A Markov Random Field Model for Object Matching under Contextual Constraints

S. Z. Li
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore 2263
szli@ntu.ac.sg

Abstract

This paper presents a Markov random field (MRF) model for object recognition in high level vision. The labeling state of a scene in terms of a model object is considered as an MRF or couples MRFs. Within the Bayesian framework, the optimal solution is defined as the maximum a posteriori *(MAP) estimate of the MRF. The posterior distribution is derived based on sound mathematical principles from theories of MRF and probability, which is in contrast to heuristic formulations. An experimental result is presented.*

1 Introduction

In object recognition, an object is usually represented by a set of primitives or features. These features are attributed by their properties and are constrained to one another by contextual inter-relations. Two issues must be addressed for successful recognition: how to use contextual constraints effectively and how to deal with uncertainties.

Markov random field (MRF) theory provides a way of encoding contextual constraints. Since 1980's, there has been considerable interest in image and vision modeling using MRF. The main reason for MRF vision modeling is that results from MRF theory provide us with tools to encode contextual constraints into the prior probability. Apart from this, MRF modeling also provides other advantages (see Preface of [2] by R. Chellappa and A.K. Jain). It has been successfully used for modeling low level vision processes. (There is a long list of references in such work. See collections of papers in [2, 9] and references therein.) A prototypical Bayesian formulation using MRF is that of Geman and Geman [4] for image restoration.

MRF modeling in high level vision such as object matching and recognition is more challenging (see *e.g.* Introduction of [9] by Mardia). Modestino and Zhang in their recent paper [10] describe their initial work in developing an MRF model for image interpretation. They define the optimal matching as the MAP estimate of the MRF. Unfortunately, the posterior probability therein is derived not by using the laws of probability; rather, it is designed directly by using some heuristic rules. This contradicts the orginal motivation and dissolves the promises of MRF vision modeling. Cooper [3] describes a coupled network for simultaneous object recognition and segmentation. MRF is used to encode prior qualitative, and possibly, quantitative knowledge in non-homogeneous and non-isotropic situations.

This work formulates an MRF model for high level object recognition based on probability theory in a rational way, in contrast to heuristic methods. Section 2 presents the MRF model, including the derivation of the posterior probability for matching to one and multiple objects and introducing coupled MRFs for the matching of different features. Section 3 presents an experimental result. Section 4 concludes the paper.

2 The MRF Model

2.1 Bayesian Labeling Based on MRF

Let $\mathbf{d} = \{1, \ldots, m\}$ be a set of m discrete *sites* and $\mathbf{D}^+ = \{0, 1, \ldots, M\}$ be a set of *labels* which include M physical labels $(1, \ldots, M)$ and a virtual NULL label (0). We want to assign a label from $\mathbf{D}^+$ to each of the sites in $\mathbf{d}$ so as to satisfy some contextual constraints. Let $f = \{f_1, \ldots, f_m\}$ be a configuration of an MRF, with $f_i \in \mathbf{D}^+$ assuming a value in the label set. It can be interpreted as a mapping $f: \mathbf{d} \longrightarrow \mathbf{D}^+$ or a labeling of $\mathbf{d}$. The set of all configurations is $\mathbf{S} = \underbrace{\mathbf{D}^+ \times \mathbf{D}^+ \cdots \times \mathbf{D}^+}_{m \text{ times}}$.

Given the likelihood function $p(\mathbf{r} \mid f)$ and the *a priori* probability $P(f)$, the posterior probability can be computed by using the Bayesian rule $P(f \mid \mathbf{r}) \propto p(\mathbf{r} \mid f)P(f)$. The *Bayesian labeling problem* is the following: given the observation $\mathbf{r}$, find the MAP configuration f^* from an admissible space $\mathbf{S}$, that is, $f^* = \arg\max_{f \in \mathbf{S}} P(f \mid \mathbf{r})$.

According to the Hammersley-Clifford theorem [1] of Markov-Gibbs equivalence, the prior probability

$P(f)$ obeys a Gibbs distribution $P(f) = Z^{-1} \times e^{-\frac{1}{T}U(f)}$ where Z is a normalizing constant, T is a global control parameter called the temperature and $U(f)$ is the *prior energy*. The prior energy has the form $U(f) = \sum_{c \in \mathcal{C}} V_c(f)$ where $\mathcal{C}$ is the set of cliques in a neighborhood system $\mathcal{N} = \{\mathcal{N}_i \mid \forall i \in \mathbf{d}\}$ for $\mathbf{d}$ in which $\mathcal{N}_i$ is the collection of sites neighboring to i.

