Recognizing Multiple Overlapping Objects in Image: An Optimal Formulation

Stan Z. Li
School of EEE, Nanyang Technological University
BLK S1, Nanyang Avenue, Singapore 639798
szli@szli.eee.ntu.edu.sg
http://markov.eee.ntu.edu.sg:8000/~szli/
Phone: +65 790-4540 Fax: +65 791-2687

Abstract

A statistically optimal formulation is presented for recognizing multiple, partially occluded objects. The optimality, in terms of the maximum *a posteriori* (MAP) principle, is with respect to all, rather than just individual, modeled objects. Various constraints are incorporated into the posterior distribution. A two-stage MAP estimation approach is proposed to reduce the computational cost.

Keywords

Contextual constraints, energy minimization, Markov random fields, maximum *a posteriori*, object recognition, object segmentation, statistical estimation.

I. Introduction

Model-based object recognition is a high level vision task by which objects in the scene are identified with reference to model objects. To recognize multiple objects under partially occlusions, the objects have to be separated from each other, and then identified. Because only partial observation is available, this task has to make use of information about local features such as points, line segments or regions and various constraints on them. The feature-based separation and identification of objects can be performed through object matching by which features correspondence is established between the scene and potential model objects.

The use of contextual constraints is essential in visual interpretation. An object can be considered as a pattern of mutually or contextually constrained features. 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 un-modeled features.

Exact and perfect solutions hardly exist for object recognition due to various uncertainties such as noise, occlusion, and unknown transformations from model objects to the scene. One usually looks for a solution that is optimal in satisfying some considered constraints. In terms of statistics, the optimal solution is defined to be one which maximizes some probability. The maximum *a posteriori* (MAP) principle is a statistical criterion 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 [1], [2], [3]. MRFs and MAP together leads 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. It has also been used for object recognition [5], [6], [3], [7].

In the past, matching between the scene and model objects is 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 with respect to a single model object [3], [7]. 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 [3], [7].

The contribution of this paper is a statistically optimal, MAP-MRF based, formulation for recognition of multiple, partially occluded objects (Section 2). The MAP solution is defined with respect to all, not just individual, model objects (Section 2B). Such a solution is overall optimal, and consistent by itself. A two-stage MAP estimation approach is proposed to reduce the computational cost (Section 2C). 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 (Section 2D). The energy functions for the two MAP estimation problems are formulated (Section 3). 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. By these, both BOCs and WOCs are incorporated into the posterior distribution. Experimental results are incorporated into the presentation to illustrate the theoretical formulation.

II. Object Recognition by MAP-MRF Estimation

Fig.1 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 ($L = 8$ in Fig.1:Left). The objects in the scene are rotated, translated, scaled, and partially occluded versions of the model objects (Fig.1:Middle). The objects and the scene are represented by features (*e.g.* corners on boundary curves in Fig.1:Right), and constraints on the feature such as unary properties and binary relations between features (the interested reader is referred to [8] 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, with respect to 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

Fig. 1. Left: Eight model jigsaw objects. Middle: A scene. Right: Three boundary curves and the corner features extracted from the scene.

the objects.

A. Matching as a Labeling Problem

Object matching can be posed as labeling problems [3], represented by a set $\mathcal{S}$ of sites, a set $\mathcal{L}$ of labels, and a mapping $f : \mathcal{S} \to \mathcal{L}$. Here, $\mathcal{S} = \{1, \ldots, m\}$ corresponds to the set of m image features. In Fig.1:Right, there are three sequences of corner features in the scene, represented by three $\mathcal{S}$ sets with $m = 24$, $m = 47$ and $m = 6$, respectively.

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 in Fig.1:Left). 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 necessity 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, $f : \mathcal{S} \to \mathcal{L}^{(all)}$. 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 non-NULL label), then image feature i belongs to an object; otherwise if $f_i^{(\alpha_i)} = 0$, it belongs to background, noise or an un-modeled 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 that image feature i is matched to.

B. MAP-MRF Estimation

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

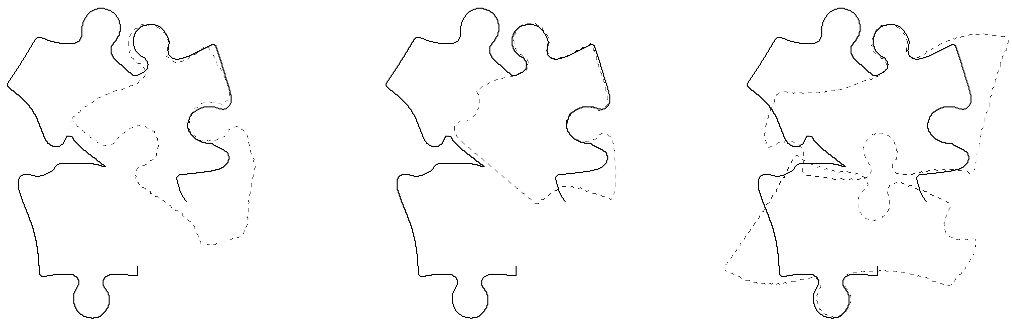
$$f^* = \arg \max_{f \in \mathbb{F}^{(all)}} P(f \mid d, \mathcal{O}^{(all)}) \tag{1}$$

