A Two-Stage Probabilistic Approach for Object Recognition

Stan Z. Li[1] and Joachim Hornegger[2]

[1] Nanyang Technological University
School of EEE, Nanyang Avenue, Singapore 639798
szli@szli.eee.ntu.ac.sg
http://markov.eee.ntu.ac.sg:8000/~szli/
[2] Stanford University, Robotics Laboratory
Gates Building 134, Stanford, CA 94305-9010, USA
jh@Robotics.Stanford.EDU

Abstract. Assume that some objects are present in an image but can be seen only partially and are overlapping each other. To recognize the objects, we have to firstly separate the objects from one another, and then match them against the modeled objects using partial observation. This paper presents a probabilistic approach for solving this problem. Firstly, the task is formulated as a two-stage optimal estimation process. The first stage, *matching*, separates different objects and finds feature correspondences between the scene and each potential model object. The second stage, *recognition*, resolves inconsistencies among the results of matching to different objects and identifies object categories. Both the matching and recognition are formulated in terms of the maximum *a posteriori* (MAP) principle. Secondly, contextual constraints, which play an important role in solving the problem, are incorporated in the probabilistic formulation. Specifically, between-object constraints are encoded in the prior distribution modeled as a Markov random field, and within-object constraints are encoded in the likelihood distribution modeled as a Gaussian. They are combined into the posterior distribution which defines the MAP solution. Experimental results are presented for matching and recognizing jigsaw objects under partial occlusion, rotation, translation and scaling.

1 Introduction

Model-based *object recognition* is a high level vision task which identifies the category of each object in the scene with reference to the model objects. There are two broad types of approaches: templet-based and feature-based. In the templet-based approach, an object is represented by a templet which may be in the form of its bitmap or the entire outline; the observation is matched to the templet based on some distance measure. This approach has been used in numerous applications such as character recognition and face recognition in which objects are well observed in the image.

Currently, the templet-based approach does not seem to offer a proper solution to the partial matching problem. A basic assumption in the templet-based approach is that the object can be observed entirely. The assumption is invalidated when objects are only partially observable due to mutual occlusions. The templet-based approach is not inherently ready to handle this situation: While it allows local deformations, it is unable to perform with missing parts. This is because of its lack of the ability to represent an object by local features.

The feature-based approach is complementary to the templet-based approach in this regard. Here, an object model is represented by local object features, such as points, line segments or regions, subject to various constraints [3,8,6,17]. *Object Matching* is performed to establish correspondences between local features in the scene (image) and those in each model object. Because it is based on the local features of an object rather than the global information of it, it is more appropriate to handle the partialness and ovelappingness, and hence provides an alternative for object recognition. This paper is aimed to investigate a formal mathematical framework for object recognition using partial observation.

From a pattern recognition viewpoint, an object is considered as a pattern of mutually or contextually constrained features. The use of contextual constraints is essential in the interpretation of visual patterns. An feature itself makes little sense when considered independently of the rest. It should be interpreted in relation to other image features in the spatial and visual context. At a higher level, a scene is interpreted based on various contextual constraints between features.

An interesting situation is when a scene contains many, possibly mutually occluded, objects, which is the case dealt with in this paper. In this situation, both the following two sources of contextual constraints are required to resolve ambiguities in the model-based matching and recognition, in our opinion: *between-object constraints* (BOCs) and *within-object constraints* (WOCs). The particular structure of an object itself is described by the WOCs of the object. Such constraints are used to identify an instance of that object in the scene. The BOCs, which describe constraints on features belonging to different objects, are used to differentiate different objects in the scene. In a sense, within-object constraints are used for evaluating similarities whereas between-object constraints are for evaluating dissimilarities. An interpretation is achieved based on the two types of constraints.

In matching and recognition, as in other image analysis tasks, exact and perfect solutions hardly exist due to various uncertainties such as occlusion and unknown transformations from model objects to the scene. Therefore, we usually look for some solution which optimally satisfies the considered constraints. A paradigm is prediction-verification [1]. It is able to solve the matching problem efficiently. In terms of statistics, we may define the optimal solution to be the most probable one. The maximum *a posteriori* (MAP) principle is a statistical criteria used in many applications and in fact has been the most popular choice in statistical image analysis.

Markov random field (MRF) theory provides a convenient and consistent way for modeling image features under contextual constraints [4,14,12], also for object recognition [15,5,12,9]. MRFs and MAP together give rise to the MAP-

MRF framework. 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.

Scene-model matching is generally performed by considering one model object at a time, and when there are multiple model objects, multiple matching results are generated. Because the matching to each model is done independently of the other models, inconsistencies can exist among the results of matching to the different objects, and must be resolved to obtain consistent and unambiguous solutions. Our formulation of a two-stage estimation offers a solution in terms of the MAP principle.

In this paper, we develop a statistically optimal formulation for object matching and recognition of a scene containing multiple overlapping objects. Matching and recognition are posed as labeling problems and are performed in two consecutive stages, each solving an MAP-MRF estimation problem. The first stage matches the features extracted from the scene against those of each model object by maximizing the posterior distribution of the labeling. This finds feature correspondences between the scene to the model objects, and separates overlapping objects from one other. It produces multiple MAP matching results, each for one model object. Inconsistencies in these results are resolved by the second estimation stage, MAP recognition. In this latter stage, the MAP matching results produced by the previous stage are examined as a whole, inconsistencies among them are resolved, and all the objects are identified unambiguously finally.

