Probabilistic Representation and Inference in Natural Scene Segmentation

Erich Leung

John Tsotsos

Technical Report CSE-2009-02

March 15 2009

Department of Computer Science and Engineering

4700 Keele Street Toronto, Ontario M3J 1P3 Canada

Abstract

This paper outlines some major ideas and principles of probabilistic reasoning that provide an essential foundation for probabilistic approaches to natural scene segmentation. Despite this immediate context, the discussion may be found relevant beyond image segmentation in a broader scope of image processing and visual analysis. The organization of the paper corresponds to two recurrent themes in the literature of probabilistic approaches to the problem: the representational framework of probabilistic graphical models and the framework of probabilistic inference in natural scene labeling. After a brief discussion of both the directed and undirected graphical models, the first part of the paper mainly focuses on the undirected graphical models of stochastic random fields, which have been widely adopted to represent the inference problems of natural scene analysis. Some details are also covered of the distinction between generative random fields and conditional random fields. Irrespective of the choice of representational language, the probability theory of inference plays a critical role in the probabilistic approaches to natural scene segmentation. The second part of the paper discusses three major themes of probabilistic inference in the literature of probabilistic approaches to object class segmentation, namely, the problems as to (1) inference of stochastic models of natural scene labeling, (2) inference of visual labeling of natural images, and (3) distribution approximation of probabilistic models.

Contents

1 Introduction **1**

2 Probabilistic Graphical Models **6**

2.1 Directed Graphical Models 7

2.2 Undirected Graphical Models 9

2.2.1 Stochastic random fields. 9

2.2.2 Gibbs random fields: normalized energy-based models. 11

2.2.3 Generative random fields. 14

2.2.4 Conditional random fields. 17

3 Probabilistic Inference **20**

3.1 Inference of Probabilistic Models 21

3.2 Optimization Problems of Natural Scene Labeling 26

3.3 Approximations of Inference Models 29

3.3.1 Stochastic distribution approximations 29

3.3.2 Deterministic distribution approximations quad 30

3.3.3 Limitations of approximations in natural scene segmentation 32

References **43**

1 Introduction

Object class segmentation seeks to segregate instances of semantic categories or object classes from a scene along the semantic boundaries pertinent to meaningful objects of human intuition. The problem is usually formulated as some form of probabilistic inference. An image is viewed as stochastic events, or samples drawn from image ensembles, which can be described, represented, analyzed and interpreted in terms of probability distributions of plausible visual interpretations, i.e., scene descriptions. In particular, object class segmentation is conceptualized as a stochastic mapping of visual patterns recoverable from image data to object class membership. This mapping encodes *a priori* knowledge of image semantics and the rules of interpretation and is completely specified by the probabilistic models of membership assignment given observation over the solution space. Image segmentation and object class segregation can then be defined by an assignment of labels to pixels in a 2D grid or subsets of pixels that form regions or super-pixels. These principles constitute a unified framework for different computational approaches to object class segmentation, which collectively referred to as the probabilistic approaches to object class segmentation.

This paper does not aim at a survey of different approaches to the problem[1]. Instead, it outlines the major ideas and principles of probabilistic reasoning, which are widely adapted in the recent development of computational frame-

[1]A review paper on the probabilistic approaches to object class segmentation is under preparation.

works for object class segmentation. The organization of the paper corresponds to two recurrent themes in the literature of probabilistic approaches to object class segmentation: the representational framework of probabilistic graphical models for formal description of the inference problems and the framework of probabilistic inference of natural scene description. Its motivation notwithstanding, the ensuing discussion will also be found relevant beyond its immediate context of image segmentation in a broader scope of image processing and visual analysis; for further discussion, see [18, 37, 89, 38, 11, 80, 87].

It is an emerging theme in the literature to encode the statistical regularities and configurations of visual patterns pertinent to object classes in the language of probabilistic graphics models, or graphical models for short [62, 31, 17, 6]. A graphical model associates a probability distribution with a graph. The sites (nodes) of the graph represent the random variables on which the distribution is defined, and the edges (or links) between these variables express the probabilistic relationship between these variables. It gives explicit expression for the dependence relations among important variables of the system, thus allowing a simple and intuitive way to visualize as well as to specify the structures of an inference problem. These structures can usually give insights into the influence of a set of random variables on the distribution over the other parts of the system. Two classes of graphical models are widely adopted for probabilistic representation of the problems of visual inference, viz, the directed graphical models, i.e., probabilistic models defined over directed graphs, and undirected graphical models, that is, probabilistic models defined over undirected graphs.

In general, directed graphical models or Bayesian networks are more useful for describing causal relationships between variables, whereas the undirected graphical models or random fields provide an efficient tool for representing contextual constraints between spatially related variables. The latter also provide more convenient graphical semantics that allows more explicit expression, and thus also more efficient way to specify the structures of conditional independence [17, 6]. Although both classes of graphical models has been applied to modeling image ensembles for image analysis, the properties of graphical semantics has made the undirected models a choice of representational framework for a wider range of problems in visual inference as well as natural scene segmentation; for further discussion, see [57, 11, 80]. After a brief discussion of the distinction between these two classes of graphical models in the context of natural scene segmentation, the first part of this paper focuses on the basic ideas of the undirected graphical models and their underlying theoretical properties. Also highlighted in the discussion are the important distinctions between the generative and conditional random fields.

A probabilistic model is a mathematical description of an inference problem. Irrespective of the choice of language, stochastic fields or otherwise, deployed to encode the problem, probabilistic inference plays a critical role in the probabilistic approaches due to their common assumption that natural scene analysis is essentially part of the perceptual processes which infer the states of the external environment from visual input. The second part of the paper shifts its focus to those theoretical principles of probabilistic inference that consti-

tute a unified theoretical framework for different probabilistic approaches to object class segmentation. The discussion is organized in terms of three major themes, namely (1) inference of probabilistic models, (2) inference of visual description, and (3) distribution approximations.

At its core, probabilistic approaches conceptualize image interpretation in object class segmentation as a stochastic mapping of visual patterns recoverable from image data to object class membership. This mapping is completely encoded by the conditional probability of membership assignment given observation over the solution space. Visual appearances of natural scenes are known to be highly complex, varying and ambiguous. It becomes overwhelming to construct a probabilistic models on a case-by-case basis for each class of natural scenes. One recurrent theme running through the current approaches of object categorization and semantic-based image analysis is to generate a predictive model for a given class of scenes from a set of parameterized model classes through adapting their structure and parameters to match empirical observation. The goal of inference is therefore to generalize the knowledge derived from a subset of interpreted data to a general model capable of evaluating observational data of the category as a whole. This empirical approach to model specification serves as a cornerstone of the computational procedures of inferring the underlying causes of observation in many probabilistic approaches to natural scene analysis. Visual inference entails two different problems, viz, model selection and interpretation selection. The first problem is concerned with searching over the space of model parameters for the best

member from a given family of models in terms of their ability to provide the best segmentation prediction for a set of training images and to generalize the performance over a general class of natural images. Interpretation selection aims at an optimal assignment of scene description that incurs the least expected risk due to misclassification, according to a trained probabilistic model. The probabilistic models that describe natural scene are usually too complex to allow exact inference. In practice, probabilistic approaches must resort to an approximation to the exact solution. This part concludes with two major classes of distributional approximation which are recurrently used in the literature of object class segmentation, viz, stochastic and deterministic distribution approximations.