The likelihood $p(\mathbf{r} \mid f)$ depends how $\mathbf{r}$ is observed. It can usually be represented in an exponential form $p(\mathbf{r} \mid f) = Z_r^{-1} \times e^{-U(\mathbf{r} \mid f)}$ where $U(\mathbf{r} \mid f)$ is the likelihood energy. Hence the posterior probability is a Gibbs distribution $P(f \mid \mathbf{r}) = Z_E^{-1} \times e^{-U(f \mid \mathbf{r})}$ with *posterior energy*

$$U(f \mid \mathbf{r}) = U(f)/T + U(\mathbf{r} \mid f) \tag{1}$$

Therefore, given an observation $\mathbf{r}$, a labeling f of sites in $\mathbf{d}$ is also an MRF on $\mathbf{d}$ with respect to $\mathcal{N}$. The MAP solution is equivalently found by

$$f^* = \arg\min_{f \in \mathbf{S}} U(f \mid \mathbf{r})$$

.

2.2 Deriving the Posterior Distribution

Neighborhood System and Cliques

In all cases, $\mathcal{N}_i$ can be the set of all the other sites $j \neq i$. This is a trivial case for MRF. In contextual matching, it can consist of all other sites which are related to i by the observed relations in $\mathbf{r}$. When the scene is very large, $\mathcal{N}_i$ needs include only those of the other sites which are within a spatial distance from i, *i.e.* $\mathcal{N}_i = \{j \neq i \mid dist(\text{feature}_j, \text{feature}_i) < \alpha, j \in \mathbf{d}\}$. The threshold α can be reasonably related to the size of the model object. The set of first order cliques is $\mathcal{C}_1 = \{\{i\} \mid i \in \mathbf{d}\}$. The set of second order cliques is $\mathcal{C}_2 = \{\{i,j\} \mid j \in \mathcal{N}_i, i \in \mathbf{d}\}$. In this paper, only cliques of up to order two are considered.

Prior Energy

The single site potential is defined as

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

where v_{10} is a constant. This definition says that if f_i is the NULL label, it incurs a penalty v_{10}; or otherwise no penalty. The two sites potential is defined as

$$V_2(f_i, f_j) = \begin{cases} v_{20} & \text{if } f_i = 0 \text{ or } \ f_j = 0 \\ 0 & \text{otherwise} \end{cases}$$

where v_{20} is a constant. This says that if either f_i or f_j is the NULL , it incurs a penalty v_{20}; or otherwise no penalty. The above clique potentials define the prior energy.

Likelihood Energy

The joint likelihood function $p(\mathbf{r} \mid f)$ 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 $\mathcal{N}$, and (3) It depends on how the model object is observed in the scene which in turn depends on the underlying transformations and noise. Assume $\mathbf{r} = (r_1, r_2)$ where $r_1 = \{r_1(i) \mid i \in \mathbf{d}\}$ and each $r_1(i)$ is a vector of K_1 unary properties; $r_2 = \{r_2(i,j) \mid i,j \in \mathbf{d}, i \neq j\}$ and each $r_2(i,j)$ is a vector of K_2 binary relations. The same assumptions are also made for $\mathbf{R} = (R_1, R_2)$. The properties and relations are assumed to be invariant under the class of underlying transformations from $\mathbf{R}$ to $\mathbf{r}$. Assuming an observation model to be $\mathbf{r} = \mathbf{R} + \mathbf{n}$ where $\mathbf{n}$ is the independent Gaussian noise, then the likelihood energy is

$$U(\mathbf{r} \mid f) = \sum_{i \in \mathbf{d}; f_i \neq 0} V_1(\mathbf{r} \mid f_i) + \sum_{i,j \in \mathbf{d}; f_i, f_j \neq 0} V_2(\mathbf{r} \mid f_i, f_j)$$

