang, Chan, and Petrou 1997). It is much better than ICM (Besag 1986), quite comparable with simulated annealing (SA) (Kirkpatrick, Gellatt, and Vecchi 1983; Geman and Geman 1984), yet requires much less computational effort than SA, thus balancing between the quality and the cost.

3.1 Matching Objects of Point Patterns

Please refer to Fig.1. There are three model objects as shown in (a) – (c) of the figure. 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: (i) Take a subset of features from each of the three objects, (ii) do a transformation (rotation and translation) on each of the subsets, (iii) mix the subsets together after that, (iv) randomly deviate the locations of the points using either Gaussian or uniform noise, and (v) 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 binary relation, *i.e.* the Euclidean distance between two points $d_2(i, i') = dist(i, i')$. The points, their sizes and the 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 parameters is chosen to be $[\sigma_2^{(1)}]^2 = 0.1$.

Fig.2 shows the matching results in which the matched object points are aligned with the scene in (d). The black points in (a) correspond to points 5, 6, 7, 11, 12, 14, 15, 17 and 19 of the object in Fig.1(a). Those in (b) correspond to points 3, 5, 8, 9, 10, 11, 12, 13, 16, 17 and 18 of the object in Fig.1(b). Those in (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.1(c). In (d) is shown the union of all the individual results. The spurious points 2, 8, 10, 27, 35 and 40 in Fig.1(d) have found no counterparts in the three objects. They are correctly classified as the NULL . There is one mis-match: point 19 in Fig.1(d) is matched to the NULL while its correct home should be point 10 of Fig.1(a).

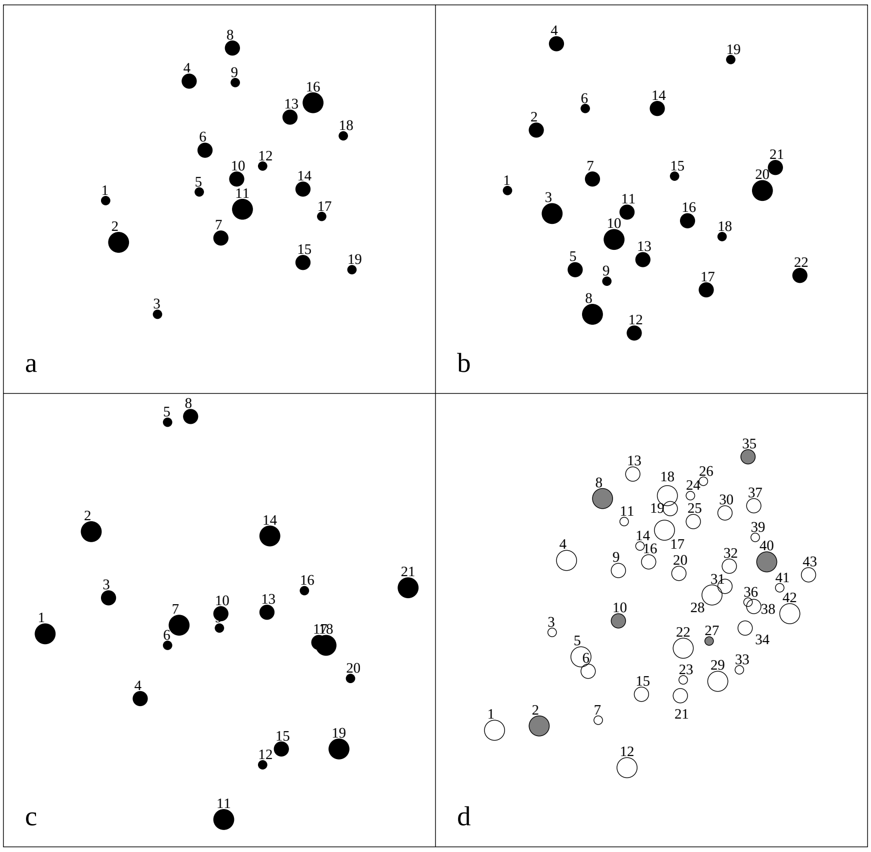

Figure 1: Matching objects of points. (a-c) The model objects. (d) The scene.