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 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

$$E^{(all)}(f) \stackrel{\triangle}{=} U(f|\mathcal{O}^{(all)}) + U(d \mid f, \mathcal{O}^{(all)}) \tag{2}$$

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 derivations of the energies will be given in Section 3). The solution to problem (1) equivalently minimizes the posterior energy: $f^* = \arg\min_{f \in \mathbb{F}^{(all)}} E^{(all)}(f)$.



$f^{(2)}$ $f^{(3)}$ $f^{(7)}$

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

The objective of (1), 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.

C. Finding Solution in Two-Stage

Stage 1 solves L sub-problems, $f^{(\alpha)} = \arg\max_{f \in \mathbb{F}^{(\alpha)}} P(f \mid d, \mathcal{O}^{(\alpha)})$ for $\alpha = 1, \dots, L$, resulting in L MAP solutions $\{f^{(1)}, \dots, f^{(L)}\}$. Then, a reduced configuration space is constructed from the L solutions. In stage 2, the solution of (1) with respect to 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 $f^{(\alpha)} = \{f_1^{(\alpha)}, \dots, f_m^{(\alpha)}\} = \arg\min_{f \in \mathbb{F}^{(\alpha)}} E^{(\alpha)}(f)$ 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 the object α from the other image features in the following way: If image feature i belongs to the object α, then $f_i^{(\alpha)} \neq 0$ (a non-NULL label); otherwise, $f_i^{(\alpha)} = 0$. Second, if $f_i^{(\alpha)} \neq 0$, $f_i^{(\alpha)}$ is the model feature that image feature i is matched to.

$f^{(\alpha)}$ is optimal with respect to 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 features belonging to different objects; that is, there may exist more than one $\alpha \in \{1, \dots, L\}$ for which $f_i^{(\alpha)} \neq 0$. For example, in Fig.2, 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. (Fig.2 also shows that an 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 (1) with respect to all the L objects, in a reduced solution space constructed from the stage 1 solutions $\{f^{(1)}, \dots, f^{(L)}\}$ (see 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. Fig.3 shows the overall optimal result, which is the output of stage 2, for the MAP recognition of the scene with respect to all the models. It is consistent by itself.

D. Reduced Solution Space

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

Fig. 3. The overall matching and recognition result.

TABLE I
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	$5^{(4)}$	$4^{(4)}$	$3^{(4)}$	$2^{(4)}$	$7^{(5)}$	$6^{(5)}$	$5^{(5)}$	$4^{(5)}$	$3^{(5)}$

- For $i \in \mathcal{S}'$, $\mathcal{L}_i^{(all)}$ consists of all the non-NULL labels assigned to i by the $f^{(\alpha)}$'s, plus the NULL label; that is $\mathcal{L}_i^{(all)} = 0 \cup \{f_i^{(\alpha)} \neq 0 \mid \alpha = 1, \ldots, L\}$. For the case in Table I, for example, $\mathcal{L}_8^{(all)} = \{0, 10^{(2)}, 1^{(4)}, 7^{(5)}\}$, $\mathcal{L}_9^{(all)} = \{0, 9^{(2)}, 6^{(5)}\}$, $\mathcal{L}_{10}^{(all)} = \{0, 7^{(2)}, 3^{(3)}, 5^{(5)}\}$, $\mathcal{L}_9^{(all)} = \{0, 4^{(3)}, 4^{(5)}\}$.
- 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 $i \notin \mathcal{S}'$, the size of $\mathcal{L}_i^{(all)}$ is one; whereas before the reduction, the size is $1 + \sum_{\alpha=1}^{L} M^{(\alpha)}$ 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 I, 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 Eq.(1) 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 allowed to also 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 Eq.(1) if $\mathcal{L}_i^{(\alpha)}$ derived from $f^{(\alpha)}$ contains the correct matching components for the original problem.