Because the noise is white, we have $U(\mathbf{r} \mid f_i) = U(r_1(i) \mid f_i)$ and $U(\mathbf{r} \mid f_i, f_j) = U(r_2(i,j) \mid f_i, f_j)$. The likelihood potentials are

$$V_1(r_1(i) \mid f_i) = \sum_{k=1}^{K_1} [r_{1,k}(i) - R_{1,k}(f_i)]^2 / 2\sigma_{1,k}^2$$

and

$$V_2(r_2(i,j) \mid f_i, f_j) = \sum_{k=1}^{K_2} [r_{2,k}(i,j) - R_{2,k}(f_i, f_j)]^2 / 2\sigma_{2,k}^2$$

where $\sigma_{n,k}^2$ ($k = 1, \ldots, K_n$ and $n = 1, 2$) are the standard deviations of the noise components. The vectors $R_1(f_i)$ and $R_2(f_i, f_j)$ are the "mean vector" for the random vectors $r_1(i)$ and $r_2(i,j)$, respectively. When the noise is correlated, there are correlating terms in the likelihood potentials. The assumption of the independent Gaussian noise made may not be accurate but offers an approximation when an accurate observation model is not available.

Posterior Energy

The posterior energy in Eq.(1) can then be derived as

$$\begin{aligned} U(f \mid \mathbf{d}) = \ & \textstyle\sum_{\{i\} \in \mathcal{C}; f_i = 0} V_{10}(f_i)/T + \\ & \textstyle\sum_{\{i,j\} \in \mathcal{C}; f_i = 0 \text{ or } f_j = 0} V_{20}(f_i, f_j)/T + \\ & \textstyle\sum_{\{i\} \in \mathcal{C}; f_i \neq 0} V_1(r_1(i) \mid f_i) + \\ & \textstyle\sum_{\{i,j\} \in \mathcal{C}; f_i \neq 0 \text{ and } f_j \neq 0} V_2(r_2(i,j) \mid f_i, f_j) \end{aligned}$$

It can be written into a compact form

$$U(f \mid \mathbf{d}) = E(f) = \sum_{i\in\mathbf{d}} E_1(f_i) + \sum_{i\in\mathbf{d}}\sum_{j\in\mathbf{d}} E_2(f_i, f_j)$$

where

$$E_1(f_i) = \begin{cases} v_{10}/T & \text{if } f_i = 0 \\ V_1(r_1(i) \mid f_i) & \text{otherwise} \end{cases}$$

and

$$E_2(f_i, f_j) = \begin{cases} v_{20}/T & \text{if } j \in \mathcal{N}_i \ \& \ (f_i = 0 \text{ or } \ f_j = 0) \\ V_2(r_2(i,j) \mid f_i, f_j) & \text{otherwise} \end{cases}$$

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

The involved parameters are the noise variances $\sigma^2_{n,k}$ and the prior penalties v_{n0}, with $T = 1$ fixed. Only the relative, not absolute, values of $\sigma^2_{n,k}$ 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. Estimation of these parameters under Gaussian and unknown noise is addressed in a separate paper [6].

2.3 Matching to Multiple Objects

The global energy $E(f)$ can be decomposed into the sum of local costs $E(f) = \sum_{i\in\mathbf{d}} V(f_i \mid \mathbf{r})$ where

$$V(f_i \mid \mathbf{r}) = E_1(f_i) + \sum_{j\in\mathbf{d}} E_2(f_i, f_j)$$

is the cost incurred by the local match $i \to f_i$. The MAP configuration f^* is the optimal mapping from the scene to the model object under consideration. In other words, it is the optimal labeling of the scene in terms of the model object. In matching to multiple model objects, one can use the local costs $V(f_i \mid \mathbf{r})$ as the basis for selecting the the best matched objects for i.

Suppose there are L potential model objects. Then L MAP solutions, $f^{(1)}, \ldots, f^{(L)}$, can be obtained after matching the scene to each of the models in turn. The local cost of $f^{(l)}$ $(l = 1, \ldots, L)$ at i is

$$V(f_i^{(l)} \mid \mathbf{r}) = E_1(f_i^{(l)}) + \sum_{j\in\mathbf{d}} E_2(f_i^{(l)}, f_j^{(l)})$$

We may decide that the image feature i is due to object l_i if the decision incurs the least cost

$$l_i = \arg \min_{l\in\{1,\ldots,L\}} V(f_i^{(l)} \mid \mathbf{r})$$

The final label for i is feature No. $f_i^{(l_i)}$ of object l_i.