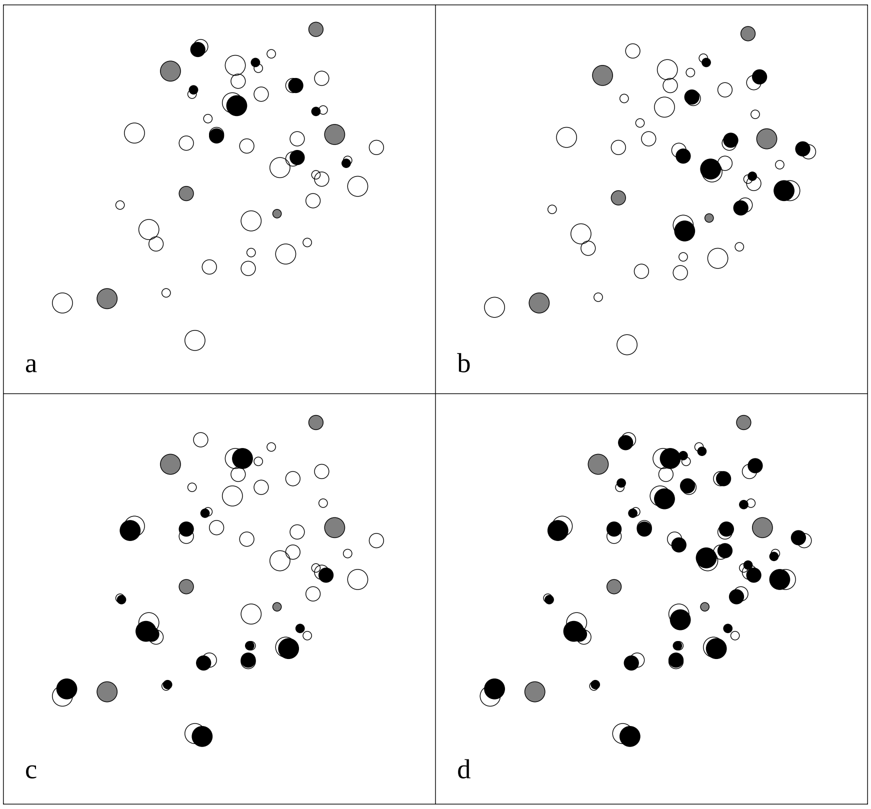

Figure 2: Results of matching objects of points. (a-c) Matched points (in black) from the respective models aligned with the scene. (d) All matched points aligned with the scene.

3.2 Matching Objects of Line Patterns

Please refer to Fig.3. There are five objects made of lines, as shown in (a) – (e). The scene in (f) consists of subset of deviated lines taken from the first three objects (a) – (c), shown in dotted lines, and spurious line features shown in 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 mid-points of the lines,

(3) $d_2^{(3)}(i, i')$: the minimum distance between the end-points of the lines, and

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

No unary relations are used. The value for the prior clique potential is $v_{20} = 0.70000$ fixed. 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 (Li 1997).

Fig.4 shows the matching results in which the matched object lines from the first three objects are aligned with the scene in (d). The solid lines in (a) correspond to lines 11, 13, 14, 16, 17, 18, 19, 21, 25, 26 and 28 of the object in Fig.3(a). Those in (b) correspond to lines 5, 8, 10, 11, 13, 14, 15, 16 and 17 of the object in Fig.3(b). Those in (c) correspond to points 1, 2, 3, 5, 8, 9, 11, 12 and 14 of the object in Fig.3(c). Objects in (d) – (e) do not have matches. In (d) is shown the union of all individual results. The spurious lines 2, 13, 14, 15, 27, 33 and 36 in Fig.3(f) have found no counterparts in the object models. They are correctly classified as the NULL .