2 Probabilistic Graphical Models

Natural scene segmentation entails giving meaning to different parts of an image, usually in terms of semantic categories. Probabilistic approaches seek the probability distributions over the space of image description, such that different parts of an image may be assigned the object class membership that is most likely under the distribution given some local patterns of visual input or responses of classifiers. An essential problem is how to represent knowledge about observation and to encode its underlying interpretation. For natural scene analysis, it oftentimes involves complex, non-linear relationships of contextual interactivity and influence between observable patterns, object class membership, explanatory constructs and *a priori* beliefs. It is a recurrent theme among the probabilistic approaches to encode the inference problem in the language of probabilistic graphical models, or graphical models for short[2]. The cornerstone of these approaches is the graphical structures of representation and inference [62, 31, 17, 6]. A probability distribution is associated with a graph, where its sites (nodes) represent random variables and its edges (links) the probabilistic relationship between these variables. The state of a site is defined by the local responses or computational decisions associated with the site. These representations provide an intuitive description of the spatial structures and contextual interactivity of image features and computational decisions among different sites. The modular structure of these relationships are captured by decomposition of the joint distribution in terms of the neigh-

[2]For other approaches, see, for examples, [40, 9, 8, 53, 51, 10].

borhood structure of the underlying graph, such that, each variable depends on only a subset of other variables in a local neighborhood. That is, given the state of its neighborhood, the variable is statistically independent of the rest of the graph.

2.1 Directed Graphical Models

Both directed graphical models and undirected graphical models have been applied to object class segmentation. The two families are distinct by their underlying graphical representation. The former associated a probability distribution with a directed acyclic graph on a set of random variables, where the joint distribution is given by the product, over all the sites of the graph, of the conditional probability one for each variable conditioned on the variables corresponding to its parents [17, 6]. A tree-structured belief network, for instance, is proposed for natural scene labeling. A scene description is inferred from observation based on a prior model of scene description and the prediction of description conditioned on observation. The latter is given by the response of a set of trained classifiers. A belief network is deployed to encode the prior model of scene descriptions across scales of granuity [16]. The description associated with a node in the belief tree is dependent only on the coarser scale description at its parent node given those at all coarser scale nodes. The common core of the latent class approaches[3], rests on their representation with

[3]Latent class approaches refer to those inference approaches that extend the probabilistic latent Semantic analysis (PLSA) [27, 28] or latent Dirichlet allocation (LDA) [7] to semantic-based visual analysis. LDA can be viewed as an extension of PLSA with an additional Dirichlet prior for the probability distribution of the latent topics.

directed graphical models of the generative processes of observation. The key idea is to explain observation in a high dimensional and usually sparse feature space by a set of latent topics populated in a lower dimensional probabilistic semantic space. The joint distribution of observation is factorized into a product of local condition distributions of visual patterns given the latent topics and the distribution of the topics. In the context of object class segmentation, each subimage, usually an over-segmented superpixel, is modeled as a mixture of the latent topics. These latent topics are inferred from a set of discriminative visual patterns extracted by image-based techniques [67, 75, 24]. The recovered topics are then mapped to object class membership. Under the framework of image parsing, a hierarchical parsing graph is deployed to represent the decomposition of an image into constitutive components in successive level of details [74]. The parsing graph is built upon a directed structure of part relationships with lateral, undirected connections between parts on the same level of decomposition for describing spatial relationships between visual patterns. The parsing graph of an image is reconfigured on the fly for an image, using reversible jump Markov Chain Monte Carlo. These approaches attest the usefulness of the language of directed graphical models in semantic-based visual analysis.

Images are more often represented as stochastic random fields, which associate a potential function for each variable corresponding to a site of an undirected graph for each maximum clique[4] [57, 11, 17, 80]. An undirected graphical

[4]A clique is a set of sites where any two different elements of the set are neighbors [80].

model can be interpreted as a normalized energy model which provides an efficient and intuitive tool for encoding statistical regularities for visual inference. It is therefore no surprise that stochastic random fields have increasingly been adopted as a common language for expressing many inference problems of natural scene segmentation; for examples see [25, 44, 26, 43, 56, 64, 81, 68, 82, 29, 66, 75]. The following discussion[5] is concerned mainly with the central concepts of undirected graphical models, that are essential to understanding the workings of many recent approaches to object class segmentation.

2.2 Undirected Graphical Models

2.2.1 Stochastic random fields.

An undirected graphical model is a random field defined over an undirected graph, $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of sites (or nodes) and $\mathcal{E}$ the set of edges (or links) of $\mathcal{G}$. Consider a set of sites, $\{\varsigma_i \in \mathcal{S}\}$, each representing a random variable defined over a space $\mathcal{X}_{\varsigma_i}$ of states, x_{ς_i}. Each of these sites corresponds to a node in the underlying graph, and their spatial interactions are encoded by the edges (or links), each of which connects a pair of sites. The variable at each site takes on a state x_{ς_i}; the totality of these states constitutes a state of the graph, and is referred to as a configuration, $x \in \mathcal{X} = \prod_{\varsigma \in \mathcal{S}} \mathcal{X}_\varsigma$. An undirected graph induces a neighborhood system[6], $\mathcal{N} = \{\mathcal{N}_i : \varsigma_i \in \mathcal{S}\}$, through which the

[5] The major sources for this brief review on the central concepts of random field modeling include Geman and Geman [20], Besag [4], Kato [36], Kopparapu and Desai [41], Li [57], Won and Gray [83], Chan and Shen [11], Bishop [6], LeCun, Chopra, Hadsell, Ranzato, and Huang [50], Winkler [80], and Grenander and Miller [22], which are not cited separately.

sites relate to each other. The families of random fields deployed in object class segmentation are characterized by two properties with respect to the neighborhood system $\mathcal{N}$. The property of positivity (RF-Pos), $pr(\mathcal{X}) > 0$, and the property of Markovianity (RF-Mar), $pr(\mathcal{X}_{\varsigma_i}|\mathcal{X}_{\mathcal{S}/\{\varsigma_i\}}) = pr(\mathcal{X}_{\varsigma_i}|\mathcal{X}_{\mathcal{N}_i})$, where $\mathcal{S}\backslash\{\varsigma_i\}$ is the set of sites of the model with $\{\varsigma_i\}$ removed, and $\mathcal{X}_{\mathcal{N}_i}$ is the states of the neighborhood $\mathcal{N}_i$. RF-Pos is a technical requirement for deriving some important theoretical properties of the random fields and can be easily fulfilled in modeling, whereas RF-Mar is concerned with the structural dependency of the random fields. Under this latter property, only neighboring states have direct interactions; the conditional probabilities of the field globally determine the joint distribution. Markovianity is therefore central to the class of random fields that are decomposable into local components with a joint distribution factorizable as a product of functions defined over the neighborhood system. The terms 'undirected graphical models' and 'random fields' hereafter refer to those models that follow these properties, unless otherwise is specified.

From the modeling point of view, random fields can be viewed as normalized energy-based models. An energy-based representation captures dependencies by associating a scalar function, oftentimes referred to as an energy function, or simply energy, to each configuration of variables. These energy functions can take on many different forms, measuring the compatibility of observations and

[6] A neighborhood $\mathcal{N}_i$ of the site ς_i is the set of sites such that (1) $\varsigma_i \notin \mathcal{N}_i$, i.e., no site ς_i is its own neighbor, (2) for any site ς_j, $\varsigma_i \in \mathcal{N}_j \iff \varsigma_j \in \mathcal{N}_i$, that is, the neighborhood relations are mutual, and (3) for any $\varsigma_j \in \mathcal{N}_i$, there exists an edge in $\mathcal{E}$ that connects ς_i and ς_j. A subset C of sites is referred to as a clique if two distinct members are neighbors, i.e., $\varsigma_i, \varsigma_j \in C \implies \varsigma_i \in \mathcal{N}_j, \varsigma_j \in \mathcal{N}_j$; a clique system, $\mathcal{C} = \cup C$, is the set of cliques.

interpretation in arbitrary units and scales. In addition to the obvious reasons that require local energies to build a global representation, different energies corresponding to different sources of observation, descriptions and semantics are combined to encode a complex scene in semantic-based image modeling. Transforming energies into probability distributions provides a consistent way to build complex energy models based on local energies. That is, an energy is mapped to the interval $(0, 1]$ such that it is summed to one over all possible configurations in the output space.