The contextual constraints are imposed in probability terms. The BOCs are encoded in the prior distribution modeled as a Markov random field (MRF). This differentiates between different objects and between an object and the background. In a way, this is similar to the line-process model [7] for differentiating edge and non-edge elements. The WOCs are encoded in the likelihood distribution modeled as a Gaussian. It compares the similarity between a model object and its corresponding part in the scene. The BOCs and the WOCs are combined into the posterior distribution. An optimal solution, either for matching or for recognition is defined as the most probable configuration in the MAP sense.

The rest of the paper is organized as follows: In Section 2, the optimal solutions for matching and recognition are formulated, which illustrates how to use probabilistic tools to incorporate various contextual constraints and how to resolve ambiguities arising from matching to individual model objects. Experiments are presented in Section 3 for matching and recognition of a scene containing multiple free-form jigsaw objects under rotations, translations, scale changes and occlusions.

2 Two Stage MAP-MRF Estimation

Object matching and recognition, like many other image analysis problems, can be posed as labeling problems. Let $\mathcal{S} = \{1, \ldots, m\}$ be a set of m sites corresponding to the features in the scene, and $\mathcal{L} = \{1, \ldots, M\}$ be a set of M labels corresponding to the features in a model object. What types of features to use to represent an object is a problem not addressed in this paper. We assume some features have been chosen which present invariance in some feature properties and relations. An example of representation is given in the experiments section for curved objects like jigsaw, which can be referred to now by the unfamiliar reader.

In addition to the M labels in $\mathcal{L}$, we introduce a virtual label, called the NULL and numbered 0. It represents everything not in the above label set $\mathcal{L}$, including features due to un-modeled objects as well as noise. By this, the label set is augmented into $\mathcal{L}^+ = \{0, 1, \ldots, M\}$. Labeling is to assign a label from $\mathcal{L}^+$ to each site in $\mathcal{S}$. Without confusion, we still use the notation $\mathcal{L}$ to denote the augmented label set unless there is a necessity to differentiate. A labeling configuration, denoted by $f = \{f_1, \cdots, f_m\}$, a mapping from the set of sites to the set of labels, *i.e.* $f : \mathcal{S} \to \mathcal{L}$, in which $f_i \in \mathcal{L}$ is the object feature matched to the image feature i. When there are more than one object, a label represents not only an object feature but also the object category.

Given the observed data d, we define the optimal labeling f^* to be the one which maximizes the posterior. The posterior is a Gibbs distribution $P(F = f \mid d) \propto \mathrm{e}^{-E(f)}$ with the posterior energy

$$E(f) \triangleq U(f \mid d) = U(f) + U(d \mid f) \tag{1}$$

The energy is a sum of the prior energy $U(f)$ (the energy in the prior distribution) and the likelihood energy $U(d \mid f)$ (the energy in the distribution of d). Hence, the MAP solution is equivalently found by minimizing the posterior energy $f^* = \arg\min_{f \in \mathbb{F}} E(f)$. An MAP estimation is performed in each of the two stages.

2.1 Stage 1: MAP Matching

This stage performs MAP matching to each model object by minimizing the energy $E(f)$ of a posterior distribution in which the prior is modeled by Markov random fields (MRFs) and the likelihood by Gaussian.

The prior is modeled as an MRF which is a Gibbs distribution $P(f) = Z^{-1} \times \mathrm{e}^{-U(f)}$ where Z is the normalizing constant. The energy $U(f)$ is of the form $U(f) = \sum_{c \in \mathcal{C}} V_c(f)$ where $\mathcal{C}$ is the set of "cliques" for a neighborhood system $\mathcal{N}$ and $V_c(f)$ are the clique potential functions. In object matching, one may restrict the scope of interaction by defining the neighborhood set of i as the set of the other features which are within a distance r from feature i, $\mathcal{N}_i = \{i' \in \mathcal{S} \mid [\text{dist}(\text{feature}_{i'}, \text{feature}_i)]^2 \leq r,\ i' \neq i\}$. The function "dist" is a suitably defined function for the distance between features. For point features, it can be chosen as the Euclidean distance between two points. It is tricky as how to define a distance between non-point features; *e.g.* for straight lines, a simple definition would be the distance between the midpoints of two straight lines. The distance threshold r may be chosen reasonably to be the maximum diameter of the model object currently under consideration.

The prior energy $U(f)$ is of the form $U(f) = \sum_{c \in \mathcal{C}} V_c(f)$ where $\mathcal{C}$ is the set of "cliques" and $V_c(f)$ are the clique potential functions. In essence, a Gibbs distribution is featured by two things: it belongs to the exponential family and its energy is defined on clique potentials. When cliques containing up to two sites are considered, the energy has the following form

$$U(f) = \sum_{\{i\} \in \mathcal{C}_1} V_1(f_i) + \sum_{\{i,i'\} \in \mathcal{C}_2} V_2(f_i, f_{i'}) = \sum_{i \in \mathcal{S}} V_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}) \quad (2)$$

where $\mathcal{C}_1 = \{\{i\} \mid i \in \mathcal{S}\}$ and $\mathcal{C}_2 = \{\{i,i'\} \mid i' \in \mathcal{N}_i, i \in \mathcal{S}\}$ are the sets of single- and pair-site cliques, respectively, and V_1 and V_2 are single- and pair-site potential functions. In defining $\mathcal{C}_2$, we assume that $\{a, b\}$ is an ordered set and so $\{i, i'\} \neq \{i', i\}$. The clique potentials are defined as

$$V_1(f_i) = \begin{cases} 0 & \text{if } f_i \neq 0 \\ v_{10} & \text{if } f_i = 0 \end{cases}, \quad V_2(f_i, f_{i'}) = \begin{cases} 0 & \text{if } f_i \neq 0 \text{ and } f_{i'} \neq 0 \\ v_{20} & \text{if } f_i = 0 \text{ or } f_{i'} = 0 \end{cases} \quad (3)$$

where $v_{10} > 0$ and $v_{20} > 0$ are penalty constants for NULL labels.