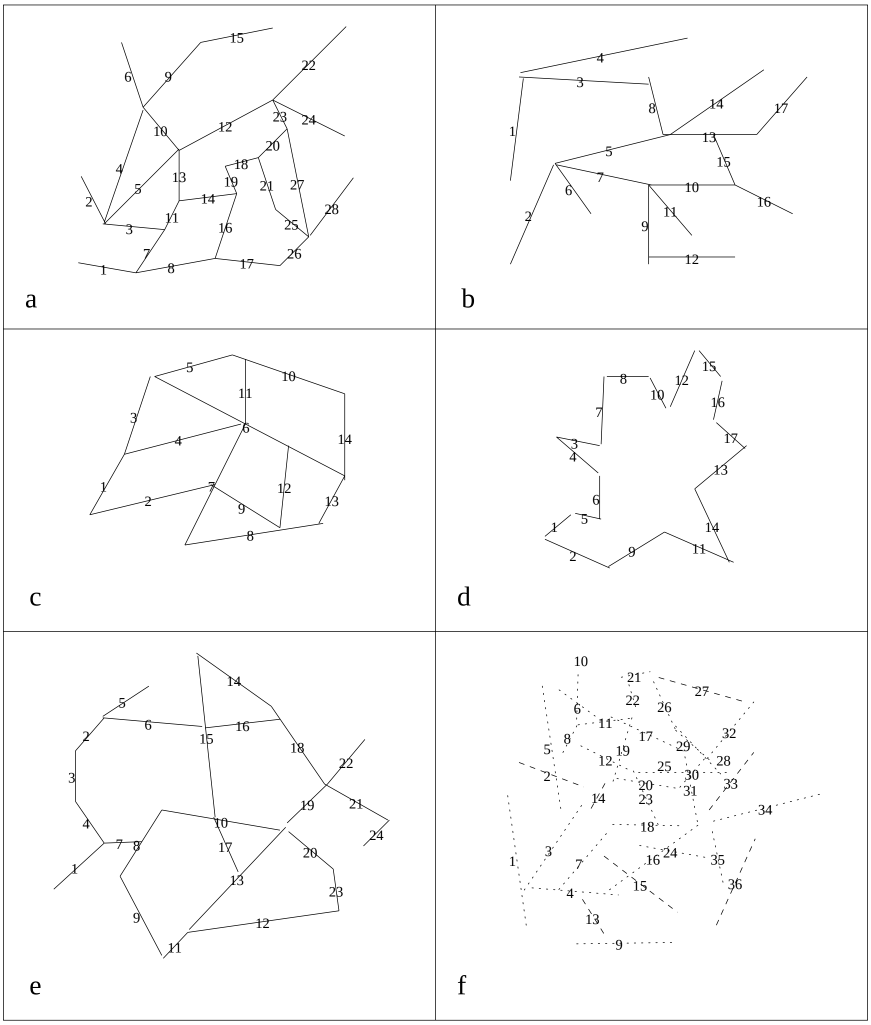

Figure 3: Matching objects of lines. (a-e) The five model objects. (f) The scene.

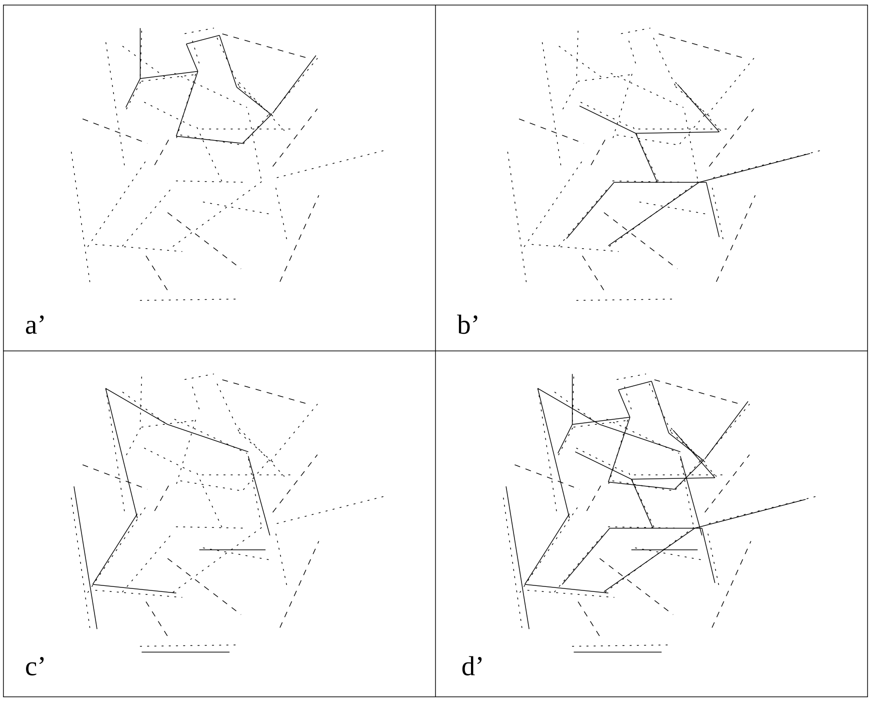

Figure 4: Results of matching objects of lines. (a-c) Matched lines (in solid) from the respective models aligned with the scene. (d) All matched lines aligned with the scene.