2.2.2 Gibbs random fields: normalized energy-based models. It is common to cast the normalized energy in an exponential form, known as Gibbs (or Boltzmann) distribution,

$$pr(\mathcal{X}) = \frac{1}{Z} \exp\Big\{ -\beta \sum_{C \in \mathcal{C}} \Psi_C\big(\mathcal{X}_C\big) \Big\}, \tag{1}$$

where $\mathcal{X}_C$ is the set of sites in the clique C, Ψ_C is the energy function defined over the clique, $\beta = \frac{1}{\tau}$ is a scaling factor, known as the inverse temperature[7], and the partition function,

$$Z = \sum_{x \in \mathcal{X}} \exp\{-\beta \sum_{C \in \mathcal{C}} \Psi_C(x)\}, \tag{2}$$

[7]The inverse temperature β controls the confidence of belief regarding probable occurrence of the states. All configurations tend to be similarly likely at a low temperature, that is, a high inverse temperature, β. At the limit, $\lim_{\beta \to \infty} pr(\mathcal{X}) = \frac{1}{|\mathcal{X}|}$, that is, all possible configurations are equally likely. As the temperature τ drops, the mass of density gets concentrated around a finite number regions, or global energy minima, corresponding to the set of most likely configurations. In many models, the temperature τ is set to 1.

is the normalizing factor. A random field with a Gibbs distribution is called a Gibbs random field. An energy, $\Psi = \sum \Psi_C$, is defined for each possible configuration in such a way that lower energies are associated with more probable configurations.

An energy function of each configuration, given in Eq. (1) is a linear combination of local energies defined for neighboring sites in a clique. Global energy is expressible in terms of local energies encoding important contextual constraints. Although the range of interaction can be adapted by the choice of cliques, in practice, most approaches to image modeling are restricted to short range interactions defined by unary and pairwise energy functions due to the complexity of inference with larger clique sizes. Very few restrictions are placed upon the forms of these local energies, where complex energy models can be generated by combining simpler ones. It is easy to encode contextual constraints originating from different sources of information and analysis. More importantly, the resulting random fields are characterized by the Markovian structure defined by the RF-Mar property according to the Markov-Gibbs equivalence theorem, often known as Hammersley-Clifford theorem: A random field characterized by the properties RF-Pos and RF-Mar with respect to a neighborhood system $\mathcal{N}$ of a graph $\mathcal{G}$ defined over $\mathcal{S}$ if and only if it is a Gibbs random field with respect to $\mathcal{N}$. Note that a random field according to Markovianity is defined in terms of the conditional probability while the definition of a Gibbs random field is based on the joint distribution. By establishing the equivalence between the two formulations, the theorem also

establishes the connection between the probabilistic models and energy based models. One can therefore model an image ensemble in terms of a set of local clique energies, knowing that a joint distribution can be expressed in a globally consistent way according to Eq. (1).

Stochastic field theory thus significantly expands the scope of admissible probabilistic models that has traditionally been restricted to a number of standard families of parametrized distributions. Flexibility is gained for describing the essential characteristics of the ensemble by capturing local behaviors through the clique energies. However, the convenience of modeling comes at the expense of the complexity of inference. In contrast to the standard families of parameterized distributions, the energy functions and consequently the loss functions[8] required for an arbitrary image are usually characterized by complex surface topologies with many extrema, both global and local. The normalizing factor, i.e., the partition function involves the sum of energies that contain the model parameters over all configurations. As a consequence, parameter estimation and model selection require the evaluation of this function for different realizations of the parameters. The evaluation is however intractable for any nontrivial state space. Since Geman and Geman [20], much research has been motivated by this problem to reduce the computational resources expended on evaluating the function or for some special model, to avoid it all together. Greater elaboration on this problem is provided in Section 3.

[8]For a discussion of loss functions, see Section 3.

2.2.3 Generative random fields. The application of random fields to image modeling is traditionally associated with the probabilistic generative framework[9] that seeks to model the joint probability of the observed data and image labels. The recent proliferation of conditional random fields in object class segmentation highlights the paradigmatic shift towards a more direct approach to the problem, building interpretations around an explicit definition of the conditional probabilistic model, $pr(\mathcal{A}|\mathcal{D})$, instead of a generative model as a whole. In general, random fields are widely adopted for modeling spatial interaction. Contextual constraints that characterize configurations of visual patterns are usually encoded by the Gibbs energy with unary and pairwise potentials, $\Psi(x) = \sum_{\{i\}\in\mathcal{C}_1} \Psi_1(x_i) + \sum_{\{s,t\}\in\mathcal{C}_2} \Psi_2(x_s, x_t)$. The unary energy encodes the compatibility of an object class membership to the local measurement at the site, usually according to some class-specific models [44, 26]. The simplest variant of the pairwise energy term, Ψ_2, used for binary decision is the well known Isling model given by $\Psi(x) = \sum_{\{s,t\}} \beta_{s,t} x_s x_t$, where $\beta_{.,.}$ are interactive coefficients. This model is usually used to bias for interpretations that give coherent foreground and background instances[44]. In multi-state labeling problems, highly fragmented solutions are discouraged by the Potts model $\Psi(x) = \sum_{\{s,t\}} \beta_{s,t} \mathbf{1}_{\{x_s = x_t\}}$, where $\mathbf{1}_{\mathcal{A}}$ is the indicator function over the

[9] Generative random fields have been widely deployed as a modeling framework in many visual applications; for related discussion, see [20, 4, 12, 83]. This class of generative models is usually referred to in the literature of image modeling as the Markov random field, a term that is used among the researchers of probabilistic graphical models to cover all the undirected graphical models, generative or otherwise. To avoid confusion, this paper adopts a less ambiguous term "generative random fields" to refer the restricted class of probabilistic generative models defined over undirected graphs.

set $\mathcal{A}$ [91]. Common among the probabilistic approaches to object class segmentation are homogeneous and isotropic models where the energy terms are independent of location. The homogeneous Isling model, for instance, is the special case with the interactive coefficients $\beta_{s,t}$ set to a constant. The energy form given by $\Psi(x) = \sum_{\{s,t\}} -\beta_{s,t}|x_s - x_t|^k$ represents a more general measure of compatibility among multi-state configurations [81, 56]. These smoothness priors can be interpreted as some form of regularization [57, 58, 59].

Prior models in the generative framework represent belief before observation and therefore incorporate no empirical terms. This complete separation of semantics and observation has been challenged in recent years. According to this view, state compatibility between sites may be better decided conditionally on local image structures in many visual tasks [46]. For instance, state transitions may be anticipated across sites where discontinuities in some feature spaces, such as color, texture, and intensity, are observed. Spatial discontinuities can be directly modeled, such as introducing hierarchical fields which represent the dual stochastic processes of intensity and line [20]. The constraint is relaxed in the discriminative approaches; their direct focus on the conditional probability models makes it possible to incorporate empirical measurement in the random fields of spatial configurations. The data-dependent interaction can be encoded by an energy form with an empirical term, which either modulates the pair-site interactions [81] or their coefficients [26, 56]. In the latter case, the energy is no longer homogeneous and isotropic, that is, the interactive coefficients, $\beta_{s,t}$, vary by sites with magnitude depending on the local measurement

of image structures. Similarly, observation can also be allowed to modulate the unary energy term. For example, the empirical distribution of states observed from the training set may be introduced to ensure approximate match in the actual distribution of labels [26]. This energy term may be viewed as an approximation of KL entropy between the distribution of states in the image ensemble and the image distribution in a particular interpretation[10].