The pair-site clique potentials $V_2(f_i, f_{i'})$ encode between-object constraints by treating the two situations differently: (i) when both features are due to the considered object ($f_i \neq 0$ and $f_{i'} \neq 0$), and (ii) when one of the features is due to the background or another object ($f_i = 0$ or $f_{i'} = 0$). This differentiates between the considered model object and another object, and between the considered model object and the background. A dissimilarity between the classes of the two features is thus evaluated. The potentials associate label pairs belonging to the considered object (more closely related) with a lower cost, and associates label pairs belonging to different objects (less closely related) with a higher cost. Therefore, the use of the properties of the pairwise interactions plays a crucial role in separating overlapping objects.

The likelihood distribution $p(d \mid f)$ describes the statistical properties of model features seen in the scene and is therefore conditioned on pure non-NULL matches $f_i \neq 0$. It depends on how the visible features are observed, and this in turn depends on the underlying transformations and noise, which is regardless of the neighborhood system $\mathcal{N}$. Denote D_1 for unary properties and D_2 for binary relations between the features of a model object. Assume (i) that the truth $D = \{D_1, D_2\}$ of the model features and the data d are composed of types of features which are invariant under the considered class of transformations (their selections are application-specific); (ii) that they are related via the observation models $d_1(i) = D_1(f_i) + e_1(i)$ and $d_2(i,i') = D_2(f_i, f_{i'}) + e_2(i,i')$ where e is additive independent zero mean Gaussian noise.[1] Then the likelihood function is a Gibbs distribution with the energy

$$U(d \mid f) = \sum_{i \in \mathcal{S}, f_i \neq 0} V_1(d_1(i) \mid f_i) + \sum_{i \in \mathcal{S}, f_i \neq 0} \sum_{i' \in \mathcal{S}_{\setminus i}, f_{i'} \neq 0} V_2(d_2(i,i') \mid f_i, f_{i'}) \quad (4)$$

[1] The assumptions of the independent Gaussian noise may not be accurate but offers an approximation when an accurate observation model is not available.

where $\mathcal{S}_{\setminus i} \triangleq \mathcal{S}-\{i\}$, and the summations are restricted to the non-NULL matches $f_i \neq 0$ and $f_{i'} \neq 0$. The likelihood potentials are

$$V_1(d_1(i) \mid f_i) = \sum_{k=1}^{K_1}[d_1^{(k)}(i) - D_1^{(k)}(f_i)]^2/\{2[\sigma_1^{(k)}]^2\} \tag{5}$$

and

$$V_2(d_2(i,i') \mid f_i, f_{i'}) = \sum_{k=1}^{K_2}[d_2^{(k)}(i,i') - D_2^{(k)}(f_i, f_{i'})]^2/\{2[\sigma_2^{(k)}]^2\} \tag{6}$$

where the vectors $D_1(f_i)$ and $D_2(f_i, f_{i'})$ are the "conditional mean" (conditioned on f) for the random vectors $d_1(i)$ and $d_2(i,i')$, respectively; K_1 and K_2 are the numbers unary properties and binary relations; $[\sigma_n^{(k)}]^2$ ($k = 1, \ldots, K_n$ and $n = 1, 2$) are the variances of the corresponding noise components.

The likelihood potentials are defined for image features belonging only to the model object under consideration ($f_i \neq 0$ and $f_{i'} \neq 0$). Therefore they encode the within-object constraints. They are used to evaluate the similarity between the model object and its corresponding part in the scene.

The constraints on both the labeling *a priori* and the observed data are incorporated into the posterior distribution with the posterior energy

$$\begin{aligned} U(f \mid d) = &\textstyle\sum_{i\in\mathcal{S}} V_1(f_i) + \sum_{i\in\mathcal{S}} \sum_{i'\in\mathcal{N}_i} V_2(f_i, f_{i'}) + \\ &\textstyle\sum_{i\in\mathcal{S}:f_i\neq 0} V_1(d_1(i) \mid f_i) + \\ &\textstyle\sum_{i\in\mathcal{S}:f_i\neq 0} \sum_{i'\in\mathcal{S}_{\setminus i}:f_{i'}\neq 0} V_2(d_2(i,i') \mid f_i, f_{i'}) \end{aligned} \tag{7}$$

Hence, the between-object constraints and the within-object constraints are combined into the posterior energy. The MAP matching is the configuration which minimizes $U(f \mid d)$.

One model object is considered at a time. Minimizing $U(f \mid d)$ for a model object results in a mapping from $\mathcal{S}$ to $\mathcal{L}^+$ for that object. The result tells us two things: (i) (separation) image features belonging (the "in-subset") and not belonging to the considered object, and (ii) (matching) correspondences between the features in the "in-subset" and the features of the model object.

The parameters MRF v_{10} and v_{20} and the likelihood parameter $\sigma^{(k)}$ have to be determined in order to completely define the MAP solution. This is done by using a supervised learning algorithm [12].