3.3 Matching Curved Objects under Similarity Transformations

Please refer to Figure 5. There are eight model jigsaw objects in this experiment. 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 considered model curve. The neighbors of a site (corner point) are defined as the 5 forward and the 5 backward corner points from the considered corner point alone the curve. 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 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 involved parameters are supplied by an automated optimal estimation procedure (Li 1997) with the following 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 final result of matching and recognition is shown in Figure 6 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 extrusions and intrusions. This means ambiguities exist extensively and have to be resolved by using context.

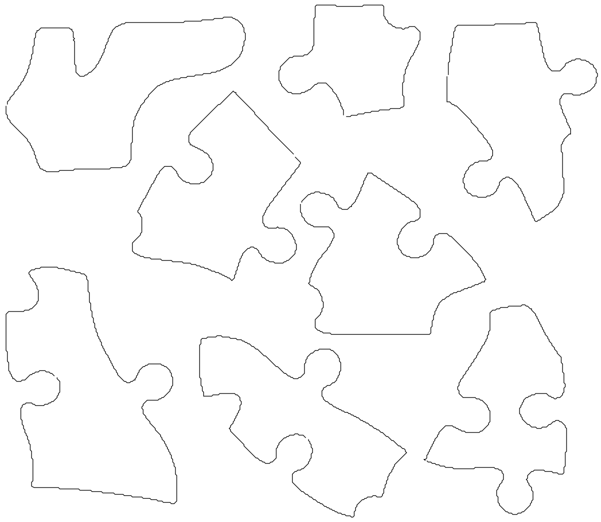

Figure 5: (Top) The eight model jigsaw objects. (Bottom) An overlapping scene.