The observation model, $pr(\mathcal{D}|\mathcal{A})$, a forward model involving stochastic processes, represents the mapping of the underlying states to empirical measurement. The degradation model, taking a classical example in image restoration for illustration, is a nonlinear transformation of image intensity corrupted by a Gaussian noise process [20]. From the generative point of view, the observation model captures our understanding of the underlying causes of observation and provides a very important test for its explanatory and predictive power. Image synthesis by drawing random samples from the model provides important evidence of how well they resemble the observed signal generated in similar conditions [60]. Image interpretation is therefore not solely dictated by what is observed but also taking into account the theoretical and semantic properties of a scene [15]. To this end, it is common for observation models to incorporate filter responses, classifiers and descriptors, for instance, mixture models, histograms, shape descriptors, principle components and others, to capture the class-specific patterns underlying the

[10]The incorporation of filter responses in the energy terms that govern the configurations of patterns were first introduced in FRAME for texture modeling [88, 90, 57, 89].

image generation [81, 44, 73]. For mathematical simplicity and computational tractability, observation models usually assume a factorized form given by $pr(\mathcal{D}|\mathcal{A}) = \prod_{i \in \mathcal{S}} pr(\mathcal{D}_i|\mathcal{A}_i)$. That is, observed data are assumed to be conditionally independent given the labels. Suppose a prior model represented by a random field, Ψ, for instance, $\Psi = \sum_{C_1} \Psi_1 + \sum_{C_2} \Psi_2$, that specifies the spatial dependencies among image labels. The posterior model is then given by $pr(\mathcal{A}|\mathcal{D}) = \frac{1}{Z} \exp\{\sum \log pr(\mathcal{D}_i|\mathcal{A}_i) + \Psi\}$, which is also a random field. The theoretical properties of the prior model are preserved after the belief is updated with the observation as long as the independence assumption holds for the generating processes. A broad range of probabilistic models is admissible for approximating the generating processes, allowing a very flexible way to represent the class-specific causes of observation.

2.2.4 Conditional random fields. The conditional approaches provide an alternative for those who find the limitations of generative models to be too stringent for semantic-based image analysis. Interpretation via image labeling is the goal of image modeling and inference. The generative approaches expend too much effort on modeling data generating processes which is, for the most part, irrelevant to predicting image labels. In many cases, these processes are unknown or impractical to specify explicitly and formally. Furthermore, the conditional independence assumption is too restrictive for many semantic-based visual tasks as data are usually contextually dependent. The explanatory and predictive power of the model is significantly curtailed by rendering inadmissible those descriptors that seek to capture interactions over

longer ranges or across scales [45, 25, 63, 46]. These considerations motivate a new class of random fields, known as conditional random fields, which were first introduced in sequence data analysis [47] and subsequently adapted for the task of object class segmentation.

According to the conditional principle, the random fields represent the conditional probability, $pr(\mathcal{X}|\mathcal{D})$, defined over an undirected graph. Conditional on the data, $\mathcal{D}$, the probability distribution of system states, $\mathcal{X}$, is characterized by RF-Pos and RF-Mar with respect to the neighborhood system $\mathcal{N}$. The definition of the new class addresses many issues raised against their generative counterparts. Only the conditional model, $pr(\mathcal{X}|\mathcal{D})$, is specified explicitly. The image labels may be inferred directly from the input pattern; thus, the system states, $\mathcal{X} = \mathcal{A}$, consist of only the unknown assignment of labeling sites. For image models with hidden variables, H, the system states, $\mathcal{X} = (\mathcal{A}, H)$, are defined over the product space of assignment and hidden processes. For simplicity and clarity of exposition, we consider for the moment the cases without hidden variables. Extensions to the other cases are conceptually straightforward. According to the Markov-Gibbs equivalence (also known as Hammersley-Clifford) theorem, the conditional model can be directly expressed in terms of a normalized energy field: $pr(\mathcal{X}|\mathcal{D}) = \frac{1}{Z}\exp\{\Psi(\mathcal{X}|\mathcal{D})\}$. The common definition of model energy[11] involves only unary and pairwise interaction terms, formally,

[11] Different visual tasks may motivate other variants of this basic form extended to incorporate, for example, hidden variables or other neighborhood systems (for examples, see [26, 34, 82, 29]). Alternatively, the conditional model may also take on a conditional form of product of experts in the form: $pr(\mathcal{X}|\mathcal{D}) = \frac{1}{Z}\prod_\alpha pr_\alpha^{\gamma_\alpha}(\mathcal{X}|\mathcal{D})$, where the parameters, γ_i, control the weights of the corresponding experts in the final decision. (for example, see [25].)

$\Psi(\mathcal{X}|\mathcal{D}) = \sum_{\{i\}\in\mathcal{C}_1} \Psi_1(\mathcal{X}_i|\mathcal{D}) + \sum_{\{s,t\}\in\mathcal{C}_2} \Psi_2(\mathcal{X}_s, \mathcal{X}_t|\mathcal{D})$. In any case, the Gibbs energy associated with each $\mathcal{X}_i$ is defined over all observation $\mathcal{D}$ as well as the system states, $\mathcal{X}_{\mathcal{N}_i}$ in the local neighborhood. Without having to assume conditional independence in the observed data, conditional random fields significantly extend the model space by representing the conditional behavior of system states. Since conditional random fields preserve the stochastic properties of undirected graphical models, both the theory and algorithms developed in the latter context are also applicable to the new class of conditional models. Conditional random fields can be interpreted as a network of pattern predictors communicating with each other to exchange information in labeling decision [34]. Many object class specific structures can be captured through a broad range of experts, such as neural and linear classifiers, textons, boundary detectors, shape descriptors, spatial maps, appearance fragments, to name just a few; for examples, see [25, 64, 26, 34, 56, 68, 82, 29, 66].

3 Probabilistic Inference

The idea of visual segmentation as probabilistic inference is central to the probabilistic approaches to object class segmentation across different frameworks. This section provides some background information on probabilistic inference as applied in the context of object class segmentation. In general, probabilistic approaches involve two distinct and yet related problems of inference, viz, interpretation selection and model selection. The purpose of interpretation selection in visual segmentation is to identify the most plausible interpretation, $\hat{\mathcal{A}}$, with respect to the posterior model, $pr(\mathcal{A}|\mathcal{D})$, with the observation $\mathcal{D}$ fixed, according to some selection (i.e., optimization) criteria. It relies on the model representing the inference problem adequately. In the literature, model structures are fixed in the definition of modeling frameworks. Model selection is therefore concerned with identifying from a family, the members with behavior that is most compatible with a given visual task[12]. The probabilistic approaches consist of a wide variety of stochastic models of varying complexity, which may be built upon layers of simpler models or classifiers, and which capture different aspects of the segmentation problem. By their unique constructions and constraints, these models dictate the choice of inference techniques. A comprehensive account of the individual techniques employed in the literature could easily turn into a compendium on statistical inference and machine learning. Rather, the following discussion is restricted

[12]These processes are often referred to in the literature as learning probabilistic models from data.

to a general review of the key concepts, common conceptual principles and major classes of probabilistic inference that are relevant to the search for an optimal interpretation of a natural image.