The present model can be compared to the coupled MRF model of [7] in that there are two coupled MRFs, one for line processes (edges) and one for intensities; and a line process variable can be on or off depending on the difference between the two neighboring intensities. The concept of "line process" in the present model is the relational bond between features in the scene. When $f_i \neq 0$ and $f_{i'} \neq 0$, i and i' are relationally constrained to each other; otherwise when $f_i = 0$ or $f_{i'} = 0$, the relational bond between i and i' is broken. The differences are: the present model makes use of relational measurements of any orders because contextual constraints play a stronger role in high level problems, whereas the model in [7] uses only unary observation. Moreover, in the present model, the

$i =$	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	$7^{(5)}$	$6^{(5)}$	$5^{(5)}$	$4^{(5)}$	$3^{(5)}$
$f^{(all)}$	0	0	0	$5^{(4)}$	$4^{(4)}$	$3^{(4)}$	$2^{(4)}$	$7^{(5)}$	$6^{(5)}$	$5^{(5)}$	$4^{(5)}$	$3^{(5)}$

Table 1. Matching and recognition of an image containing $m = 12$ features to $L = 5$ objects.

neighborhood system is non-homogeneous and anisotropic, as opposed to the image case in which pixels are equi-spaced.

The MAP matching of the scene is performed to each potential object one by one. In this case, the complexity of the search is linear in the number of models. Some fast screening heuristics may be imposed to quickly rule out unlikely models, but this is discussed in this paper. We concentrate on the MAP formalism.

2.2 Stage 2: MAP Recognition

After matching to each potential object one by one, we obtain a number of MAP solutions. However, inconsistencies may exist among them: Assuming that there are L potential model objects, we have L MAP solutions, $f^{(1)}, \ldots, f^{(L)}$ where $f^{(\alpha)} = \{f_1^{(\alpha)}, \ldots, f_m^{(\alpha)}\}$ is the MAP labeling f^* for matching the scene to model object $\alpha \in \{1, \ldots, L\}$ obtained in stage 1, and $f_i^{(\alpha)}$ denotes feature number $I = f_i$ of object α. Since each $f^{(\alpha)}$ is the optimal labeling of the scene in terms only of model object α but not of the other objects, inconsistencies may exist among the L results in the sense below.

A feature $i \in S$ in the scene may have been matched to more than one model feature in different objects; that is, there may exist more than one $\alpha \in \{1, \ldots, L\}$ for which $f_i^{(\alpha)} \neq 0$. Table 1 illustrates an example of results for matching an image with $m = 12$ features to $L = 5$ model objects, where $f^{(\alpha)}$ $(\alpha = 1, \ldots, 5)$ are the MAP solutions for matching to the five objects. For example, image feature $i = 8$ has been matched to $10^{(2)}$ (feature No. 10 of object 2), $1^{(4)}$ and $7^{(5)}$. However, any feature in the scene should be matched to at most one non-NULL model feature; that is, for a specific i, there should be that either $f_i^{(\alpha)} = 0$ for all α or $f_i^{(\alpha)} \neq 0$ for just one α value. When this is not the case, the inconsistencies should be resolved in order to unambiguously identify the object category to which each image feature uniquely belongs. A possible consistent final result is given as $f^{(all)}$ in the table.

The recognition stage is to make the matching results consistent and to identify the categories of objects in the scene. Again, this stage is also formulated as an MAP estimation. Denote object α as $\mathcal{O}^{(\alpha)}$. The posterior derived previously for matching to $\mathcal{O}^{(\alpha)}$ can be explicitly expressed as $P(f \mid d, \mathcal{O}^{(\alpha)})$. Denoting the posterior probability for matching to all the L objects as $P(f \mid d, \mathcal{O}^{(all)})$ where $\mathcal{O}^{(all)}$ is short for $\mathcal{O}^{(1)}, \cdots, \mathcal{O}^{(L)}$, the MAP recognition is then defined as $f^* = \arg\max_{f \in \mathbb{F}^{(all)}} P(f \mid d, \mathcal{O}^{(all)})$. The configuration space $\mathbb{F}^{(all)}$ consists of $(1 + \sum_{\alpha=1}^{L} M^{(\alpha)})^m$ elements, where $M^{(\alpha)}$ is the number of labels in model α, when all the labels in all the models are admissible.

The posterior, $P(f \mid d, \mathcal{O}^{(all)}) \propto P(f \mid \mathcal{O}^{(all)}) p(d \mid f, \mathcal{O}^{(all)})$, is a Gibbs distribution because of the Markov property of the labels. Similar to that in the matching stage, the prior energy is

$$U(f \mid \mathcal{O}^{(all)}) = \sum_{i \in \mathcal{S}} V_1(f_i \mid \mathcal{O}^{(all)}) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'} \mid \mathcal{O}^{(all)}) \tag{8}$$

and the likelihood energy is

$$\begin{aligned} U(d \mid f, \mathcal{O}^{(all)}) = & \sum_{i \in \mathcal{S}, f_i \neq 0} V_1(d_1(i) \mid f_i, \mathcal{O}^{(all)}) + \\ & \sum_{i \in \mathcal{S}, f_i \neq 0} \; \sum_{i' \in \mathcal{S}_{\setminus i}, f_{i'} \neq 0} V_2(d_2(i, i') \mid f_i, f_{i'}, \mathcal{O}^{(all)}) \end{aligned} \tag{9}$$