III. Formulation of Energy Functions

A. Energy Functions for Stage 1

The prior distribution of $f^{(\alpha)} = \{f_1^{(\alpha)}, \ldots, f_m^{(\alpha)}\}$ for matching with respect to 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 following form [3], [7]

$$U(f^{(\alpha)} \mid \mathcal{O}^{(\alpha)}) = \sum_{i \in \mathcal{S}} V_1(f_i^{(\alpha)}) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)}) \tag{3}$$

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 as

$$V_1(f_i^{(\alpha)}) = \begin{cases} 0 & \text{if } f_i \neq 0 \\ v_{10} & \text{if } f_i = 0 \end{cases}, \quad V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)}) = \begin{cases} 0 & \text{if } f_i^{(\alpha)} \neq 0 \text{ and } f_{i'}^{(\alpha)} \neq 0 \\ v_{20} & \text{if } f_i^{(\alpha)} = 0 \text{ or } f_{i'}^{(\alpha)} = 0 \end{cases} \tag{4}$$

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 [4] for differentiating edge and non-edge 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 non-NULL matches ($f_i^{(\alpha)} \neq 0$) only. The likelihood is a Gibbs distribution with the energy function [3], [7]

$$U(d \mid f^{(\alpha)}, \mathcal{O}^{(\alpha)}) = \sum_{i \in \mathcal{S}, f_i^{(\alpha)} \neq 0} V_1(d_1(i) \mid f_i^{(\alpha)}) + \sum_{i \in \mathcal{S}, f_i^{(\alpha)} \neq 0} \sum_{i' \in \mathcal{S}_{\setminus i}, f_{i'}^{(\alpha)} \neq 0} V_2(d_2(i,i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha)}) \tag{5}$$

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 [3]). 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 $E^{(\alpha)}(f^{(\alpha)}) = U(f^{(\alpha)} | \mathcal{O}^{(\alpha)}) + U(d \mid f^{(\alpha)}, \mathcal{O}^{(\alpha)})$.

B. Energy Functions for Stage 2

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 to each other when $\alpha \neq \alpha'$.

The prior energy is

$$U(f \mid \mathcal{O}^{(all)}) = \sum_{i \in \mathcal{S}} V_1(f_i \mid \mathcal{O}^{(all)}) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'} \mid \mathcal{O}^{(all)}) \tag{6}$$

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 Eq.(4) 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

$$V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha')} \mid \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')}) = \begin{cases} V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)}) & \text{if } \alpha = \alpha' \\ v_{20} & \text{otherwise} \end{cases} \tag{7}$$

where $V_2(f_i^{(\alpha)}, f_{i'}^{(\alpha)})$ are as defined in (4) for matching to object α. Substituting Eq.(4) into Eq.(7) yields

$$V_2(f_i, f_{i'} \mid \mathcal{O}^{(all)}) = \begin{cases} 0 & \text{if } (\alpha = \alpha') \text{ and } (f_i \neq 0) \text{ and } (f_{i'} \neq 0) \\ v_{20} & \text{otherwise} \end{cases} \tag{8}$$

where f_i is associated with object α and $f_{i'}$ with object α'. The above definitions are an extension of Eq.(4) for dealing with multiple objects: In Eq.(4), features due to other objects (not belonging to object α) are all labeled as NULL ; Eq.(8) simply takes this into consideration.

The likelihood energy is

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

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 (5). 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

$$V_2(d_2(i,i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha')}, \mathcal{O}^{(\alpha)}, \mathcal{O}^{(\alpha')}) = \begin{cases} V_2(d_2(i,i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha)}) & \text{if} \quad \alpha = \alpha' \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $V_2(d_2(i,i') \mid f_i^{(\alpha)}, f_{i'}^{(\alpha)})$ are defined in the same way as for (5) when for 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 [3].

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 [9] is preferable in terms of the minimized energy value and computational costs [3] and is used in our 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.

IV. Summary

This work has presented an MAP formulation for the recognition of multiple overlapping objects. In contrast to the past work, where matching between a scene containing multiple objects is performed by considering one model object at a time, the MAP solution here is formulated with respect to all the model objects and is therefore overall optimal. Moreover, it is overall consistent by itself, with no need for *ad hoc* post-processing. The computational cost may be lowered by using a two-stage MAP estimation approach.

References

[1] R. Chellappa and Anil Jain, Eds., *Markov Random Fields: Theory and Applications*, Academic Press, 1993.

[2] Kanti V. Mardia and Gopal K. Kanji, Eds., *Statistics and Images: 1*, Advances in Applied Statistics. Carfax, 1993.

[3] S. Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, New York, 1995.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, November 1984.

[5] J. W. Modestino and J. Zhang, "A Markov random field model-based approach to image interpretation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 6, pp. 606–615, 1992.

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

[7] S. Z. Li, "Bayesian object matching", *Journal of Applied Statistics*, vol. 25, no. 3, pp. 425–443, 1998.

[8] S. Z. Li, "Invariant representation, recognition and pose estimation of 3d space curved under similarity transformations", *Pattern Recognition*, vol. 30, no. 3, pp. 447–458, 1997.

[9] R. A. Hummel and S. W. Zucker, "On the foundations of relaxation labeling process", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 3, pp. 267–286, May 1983.