3.1 Inference of Probabilistic Models[13]

The purpose of inference under the probabilistic generative framework is to revise the knowledge of some unknown quantities on the basis of observation. The challenge of the task involves the computation of a posterior probability distribution over the space of unobservable variables, $\mathcal{U}$, the subject of interest in various stages of the search for the optimal interpretation of the observed data. These distributions encode the knowledge and belief about the unknown quantities (or hidden variables) in light of past and present observations. The problem of inference corresponds to identifying the conditional distribution of the current configuration of observed and hidden variables given the configurations corresponding to the sequence of past observations. Probabilistic models of object class segmentation capture the prior knowledge and the information of past observation. If a generative model, $\mathcal{M}$, admits a representation of the generative processes and the prior belief of plausible interpretation, the corresponding posterior conditional on the observation, $\mathcal{D}$, is given by the Bayes'

[13]It should be noted that inference problems may involve continuous or discrete variables. The following discussion is equally applicable to both cases. When "information pooling" is involved, as in the case for normalization, marginalization and others, the equations for only the discrete case are given for the sake of compactness. Unless otherwise mentioned, they can be easily translated into the corresponding discrete version by replacing the summation with the integration over the relevant space.

rule [19, 13, 2, 55, 6]:

$$pr(\mathcal{U}|\mathcal{D},\mathcal{M}) = \frac{pr(\mathcal{D}|\mathcal{U},\mathcal{M})pr(\mathcal{U}|\mathcal{M})}{pr(\mathcal{D}|\mathcal{M})}. \tag{3}$$

The prior expectation, $pr(\mathcal{U}|\mathcal{M})$, represents the knowledge of the unknown variables before taking into account the present observation. The likelihood (or observation) model, $pr(\mathcal{D}|\mathcal{U}.\mathcal{M})$, which captures the stochastic processes that give rise to the observation, is a measure of how well the model predicts the observed data. The evidence, also known as marginal density of the data, $pr(\mathcal{D}|\mathcal{M})$, is a normalization constant. The posterior, $pr(\mathcal{U}|\mathcal{D},\mathcal{M})$, can be viewed as the inversion of the observation model, $pr(\mathcal{D}|\mathcal{U}.\mathcal{M})$, or updated belief or expectation from the prior in light of the new observation. In either case, it represents what is known about the hidden variables after all measurements (or observation) are taken into account. It is essential to distinguish two sets of unknown variables, $\mathcal{U} = H \cup \Theta$, that constitute the interest of inference in object class segmentation, where H denotes the hidden states of the world and Θ is the set of the model parameters given the model family, $\mathcal{M}$. The most important quantities (variables) of interest in H for object class segmentation are the assignment, $\mathcal{A}$, of object class membership, that constitutes the interpretation of an image. In the context of object class segmentation, the configurations of image labels, $H_{\mathcal{A}}$, define image interpretations in terms of object class membership. In general, $H = H_{\mathcal{A}} \cup H_{\mathcal{O}}$, where $H_{\mathcal{A}}$ is the set of variables associated to segmentation labeling, whereas $H_{\mathcal{O}} = H \setminus H_{\mathcal{A}}$ is the

set of other hidden variables. These hidden variables (of the latter subset) include the intermediate, explanatory variables deployed to explain observation, such as those related to, for instance, object part labeling [44, 43, 82], pattern descriptors [73, 81], variability of visual appearance vis-à-vis some canonical representation [81, 73, 82, 29], object localization [81, 54, 52, 91] and others. The second class of quantities, Θ, of inference arises from the modeling techniques adopted in probabilistic approaches to image segmentation. It is common to define a family of probabilistic models in terms of parametrized functions. Thus the common structure, $\mathcal{M}_{\mathcal{S}}$, can be conveniently expressed in terms of functional forms and the specific "shape" of individual models of the family via parameter specification. Well known parametric models such as the exponential families [19, 31, 79, 70] and mixture models [19, 6] are common tools for capturing the stochastic properties of visual processes in object class modeling. It is also common to express the conditional distributions of more complex processes in terms of parametrized energy functions in the graphical models. In general, the exact configuration of the model parameters cannot be determined *a priori*. Heuristics including trial and error may be involved to tune the model. Probabilistic approaches primarily follow a more recent alternative which sees the model parameters as unknowns and infers the optimal configuration from experimental results or training sets empirically. Thus, model parameters are represented by a set of random variables [3, 17]. Probabilistic models or distributions defined over the parameters capture belief about their plausible configurations, encoding prior knowledge and all infor-

mation derived from past observation.

Consider the interpretative model as given by

$$pr(H|\mathcal{D}, \mathcal{Z}_{\mathcal{T}}, \mathcal{M}_{\mathcal{S}}) = \sum_{\Theta} pr(H|\mathcal{D}, \Theta, \mathcal{M}_{\mathcal{S}}) pr(\Theta|\mathcal{Z}_{\mathcal{T}}, \mathcal{M}_{\mathcal{S}}) \tag{4}$$

where $\mathcal{Z}_{\mathcal{T}}$ denotes the knowledge given by past observation, experimental results or interpretative exemplars. In many cases, interpretation selection is concerned only with the configuration of image labeling. The conditional distribution of image labeling, $pr(H_{\mathcal{A}}|\mathcal{D}, \mathcal{Z}_{\mathcal{T}}, \mathcal{M}_{\mathcal{S}}) = \sum_{H_{\mathcal{O}}} pr(H_{\mathcal{A}}, H_{\mathcal{O}}|\mathcal{D}, \mathcal{Z}_{\mathcal{T}}, \mathcal{M}_{\mathcal{S}})$, captures what is known about the plausible interpretation of an image. In the absence of any intermediate latent variables, that is, $H_{\mathcal{O}} = \emptyset$, the parameter model, $pr(\Theta|\mathcal{Z}_{\mathcal{T}}, \mathcal{M}_{\mathcal{S}})$ depends upon the training data only. This allows a simple solution to the inference problems for many probabilistic approaches [25, 64, 68, 82, 66]. That is, a model can be selected prior to inferring the model parameters independently of the image labeling.

In the presence of intermediate, latent variables, it is common to compute the optimal configurations of hidden variables and model parameters either under some iterative framework [44, 43, 81, 26] or with some stochastic search procedure [65, 73, 72, 74] using Monte Carlo methods. Iterative solutions

[13] As previously mentioned, if the model, $\mathcal{M}$, admits a representation of the generative processes and the *a priori* belief of plausible configurations in the hidden variables, the conditional model can be obtained through Bayes' law:

$$pr(H|\mathcal{D}, \Theta, \mathcal{M}_{\mathcal{S}}) = \frac{pr(\mathcal{D}|H, \Theta, \mathcal{M}_{\mathcal{S}}) pr(H|\Theta, \mathcal{M}_{\mathcal{S}})}{\sum_{H} pr(\mathcal{D}|H, \Theta, \mathcal{M}_{\mathcal{S}}) pr(H|\Theta, \mathcal{M}_{\mathcal{S}})}$$

find their inspiration in the framework of the expectation maximization (EM) algorithm, a two-stage optimization technique for solving the inference problems with hidden variables [14, 19, 6]. Starting with an initial configuration of model parameters, the algorithm updates the estimates interactively until convergence in two steps: (1) the E-step computes the conditional distribution of hidden variables given the observed data and the current estimate of the parameters, and (2) the M step re-estimates the parameters by maximizing the expected complete data log likelihood model with respect to the conditional distribution given by the E-step. This strategy forms the basis of many other variants developed to cope with the challenge of complex models [14, 19, 61, 17]. In contrast, stochastic solutions rely on a class of stochastic sampling algorithms. The model parameters and the unknown states are inferred from a correlated sequence of samples generated from the first order Markov process where the stationary distribution of the process, $pr_{\Theta}(\Theta|\mathcal{Z}_{\mathcal{T}})$, converges to the target distribution $pr(\Theta|\mathcal{Z}_{\mathcal{T}})$, independent of the initial configurations. Many approaches integrate some discriminative/classification steps to approximate the probabilistic models based on local features and thus to restrict the search space of inference to the most plausible solution regions for tractability and efficiency [23, 89, 44, 73, 43, 72, 74] . To provide coarse regions of plausible solutions or approximations to the inference problems, these discriminative models are trained separately from the training data.