The single-site potential are defined as $V_1(f_i^{(\alpha)} \mid \mathcal{O}^{(all)}) = V_1(f_i^{(\alpha)} \mid \mathcal{O}^{(\alpha)})$ which is the same as that in (3) for matching to a single model object α. The pair-site potential are defined as

$$V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha')} \mid \mathcal{O}^{(all)}) = \begin{cases} V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha')} \mid \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')}) & \text{if } \alpha = \alpha' \\ v_{20} & \text{otherwise} \end{cases} . \tag{10}$$

where $V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha')} \mid \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')}) = V_2(f_i, f_{i'})$ is the same as that in (3). The above definitions are a straightforward extension of (3): In (3), features due to other objects (as opposed to the one currently under consideration) are all labeled as NULL ; (10) simply takes this principle into consideration for recognizing multiple objects. Using (3), we obtain

$$V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha')} \mid \mathcal{O}^{(all)}) = \begin{cases} 0 & \text{if } (\alpha = \alpha') \text{ and } (f_i^{(\alpha)} \neq 0) \text{ and } (f_{i'}^{(\alpha')} \neq 0) \\ v_{20} & \text{otherwise} \end{cases} \tag{11}$$

The single-site likelihood potentials are $V_1(d_1(i) \mid f_i^{(\alpha)}, \mathcal{O}^{(all)}) =$ $V_1(d_1(i) \mid f_i^{(\alpha)}, \mathcal{O}^{(\alpha)})$ which is the same as (5). The pair-site likelihood potentials are

$$V_2(d_2(i, i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha')}, \mathcal{O}^{(all)}) = \begin{cases} V_2(d_2(i, i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha')}, \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')}) \\ \qquad\qquad\qquad\qquad \text{if } \alpha = \alpha' \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{12}$$

where $V_2(d_2(i,i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha')}, \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')}) = V_2(d_2(i,i') \mid f_i, f_{i'})$ is the same as (6). The posterior energy is obtained as $U(f \mid d, \mathcal{O}^{(all)}) = U(f \mid \mathcal{O}^{(all)}) + U(d \mid f, \mathcal{O}^{(all)})$.

When L MAP matching solutions are available, the configuration space $\mathbb{F}^{(all)}$ can be reduced to a great extent. Let $\mathcal{S}' \subset \mathcal{S}$ be the set of sites which were previously matched to more than one non-NULL label, $\mathcal{S}' = \{i \in \mathcal{S} \mid f_i^{(\alpha)} \neq 0$ for more than one different $\alpha\}$. For the case of Table 1, $\mathcal{S}' = \{8, 9, 10, 11\}$. Only those labels in $i \in \mathcal{S}'$ are subject to changes in the recognition stage. Therefore, the configuration space can be reduced to $\mathbb{F}^{(all)} = \mathcal{L}_1^{(all)} \times \mathcal{L}_2^{(all)} \times \cdots \times \mathcal{L}_m^{(all)}$ where $\mathcal{L}_i^{(all)}$ is constructed in the following way:

- For $i \in \mathcal{S}'$, $\mathcal{L}_i^{(all)}$ consists of all the non-NULL labels previously assigned to i, $\{f_i^{(\alpha)*} \neq 0 \mid \alpha = 1, \ldots, L\}$, plus the NULL label;
  e.g. $\mathcal{L}_8^{(all)} = \{0, 10^{(2)}, 1^{(4)}, 7^{(5)}\}$ for Table 1.
- For $i \notin \mathcal{S}'$, $\mathcal{L}_i^{(all)}$ consists of the non-NULL label if there is a non-NULL label, or $\mathcal{L}_i^{(all)} = \{0\}$ otherwise; *e.g.* $\mathcal{L}_6^{(all)} = \{3^{(4)}\}$, and $\mathcal{L}_3^{(all)} = \{0\}$.

Each involved object α contributes one or zero label to $\mathcal{L}_i^{(all)}$, as opposed to $M^{(\alpha)}$ labels before the reduction, and therefore the size of $\mathcal{L}_i^{(all)}$ is at most $L+1$ as opposed to $1 + \sum_{\alpha=1}^{L} M^{(\alpha)}$ before. The size of $\mathcal{L}_i^{(all)}$ is one for $i \notin \mathcal{S}'$. Therefore, the MAP recognition is thus reduced to the following: (i) It is performed over the reduced configuration space $\mathbb{F}_{\mathcal{S}'}^{(all)} = \prod_{i \in \mathcal{S}'} \mathcal{L}_i^{(all)}$; (ii) it is to maximize 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, and $f_{\mathcal{S}-\mathcal{S}'} = \{f_i \mid i \in \mathcal{S} - \mathcal{S}'\}$ is the set of labels which are fixed during the maximization. It is equivalently to minimize the conditional posterior energy $U(f_{\mathcal{S}'} \mid d, f_{\mathcal{S}-\mathcal{S}'}, \mathcal{O}^{(all)})$.

After the reduction, only one or just a small number of labels remain admissible for each site and the search space becomes very small. For example, for the case of Table 1, the reduced label sets of size larger than one are $\mathcal{L}_8^{(all)} = \{0, 10^{(2)}, 1^{(4)}, 7^{(5)}\}$, $\mathcal{L}_9^{(all)} = \{0, 9^{(2)}, 6^{(5)}\}$, $\mathcal{L}_{10}^{(all)} = \{0, 7^{(2)}, 3^{(3)}, 5^{(5)}\}$, $\mathcal{L}_9^{(all)} = \{0, 4^{(3)}, 4^{(5)}\}$, and the sizes of $\mathcal{L}_i^{(all)}$ are one for $i \notin \mathcal{S}'$; the previous size of $\sum_{\alpha=1}^{5} (M^{(\alpha)} + 1)^{12}$ configurations (say, $M^{(\alpha)} = 10$) is then reduced to $4 \times 3 \times 4 \times 3 = 144$, so small that an exhaustive search may be plausible.