2.4 Coupled MRFs for Matching of Different Features

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

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

If there are two different types of features, then all the physical can be partitioned into two admissible sets, with each set consisting of indices to a single type of features. In the most general case, each of the m sites has its own set of labels $\mathbf{D}_i$ $(i = 1, \ldots, m)$, each $\mathbf{D}_i$ being determined using the symbolic unary constraints; and the label for site i assumes a value $f_i \in \mathbf{D}_i^+$ where $\mathbf{D}_i^+ = \{0\} \cup \mathbf{D}_i$. Then, the configuration space is defined as $\mathbf{S} = \mathbf{D}_1^+ \times \mathbf{D}_2^+ \times \cdots \times \mathbf{D}_m^+$. In this situation, the energy $E(f)$ has the same form as usual and the solution is still found by $f^* = \arg\min_{f\in\mathbf{S}} E(f)$. The only difference is in the definition of the configuration space $\mathbf{S}$ in which the solution is searched for.

3 Experiment Result

Here we present a result of recognizing 2D jigsaw objects under translation, rotation and scaling transformations using the MRF model. There are eight model jigsaw objects in this experiment. An overlapping scene of these objects is taken. It is composed of rotated, translated and scaled parts of the model objects; some of them are considerably occluded. Boundaries of the scene and model objects are extracted using canny detector followed by hysteresis and edge linking and then corners are detected.

In this case of curve matching, the sites correspond to the feature points (corners) on the scene curve and

Figure 1: A recognition result.

the labels correspond to the feature points on the considered model curve. The neighbors of a site is defined as the 5 forward points and the 5 backward points. Invariant relations are derived from the boundaries as well as the corners based on a similarity-invariant representation of curves [7]. Only binary relations are used, which are $r_{2,1}(i,j)$ the ratio of estimated curvature values at feature points i and j, $r_{2,2}(i,j)$ the ratio of arc-length to cord-length, $r_{2,3}(i,j)$ the invariant coordinate vector, $r_{2,4}(i,j)$ the invariant radius vector and $r_{2,5}(i,j)$ the invariant angle vector.

The parameters are learned from a humanly-supplied example matching by an automated procedure [6], which gives the following values $v_{20} = 0.7$, $1/\sigma_{2,1} = 0.00366$, $1/\sigma_{2,2} = 0$, $1/\sigma_{2,3} = 0.09466$, $1/\sigma_{2,4} = 0.00251$ and $1/\sigma_{2,5} = 0.03669$. The minimal labeling f^* is searched for by using a deterministic relaxation labeling algorithm [5]. The final result of recognition is shown in Figure 1 in which model objects are aligned with the corresponding parts in the scene. Note that most of the model objects share common structures of round extrusion and intrusion, which means ambiguities exist extensively and have to be resolved by using context.

4 Conclusion

MRF modeling has been recognized as a rational approach to solving a variety of computer vision problems. This work contributes to this area by presenting an MRF model for high level object recognition under contextual, relational constraints. In contrast to previous work in MRF modeling for the recognition problem, the present MRF model is formulated based on sound probability principles, in deriving the posterior distribution for MAP solutions.

The present model has close relationships to existing low level MRF vision models and their similarities imply a unified MRF framework for computer vision [8]. In this framework, we may categorize MRF vision models in terms of the spatial "homogeneity" of (sites and) neighborhood systems and the "continuity" of labels. Such categorization could be significant in the understanding of MRF modeling in computer vision and in the design of vision computation architectures.

References

[1] J. Besag. "Spatial interaction and the statistic analysis of lattice systems". *J. Royal. Statist. Soc. B*, 36:192–293, 1974.

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

[3] 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.

[4] G. Geman and D. Geman. "Stochastic relaxation, gibbs distribution and bayesian restoration of images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, November 1984.

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

[6] S. Z. Li. "MRF parameter estimation for object recognition". *submitted*, 1993.

[7] S. Z. Li. "Similarity invariants for 3D space curve matching". In *Proceedings of the First Asian Conference on Computer Vision*, Osaka, Japan, November 1993.

[8] S. Z. Li. "Markov random field models in computer vision". In *Proceedings of the European Conference on Computer Vision*, Stockholm, Sweden, May 1994.

[9] K. V. Mardia, editor. Special Issue on Statistic Image Analysis. *Journal of Applied Statistics*, volume 16, no.2. 1989.

[10] 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.