3.2 Optimization Problems of Natural Scene Labeling

The primary interest of interpretation selection in the context of object class segmentation is an optimal assignment that incurs the least expected risk due to misclassification, according to a probabilistic model. A similar decision may be required in model selection with respect to the model parameters that provide the best explanation of the training set. The decision can be formally viewed as a decision rule, $z \longrightarrow x$, from the observed to an estimate of the unobserved according to some optimality criteria. These criteria are defined as a loss function (also known as a cost function), $L(x, \hat{x}) > 0$, which measures the cost of estimating a true value x by $\hat{x}$, and by convention, $L(x, x) = 0$ [80]. The optimal decision rule aims to minimize the total loss incurred by a decision [6]. In the absence of *a priori* knowledge of the true value x, the optimal solution is the one that minimizes the average loss with respect to the conditional probability of the true value given an observation; formally, $\hat{x}_{opt} = \arg\min \mathbb{E}[L(x, \hat{x})|z]$, where the *a posteriori* expected loss is defined as $\mathbb{E}[L(x, \hat{x})|z] = \int L(x, \hat{x}) pr(x|z) dx$. Therefore the optimal decision rule depends on the definition of the loss function. The " 0/1" loss function[14], for instance, leads to the maximizer of the *a posteriori* mode and its additive version[15] yields the maximizer of the posterior marginals (MPM) [80]. These

[14]The "0/1" loss function is defined as $L_\epsilon(x, \hat{x}) = \mathbf{1}_{|x-\hat{x}|\geq\epsilon}$, for some $\epsilon > 0$, for scalar continuous variables and $L(x, \hat{x}) = \mathbf{1}_{|x\neq\hat{x}|}$ for scalar discrete variables, where $\mathbf{1}_A$ maps to one if A is true and zero otherwise.

[15]The additive " 0/1" loss function is the decomposable function defined as $L(x, \hat{x}) = \sum_{i=1}^{M} L_i(x_i, \hat{x}_i)$, where each summand L_i is a " 0/1" loss function.

decision rules are generally adopted in both interpretation and model selection in object class segmentation[16]. For other loss functions, see [55, 50].

Challenges often arise in computing the inference models which usually involve components, such as computing marginal distributions[17], and normalization, that are not always tractable due to their non-local scope of information consideration, [19, 13, 76]. Observation models or likelihoods are marginal probabilities of the observed data, and thus their computation is a special case of marginal probability computation. Computing the conditional distribution, $pr(\mathcal{X}_\alpha|\mathcal{X}_\beta)$, for disjoint subsets, $\mathcal{X}_\alpha$ and $\mathcal{X}_\beta$, of the state space also involves marginalization. Normalization, a general problem in cases where the posterior involves a partition (normalization) function, requires integrating the energy responses to every configuration in the solution space. In general, the partition function is a function of the model parameters, where all parameters are interdependent. Computation of this type involves integrating information over an exponential number of configurations and is therefore generally intractable in undirected graphical models [78, 17, 39].

In some special cases, the complexity of the problem can be reduced by exploiting the probabilistic structures of specific models. If a probabilistic graphical model can be described with a tree, for instance, each node and its descendants induce a subgraph which is also a tree. This restricted class of tree-structured

[16]These rules are discussed for the cases of object class labeling, where $\hat{\mathcal{X}}^{MAP} = \hat{\mathcal{A}}^{MAP}$, and $\hat{\mathcal{X}}^{MPM} = \hat{\mathcal{A}}_i^{MPM}$.

[17]The marginal distribution, $pr(\mathcal{X}_\alpha)$, is defined over a subset $\mathcal{X}_\alpha \in \mathcal{X}$ of the state space or space of unobserved variables.

distributions allows breaking down an inference problem over a graph into a set of problems over the subgraphs recursively. Efficient inference can therefore be implemented by means of exchanging information among local neighboring nodes via message passing using dynamic programming of belief propagation [62, 84, 79, 77, 42, 6, 76]. Ignoring the existence of loops, one may iterate a belief propagation procedure to graphs with cycles until convergence – hence these solutions are generally referred to as 'loopy belief propagation'. Although supported by many successful applications, loopy belief propagation may not converge to a stable equilibrium, and the resulting marginals may not be accurate [62, 84]. The clustering inference algorithm is a logical extension of belief propagation to graphs with cycles, which transforms the model into a probabilistically equivalent polytree through node merging. A methodical and efficient way to implement this approach for probabilistic inference is the junction tree algorithm, which projects the underlying graph of the model into a clique tree subject to constraints of consistency[18] across cliques; for details, see [48, 33, 77, 32, 42, 76]. However, the computational complexity of the algorithm is exponential in the size of the maximal cliques in the junction tree, and thus it is feasible only for models with small cliques. For nontrivial inference problems, the tree algorithm is generally prohibitively complex due to enormously large state cardinalities [77, 76].

[18]These constraints are required since a given node in the model may appear in multiple cliques. It is necessary to ensure that the assignment of marginals to these nodes in different cliques is consistent.

3.3 Approximations of Inference Models

Many inference models used in probabilistic approaches cannot be reduced to these special cases. It is computationally impractical to produce an exact inference due to the dimensionality of the solution space or due to the complex functional forms of the distributions. Instead, a variety of approximation schemes are adopted to provide approximations to the exact solution. These schemes fall naturally into two categories: stochastic and deterministic distribution approximations.

3.3.1 Stochastic distribution approximations The first class of approximations maps a target distribution to a randomly-generated approximation via numerical sampling of the distribution [6, 69]. Markov Chain Monte Carlo (MCMC) is a framework commonly adopted in many inference problems of visual segmentation to sample from a complex distribution of high-dimensional data. The basic idea of the approach is to generate samples by exploring the state space in a Markov chain process that allows the chain to realize the target model distribution as its limiting distribution. After a partial realization of the limiting distribution, the chain spends more time in the highly probable regions and consequently generates a sample which mimics the one drawn from the target distribution [1, 5, 6, 80]. The Metropolis-Hastings (MH) algorithm and its many variants have become the best known sampling strategies for constructing a convergent chain. Many inference problems of object class segmentation of natural images dwell in a varying dimensional space. Except for some special cases, there is no *a priori* knowledge about the

number of distinct regions in an image. A scene may be labeled in terms of different object classes, which are represented by diverse families of class models in terms of varying number of explanatory factors. As a consequence, these inference problems call for approximation schemes based on trans-dimensional sampling techniques. Reversible jump MCMC is a an extension of MCMC that constructs a reversible Markov chain according to a transition probability between subspaces of varying dimensions so that the chain with jumps across different subspaces converges to the target distribution at its equilibrium. As previously discussed, stochastic search looks for the most coherent interpretation by sampling different configurations of image partition. According to a set of transition kernels, image parsing, for instance, specifies the reconfiguration dynamics as a series of reversible Markov chain jumps governed by a set of graph transformation operators, which either change pattern models or node attributes, including the birth and death of visual patterns, splitting and merging of regions, pattern model switching and boundary evolution. The reversible Markov chain in the space of parsing graphs ensures that fair samples are generated from the invariant probability corresponding to the posterior model.

3.3.2 Deterministic distribution approximations quad Alternatively, deterministic approaches seek analytical approximations to a target distribution. Point estimators, such as the ML and MAP, may be viewed as the simplest and the limiting variants[19] of this general principle. Laplace ap-

[19]The point estimate of unknown variables or parameters, denoted by u, may be interpreted as an approximation of the distribution, $pr(u)$, by the function $\delta(u - \hat{u})$, where $\hat{u}$ is

proximation seeks a Gaussian approximation to a probability density defined over a set of continuous variables using the second order Taylor expansion about the modes of the function. These schemes may fail to capture significant, global properties due to their dependence on the local behavior of the density function around particular configurations [49, 6]. Consequently, they are not sufficient for the inference problems of natural scene segmentation due to the fact that any adequate representation of the problems generally involve complex probability distributions defined over discrete spaces.