2.3 Minimization Methods

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 information. Many algorithms are available for this [12]. The ICM algorithm [2] iteratively maximizes local conditional distributions in a way as a "greedy method". Global optimizers such as simulated annealing (SA) [11,7] also iterate based on local energy changes. Relaxation labeling algorithms [10, 16] provide yet another choice. It is desirable to find globally good solution with a reasonable cost. A comparative study [13] shows that the Hummel-Zucker relaxation labeling algorithm [10] is preferable in terms of the minimized energy value and computational costs. Therefore, the Hummel-Zucker algorithm is used for computing the MAP solutions in our experiments. As the result of unambiguous relaxation labeling, there is a unique f_i for any i in the scene while the global energy reaches a local minimum.

The computational time is dominated by relaxation labeling, in the first stage, which matches the scene to each model, and hence so is the complexity of the system. The Hummel-Zucker relaxation labeling algorithm converges after dozens of iterations.

3 Experimental Results

The following experiment demonstrates the use of the present approach for the MAP object matching and recognition of partially observed and overlapping objects. There are 8 model jigsaw objects in the model-base which can be seen later in the results. All the models share the common structure of round extrusions and intrusions, and such ambiguities can cause difficulties in matching and recognition. In a scene, the objects are rotated, translated, scaled, partially occluded and overlapping, as shown in Fig.1. Boundaries are computed from the image using the Canny detector followed by hysteresis thresholding and edge linking, which results in three broken edge sequences. After that, corners of the boundaries, which are defined as curvature extrema and tangent discontinuities, are located, and used as feature points (Fig.1). Some model feature points are missing and the boundary of the object in the center is broken into two pieces. The sites in $\mathcal{S}$ correspond to the corners on a image curve and the labels in $\mathcal{L}$ correspond to such feature points on a model curve.

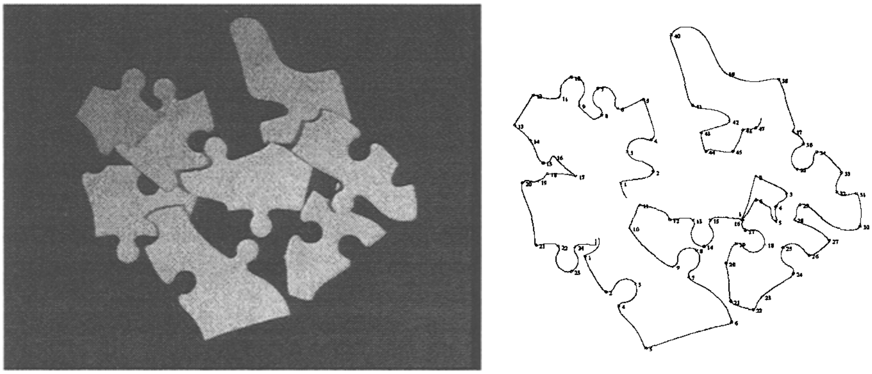

Fig. 1. A jigsaw image (left), the detected boundaries and corners (right).

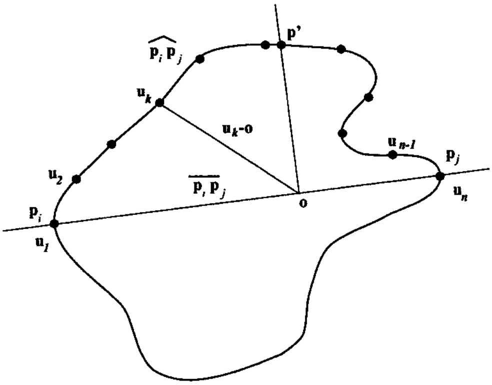

Fig. 2. Deriving similarity invariants for each curve segment bounded by two feature points p_i and p_j.

3.1 Invariant Features

A similarity invariant representation is used to encode constraints on the feature points ($d_1(i)$) and on the curve segments between the feature points ($d_2(i,j)$). The invariant unary property $d_1(i)$ is chosen to be the sign of the curvature $\kappa(p_i)$ for corner i. Similarity invariant relations $d_2(i,j)$ are derived to describe the curve segment between the pair of corner i and j, as follows: Consider the curve segment $\widehat{p_ip_j}$ between p_i and p_j and the straight line $\overline{p_ip_j}$ that passes through the points, as illustrated in Fig.2. The ratio of the arc-length $\widehat{p_ip_j}$ and the chord-length $\overline{p_ip_j}$: $d_2^{(1)}(i,j) = \frac{\text{arclength}(\widehat{p_ip_j})}{\text{chordlength}(\overline{p_ip_j})}$ is an invariant scalar. The ratio of curvature at p_i and p_j: $d_2^{(2)}(i,j) = \frac{\kappa(p_i)}{\kappa(p_j)}$ is also an invariant scalar. Two n-position-vectors of invariants are derived to utilize the constraints on the curve segment: First, find the mid-point, denoted by p', of $\widehat{p_ip_j}$ such that curve segments $\widehat{p_ip'}$ and $\widehat{p'p_j}$ have the equal arc-length. Next, find the point, denoted by o, on $\overline{p_ip_j}$ such that line $\overline{op'}$ is perpendicular to $\overline{op_j}$. Both p' and o are unique for $\widehat{p_ip_j}$. Then sample the curve segment at the n equally spaced (in arc-length) points $u_1, \ldots, u_n$. This is equivalent to inserting $n-2$ points between p_i and p_j. The vector of normalized radii is defined as $d_2^{(3)}(i,j) = [r_k]_{k=1}^n$ where $r_k = \frac{\|u_k - o\|}{\|op'\|}$ is similarity invariant. The vector of angles is defined as $d_2^{(4)}(i,j) = [\theta_k]_{k=1}^n$ where $\theta_k = \angle u_k o p_i$ is also similarity invariant. Now, $d_2(i,j)$ consists of four types ($K_2 = 4$) of $2n+2$ similarity invariant scalars.