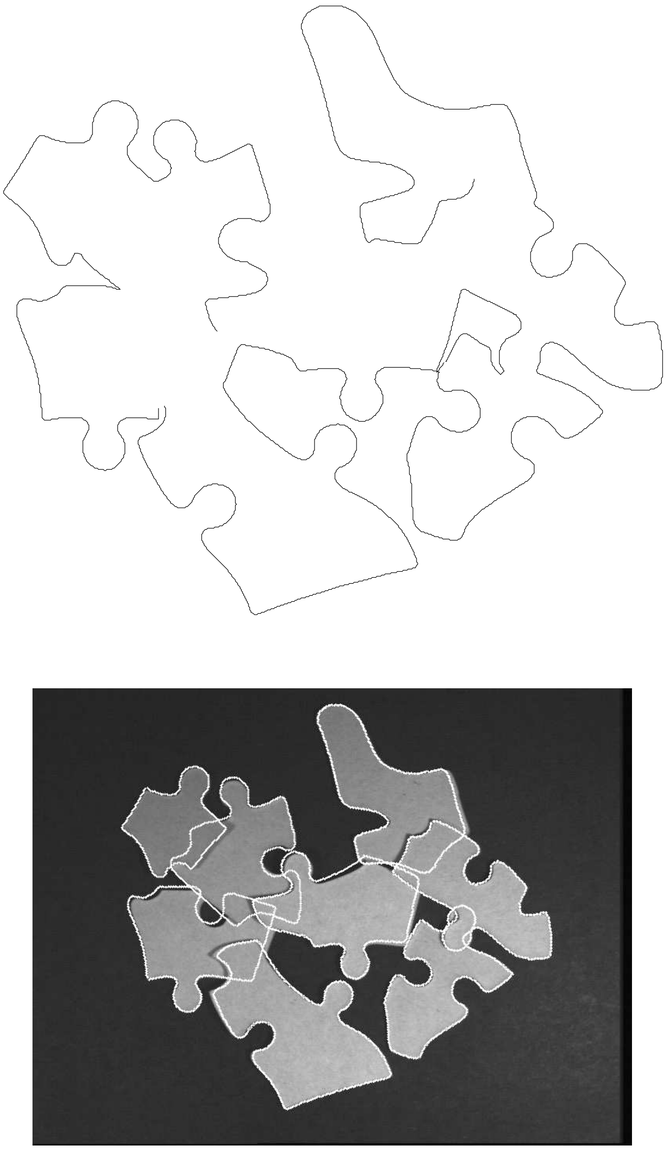

Figure 6: (Top) Boundaries detected from the scene. (Bottom) Matched objects are aligned with the scene.

4 Conclusions

In this paper, we have presented an MAP-MRF approach for model-based object matching in high level vision. An object or a scene is represented by features constrained one another by relations. Matching is posed as a labeling problem in which each site (an image feature) is assigned a a label (a model feature). MRFs are used to define the prior distribution of the labels. The conditional distribution of the observed features given the MRF labels encodes the particular structure of each model object. The MAP solution to the matching problem is defined based on the two distributions.

In the present model, the between-object constraints are modeled by the MRF prior distribution. This is done via the interactions between NULL and non-NULL labels in the prior. Although the prior distribution also imposes within-object constraints, it is done in a weak manner. Currently, the within-object constraints are mainly imposed by the conditional distribution of properties and relations. How to use MRFs to encode within-model constraints into the prior distribution in a more efficient way is a topic in a feature research.

References

Baddeley, A. J. and J. Moller (1989). "Nearest-neighbour Markov processes and random sets". *International Statistical Review 57*, 89–121.

Baddeley, A. J. and M. N. M. van Lieshout (1992). "Object recognition using Markov spatial processes". In *Proceedings of International Conference Pattern Recognition*, Volume B, pp. 136–139.

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

Besag, J. (1986). "On the statistical analysis of dirty pictures" (with discussions). *Journal of the Royal Statistical Society, Series B 48*, 259–302.

Blake, A. and A. Zisserman (1987). *Visual Reconstruction.* Cambridge, MA: MIT Press.

Chellappa, R. and A. Jain (Eds.) (1993). *Markov Random Fields: Theory and Applications.* Academic Press.

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

Hammersley, J. M. and P. Clifford (1971). "Markov field on finite graphs and lattices". unpublished.

Hummel, R. A. and S. W. Zucker (1983). "On the foundations of relaxation labeling process". *IEEE Transactions on Pattern Analysis and Machine Intelligence 5*(3), 267–286.

Kirkpatrick, S., C. D. Gellatt, and M. P. Vecchi (1983). "Optimization by simulated annealing". *Science 220*, 671–680.

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

Li, S. Z. (1992b). "Matching: invariant to translations, rotations and scale changes". *Pattern Recognition 25*(6), 583–594.

Li, S. Z. (1992c). "Object recognition from range data prior to segmentation". *Image and Vision Computing 10*(8), 566–576.

Li, S. Z. (1993). "Similarity invariants for 3D space curve matching". In *Proceedings of the First Asian Conference on Computer Vision*, Osaka, Japan, pp. 454–457.

Li, S. Z. (1995). *Markov Random Field Modeling in Computer Vision.* New York: Springer-Verlag.

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

Li, S. Z., H. Wang, K. L. Chan, and M. Petrou (1997). "Minimization of MRF energy with relaxation labeling". *Journal of Mathematical Imaging and Vision 7*, 149–161.

Mardia, K. V. and G. K. Kanji (Eds.) (1993). *Statistics and Images: 1.* Advances in Applied Statistics. Carfax.

Marroquin, J. L. (1985). *Probabilistic Solution of Inverse Problems.* Ph. D. thesis, MIT AI Lab.

Ripley, B. D. (1981). *Spatial Statistics.* New York: Wiley.

Ripley, B. D. and F. P. Kelley (1977). "Markov point process". *Journal of London Mathematical Society 15*, 188–192.

Professor Kanti Mardia

Editor, Journal of Applied Statistics

Department of Applied Statistics

University of Leeds

Leeds LS2 9JT

United Kingdom