Instead, many probabilistic approaches turn to a framework of deterministic distributional approximations, called variational inference, that are applicable to more general situations of probabilistic inference according to more global approximation criteria. Variational inference turns an approximation into an optimization problem by searching over a space of distributional forms for an optimal member that is close to the target distribution. It also admits an efficient evaluation scheme. The optimal approximation occurs at the minimum of the Kullback-Leiber or KL entropy between the approximation and the target distribution. With an unrestricted space, the variational free energy is maximized where the KL entropy vanishes, leading to an exact form of the target distribution. Due to computational concerns, it is often necessary to restrict the space to those distributional forms that admit an efficient evaluation. The common classes of approximation in natural scene segmentation are those of factorized distributional forms, which in one way or another

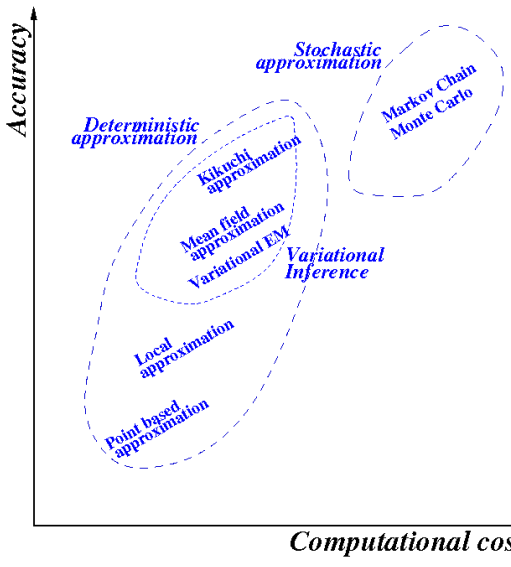
the chosen point estimate of u.

postulate some independence structures among the model components, i.e., independent groups of stochastic variables. Thus, individual components can be inferred independently or interactively; for details, see [6]. Underlying many inference algorithms used in object class segmentation is the reinterpretation of the EM algorithm [21, 61, 30, 2, 17, 6] and belief propagation [86, 30, 35, 71, 79, 84, 85, 76] as a variational approximation, with the standard algorithms being the special case when the approximation distribution equals the target distribution. Many approximation variants result from this analysis for large classes of distributions, where no exact solutions are possible under the standard frameworks.

3.3.3 Limitations of approximations in natural scene segmentation

The choice between stochastic approximation and variational inference is usually driven by an efficiency versus accuracy trade-off. Given adequate computational resources, stochastic sampling techniques can yield an approximation to any arbitrary level of accuracy but they can be computationally demanding. Inference, according to these schemes, may be too slow for many practical problems of signal and visual analysis. As a consequence, their application is usually limited to inference problems of small scales, otherwise the veridicality of representation is compromised under simplifying assumptions [73]. It is also non-trivial to determine whether the sample is generated from the target distribution – an underlying factor for the accuracy of approximation. The recent interest in variational inference is primarily motivated by the rising computational cost of inference using Monte Carlo methods. This alterna-

tive framework approaches the problem with analytical approximation to the target distribution and seeks a configuration that is reasonably close to the exact solution. The accuracy of approximation is largely limited by the choice of tractable distributions available for the approximation. Determining an appropriate family of distributions for approximation is non-trivial, and oftentimes the decision remains more art than science. In many cases, common approximations take advantage of the independence structures of the distributional forms for tractability and efficiency of evaluation. Thus, the choice of approximation family depends on how much dependence and correlation among the stochastic variables can be ignored with acceptable accuracy. Figure 1



Adopted from [69]

Figure 1: Accuracy versus computational cost trade off in distributional approximation.

illustrates the trade-off between the two approaches; for further details, see

[69, 76, 6].

This compromise is indeed not a unique problem of approximation inference. In visual segmentation, models are designed with computational issues in mind. By exploiting simplifying assumptions of independence for computational convenience, natural scene segmentation models are themselves, at best, approximations at the very core of the representation. One may expect that performance degrades when these models fail to account for some essential dependent and correlative structures of the underlying processes. Indeed, the independence assumptions used in many segmentation models are inconsistent with the general properties of a natural image. For instance, one may expect that color, intensity, texture, deformation fields or other filter responses should be highly correlated over an object-based subimage. The presence of redundant and strongly coupled structures in natural images stands in sharp contrast to the common assumption of cross-pixel independence; for instance see [81, 17]. Despite the many successes reported using these models, there is reasonable ground for questioning their generalizability when considering the myriad of possible scenes in their full complexity of visual appearance.

[19]Similar comparisons can be found in [69].

References

[1] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50:5–43, 2003.

[2] M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[3] J. Bernardo and A. Smith. *Bayesian Theory*. J. Wiley & Sons, Chichester, UK, 1994.

[4] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.

[5] Julian E. Besag. Markov chain Monte Carlo for statistical inference. Technical report, Center for Statistics and the Social Sciences, University of washington, 2004.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[8] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *IEEE Computer Society Workshop on Perceptual Organization in Computer Vision*, page 46, 2004.

[9] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *European Conference on Computer Vision*, page II: 109 ff., 2002.

[10] A. Bosch, X. Munoz, and J. Freixenet. Segmentation and description of natural outdoor scenes. *Image and Vision Computing*, 25(5):727–740, 2007.

[11] Tony Chan and Jianhong Shen. *Image Processing And Analysis: Variational, PDE, Wavelet, And Stochastic Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.

[12] Rama Chellappa, editor. *Digital Image Processing*. Institute of Electrical & Electronics Engineers, Inc, Los Alamitos, CA, 1992.

[13] Rizwan A. Choudrey. *Variational Methods for Bayesian Independent Component Analysis*. PhD thesis, Pattern Analysis and Machine Learning, Robotics Research Group, University of Oxford, 2002.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[15] Robert A. Dunne. *A Statistical Approach to Neural Networks for Pattern Recognition*. Wiley-Interscience, Hoboken, N.J., 2007.

[16] X. Feng, C. K. I. Williams, and S. N. Felderhof. Combining belief networks and neural networks for scene segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):467–483, 2002.

[17] Brendan J. Frey and Nebojsa Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1392–1416, 2005.

[18] Karl Friston. Functional integration and inference in the brain. *Progress in Neurobiology*, 68(2):113–143, October 2002.

[19] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Chapman & Hall, New York, 1995.

[20] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[21] Z. Ghahramani and M.J. Beal. Graphical models and variational methods. In *Advanced Mean Field methods - Theory and Practice*. MIT Press, 2000.

[22] Ulf Grenander and Michael I. Miller. *Pattern theory : from representation to inference*. Oxford University Press, Oxford ; New York, 2007.

[23] Cheng-En Guo, Song-Chun Zhu, and Ying Nian Wu. Modeling visual patterns by integrating descriptive and generative methods. *International Journal of Computer Vision,*, 53(1):5–29, 2003.

[24] Xuming He and Richard Zemel. Latent topic random fields: Learning using a taxonomy of labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[25] Xuming He, Richard Zemel, and Miguel Carreira-Perpinan. Multiscale conditional random fields for image labelling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II: 695–702, 2004.

[26] Xuming He, Richard S. Zemel, and Debajyoti Ray. Learning and incorporating top-down cues in image segmentation. In *European Conference on Computer Vision*, pages I: 338–351, 2006.

[27] Thomas Hofmann. Probabilistic latent semantic analysis. In *myUAI*, 1999.

[28] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *myML*, 42(1):177–196, 2001.

[29] D. Hoiem, Carsten Rother, and J. Winn. 3D layoutCRF for multi-view object class recognition and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[30] T. Jaakkola. Tutorial on variational approximation methods. In *Advanced Mean Field Methods: Theory and Practice*, pages 129– 159. MIT Press, Cambridge, MA, 2001.

[31] Tony Jebara. *Machine Learning: Discriminative and Generative*. Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[32] Michael I. Jordan. Graphical models. *Statistical Science*, 19:140–155, 2004.