3.2 Matching and Recognition

In the matching stage, an image curve is matched against each of the eight model jigsaws. Fig.3 shows the solutions of matching one of the image boundary curves (in solid) to each of the eight model jigsaws, *i.e.* the MAP estimates

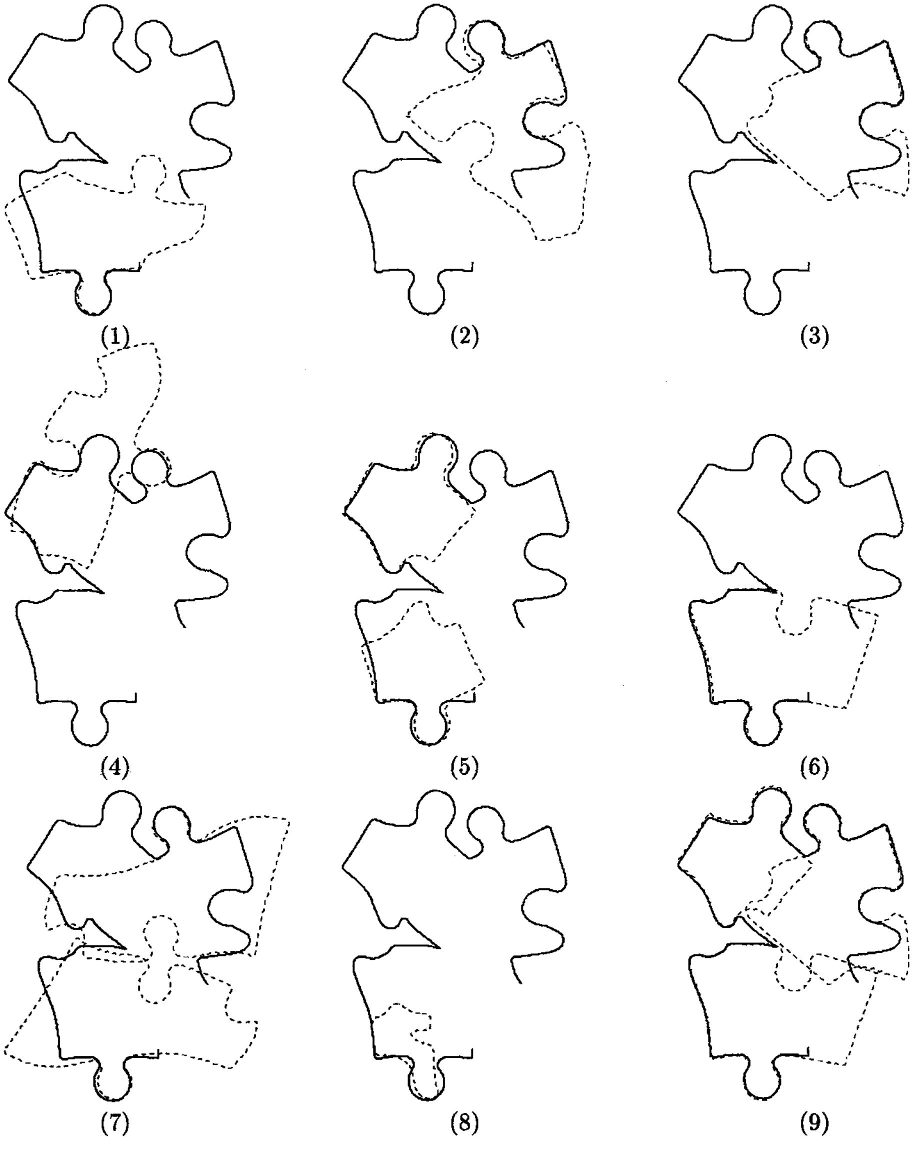

Fig. 3. Results from matching and recognition stages. (1) MAP solution $f^{(1)*}$ for matching the boundary curve (in solid) to model jigsaw No.1 (in dashed). The overlay of the model jigsaw on the input boundary curve indicates the correspondence. (2)–(8) MAP solutions for matching the boundary curve to models Nos.2–8. (9) The final MAP recognition result where the three recognized model jigsaws are overlayed on the input boundary curve.

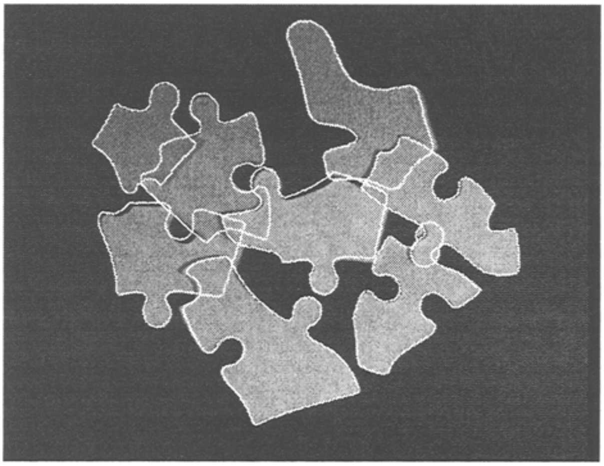

Fig. 4. The overall matching and recognition result.