[33] Michael I. Jordan and Yair Weiss. Graphical models: Probabilistic inference. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, 2nd ed. edition, 2003.

[34] A. Kapoor and J. Winn. Located hidden random fields: Learning discriminative parts for object detection. In *European Conference on Computer Vision*, 2006.

[35] H. Kappen and W. Wiegerinck. Mean field theory for graphical models. In Manfred Opper and David Saad, editors, *Adavanced Mean Field Theory – Theory and Practice*, pages 37–49. The MIT Press, February 2001.

[36] Zoltan Kato. *Modélisations markoviennes multirésolutions en vision par ordinateur. Application r' la segmentation d'images SPOT*. PhD thesis, INRIA, Sophia Antipolis, France, December 1994.

[37] D. Kersten and P. R. Schrater. Pattern inference theory: A probabilistic approach to vision. In *Perception and the physical world : psychological and philosophical issues in perception*. Wiley, Chichester, 2002.

[38] Daniel Kersten, Pascal Mamassian, , and Alan Yuille. Object perception as Bayesian inference. *Annual Review of Psychology*, 55:271–304, 2004.

[39] Vladimir Kolmogorov and Martin Wainwright. On the optimality of tree-reweighted max-product message passing. In *Uncertainty in Articial Intelligence*, pages 316–32, 2005.

[40] S. M. Konishi and A.L. Yuille. Statistical cues for domain specific image segmentation with performance analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 125–132, 2000.

[41] S.K. Kopparapu and U.B. Desai. *Bayesian Approach to Image Interpretation*, volume 616 of *The Springer International Series in Engineering and Computer Science*. Kluwer Academic Publishers, Boston, 2001.

[42] Kevin B. Korb and Ann E. Nicholson. *Bayesian Artificial Intelligence*. Chapman & Hall/CRC, New York, 2004.

[43] M. P. Kumar, P. H. S. Torr, and A. Zisserman. An object category specific MRF for segmentation. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 596–616. Springer, 2006.

[44] M. Pawan Kumar, Philip H. S. Torr, and A. Zisserman. OBJ CUT. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 18–25, 2005.

[45] Sanjiv Kumar and Martial Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *International Conference on Computer Vision*, pages II: 1150 – 1157, 2003.

[46] Sanjiv Kumar and Martial Hebert. Discriminative random fields. *International Journal of Computer Vision,*, 68(2):179–201, 2006.

[47] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289, 2001.

[48] Lauritzen, S. L. and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2): 157–224, 1988.

[49] Neil D. Lawrence. *Variational Inference in Probabilistic Models*. PhD thesis, University of Cambridge, 2000.

[50] Yann. LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. Tutorial on energy-based learning. In Gükhan H. Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S. V. N. Vishwanathan, editors, *Predicting Structured Data*, Neural Information Processing, pages 191–246. The MIT Press, 2007.

[51] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV Workshop on Statistical Learning in Computer Vision*, pages 17–32, 2004.

[52] B. Leibe, K. Mikolajczyk, and B. Schiele. Segmentation based multi-cue integration for object detection. In *British Machine Vision Conference*, page III: 1169, 2006.

[53] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *British Machine Vision Conference*, pages 759–768, Norwich, UK, Sept. 2003.

[54] B. Leibe and B. Schiele. Interleaving object categorization and segmentation. In *Cognitive Vision Systems – Sampling the Spectrum of Approaches*, Lecture Notes on Computer Science, 3948, pages 145–161. Springer, 2006.

[55] Jörg C. Lemm. *Bayesian field theory*. Johns Hopkins University Press, Baltimore, MD, 2003.

[56] Anat Levin and Yair Weiss. Learning to combine bottom-up and top-down segmentation. In *European Conference on Computer Vision*, pages 581–594, 2006.

[57] Stan Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[58] Jose Marroquin. *Probabilistic Solution of Inverse Problem*. PhD thesis, EECS, MIT, September 1985.

[59] Jose Marroquin, S Mitter, and Tomaso Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82(397):76–89, March 1987.

[60] David Mumford. Pattern theory: the mathematics of perception. In *International Congress of Mathematicians*, volume 1, 2002.

[61] Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

[62] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[63] Yuan Qi, Martin Szummer, and Thomas P. Minka. Bayesian conditional random fields. In *International Workshop on Artificial Intelligence and Statistics*, 2005.

[64] Xiaofeng Ren, Charless C. Fowlkes, and Jitendra Malik. Cue integration in figure/ground labeling. In *Advances in Neural Information Processing Systems*, 2005.

[65] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *International Conference on Computer Vision*, pages I: 10–17, 2003.

[66] Jordan Reynolds and Kevin Murphy. Figure-ground segmentation using a hierarchical conditional random field. In *Canadian Conference on Computer and Robot Vision*, 2007.

[67] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1605–1614, Washington, DC, USA, 2006. IEEE Computer Society.

[68] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision*, 2006.

[69] Václav Smídl and Anthony Quinn. *The Variational Bayes Method in Signal Processing*. Signals and Communication Technology. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[70] Erik B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, Department of Electrical Engineering and Computer Sceince. Massachusetts Institute of Technology, May 2006.

[71] Y. W. Teh. *Bethe Free Energy and Contrastive Divergence Approximations for Undirected Graphical Models*. PhD thesis, Department of Computer Science, University of Toronto, 2003.

[72] Zhouwen Tu, Xiangrong Chen, Alan L. Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. Technical report, Department of Statistics, UCLA. Department of Statistics, January 2005.

[73] Zhuowen Tu, Xiangrong Chen, Alan L. Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision,*, 63(2):113–140, 2005.

[74] Zhuowen Tu and Song-Chun Zhu. Parsing images into regions, curves, and curve groups. *International Journal of Computer Vision,*, 69(2):223–249, 2006.

[75] B. Verbeek, J.and Triggs. Region classification with Markov field aspect models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 – 8, 2007.

[76] M. Wainwright and M. Jordan. A variational principle for graphical models. In Simon Haykin, José C. Príncipe, Terrence J. Sejnowski, and John McWhirter, editors, *New Directions in Statistical Signal Processing: From Systems to Brains*, chapter 11. The MIT Press, Cambridge, Massachusetts, 2006.

[77] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on Information Theory*, 45(9):1120–1146, 2003.

[78] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. Tree-reweighted belief propagation and approximate ML estimation by pseudo-moment matching. In *International Workshop on Artificial Intelligence and Statistics*, 2003.

[79] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, UC Berkeley, Dept. of Statistics, September 2003.

[80] Gerhard Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*. Stochastic Modelling and Applied Probability. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[81] J. Winn and N. Joijic. LOCUS: Learning object classes with unsupervised segmentation. In *International Conference on Computer Vision*, pages I: 756–763, 2005.

[82] John Winn and Jamie Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 37–44, 2006.

[83] Chee Sun Won and Robert M. Gray. *Stochastic image processing*. Kluwer Academic/Plenum Publishers, New York, 2004.

[84] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, pages 239–269. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[85] J.S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, July 2005.

[86] J.S. Yedidia, W.T.Freeman, and Y. Weiss. Bethe free energy, Kikuchi approximations and belief propagation algorithms. In *Advances in Neural Information Processing Systems*, volume 13, December 2000.

[87] Alan Yuille and Daniel Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, July 2006. Special issue: Probabilistic models of cognition.

[88] Song Chun Zhu. *Statistical and computational theories for image segmentation, texture modeling and object recognition*. PhD thesis, Harvard University, 1996.

[89] Song-Chun Zhu. Statistical modeling and conceptualization of visual patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):691– 712, June 2003.

[90] Song Chun Zhu, Zing Nian Wu, and David Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9 (8):1627–1660, November 1997.

[91] Thomas Zöller and Joachim M. Buhmann. Robust image segmentation using resampling and shape constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1147–1164, July 2007.