$f^{(\alpha)^*} = \max_f P(f \mid d, \mathcal{O}^{(\alpha)})$ for $\alpha = 1, \ldots, 8$, where each model jigsaw (in dashed) is aligned to the curve. The overlapping portion indicates the correspondences whereas the image corners in the non-overlapping portion of the scene are assigned the NULL label.

The matching stage does two things: (i) It classifies the corners in the scene into two groups, non-NULL and NULL , or in other words, those belonging to the considered object and those not. (ii) For the non-NULL group, it gives the corresponding model corners. Therefore, the matching stage not only finds the feature correspondences between the scene and the considered model, but also does the separation of feature belonging to the considered object from those not.

Despite the ambiguities caused by the common structure of round extrusions and intrusions, the MAP matching has successfully distinguished the right model using information about the other part of the model. Also, it allows multiple instances of any model object, as in $f^{(5)^*}$ and $f^{(7)^*}$ of Fig.3 where each contains two instances of a model.

Although each MAP matching result is reasonable by itself, it may be inconsistent with others. For example, $f^{(2)^*}$, $f^{(3)^*}$ and $f^{(7)^*}$ in Fig.3 compete for a common part. This is mostly due to the common structures mentioned above. The inconsistencies have to be resolved. The MAP recognition stage identifies the best model class for each corner in the scene. The final recognition result is shown in the lower-right corner of Fig.3. Fig.4 shows the overall result for matching and recognizing the three boundary curves in the scene to all the models.

There are a number of parameters involved in the definition of the MAP solutions. Parameters $v_{20} = 0.7$ is fixed for all the experiments. Parameters $[\sigma_2^{(k)}]^2$ in the likelihood are estimated by using a supervised learning procedure [12]. The estimated values are $1/\sigma_2^{(1)} = 0.00025$, $1/\sigma_2^{(2)} = 0$, $1/\sigma_2^{(3)} = 0.02240$

and $1/\sigma_2^{(4)} = 0.21060$ for the likelihood. For the unary properties, we set $v_{10} = 0$ for the prior and $V_1(d_1(i) \mid f_i) = 0$ for the likelihood. The reason for setting $v_{10} = 0$ is that the influence of the single-site prior on the result is insignificant as compared to the pair-site one. The reason for $V_1(d_1(i) \mid f_i) = 0$ (in other words, we set $\sigma_1^k = \infty$) is because we are unable to compute unary constraints which are both invariant and reliable.

4 Conclusions

A two stage MAP estimation approach has been presented for solving the problems of model-based object separation, feature correspondence and object recognition using partial and overlapping observation. Contextual constraints are considered important for solving the problem. The particular structure of an object itself is described by the within-object constraints of the object. Such constraints are used to identify an instance of that object in the scene. Overlapping objects are separated by using the between-object constraints which differentiate between features belonging to an object and those not belonging to. The MAP estimate problem is formulated by taking both types of contextual constraints into consideration. Currently, the within-object constraints are mainly imposed by the likelihood, *i.e.* the distribution of properties and relations conditioned on non-NULL labels. How to use MRFs to encode within-model constraints into the prior distribution in a more efficient way is a topic in future research.

References

1. N. Ayache and O. D. Faugeras. "HYPER: A new approach for the representation and positioning of two-dimensional objects". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1):44–54, January 1986.
2. J. Besag. "On the statistical analysis of dirty pictures" (with discussions). *Journal of the Royal Statistical Society, Series B*, 48:259–302, 1986.
3. P. J. Besl and R. C. Jain. "Three-Dimensional object recognition". *Computing Surveys*, 17(1):75–145, March 1985.
4. R. Chellappa and A. Jain, editors. *Markov Random Fields: Theory and Applications.* Academic Press, 1993.
5. P. R. Cooper. "Parallel structure recognition with uncertainty: Coupled segmentation and matching". In *Proceedings of IEEE International Conference on Computer Vision*, pages 287–290, 1990.
6. O. Faugeras. *Three-Dimensional Computer Vision – A Geometric Viewpoint.* MIT Press, Cambridge, MA, 1993.
7. S. Geman and D. Geman. "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.
8. W. E. L. Grimson. *Object Recognition by Computer – The Role of Geometric Constraints.* MIT Press, Cambridge, MA, 1990.
9. J. Hornegger and H. Niemann. "Statistical learning, localization and identification of objects". In *Proceedings of IEEE International Conference on Computer Vision,* pages 914–919, MIT, MA, 1995.

10. R. A. Hummel and S. W. Zucker. "On the foundations of relaxation labeling process". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(3):267–286, May 1983.
11. S. Kirkpatrick, C. D. Gellatt, and M. P. Vecchi. "Optimization by simulated annealing". *Science*, 220:671–680, 1983.
12. S. Z. Li. *Markov Random Field Modeling in Computer Vision.* Springer-Verlag, New York, 1995.
13. S. Z. Li, H. Wang, K. L. Chan, and M. Petrou. "Minimization of MRF energy with relaxation labeling". *Journal of Mathematical Imaging and Vision*, 7:149–161, 1997.
14. K. V. Mardia and G. K. Kanji, editors. *Statistics and Images: 1.* Advances in Applied Statistics. Carfax, 1993.
15. J. W. Modestino and J. Zhang. "A Markov random field model-based approach to image interpretation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(6):606–615, 1992.
16. C. Peterson and B. Soderberg. "A new method for mapping optimization problems onto neural networks". *International Journal of Neural Systems*, 1(1):3–22, 1989.
17. Ullman. *High-Level Vision: Object Recognition and Visual Cognition.* MIT Press, 1996.