Perturb-and-MAP Random Fields: Using Discrete Optimization to Learn and Sample from Energy Models

George Papandreou[1] and Alan L. Yuille[1,2]
[1]Department of Statistics, University of California, Los Angeles
[2]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
[gpapan,yuille]@stat.ucla.edu

Abstract

We propose a novel way to induce a random field from an energy function on discrete labels. It amounts to locally injecting noise to the energy potentials, followed by finding the global minimum of the perturbed energy function. The resulting Perturb-and-MAP random fields harness the power of modern discrete energy minimization algorithms, effectively transforming them into efficient random sampling algorithms, thus extending their scope beyond the usual deterministic setting. In this fashion we can enjoy the benefits of a sound probabilistic framework, such as the ability to represent the solution uncertainty or learn model parameters from training data, while completely bypassing costly Markov-chain Monte-Carlo procedures typically associated with discrete label Gibbs Markov random fields (MRFs). We study some interesting theoretical properties of the proposed model in juxtaposition to those of Gibbs MRFs and address the issue of principled design of the perturbation process. We present experimental results in image segmentation and scene labeling that illustrate the new qualitative aspects and the potential of the proposed model for practical computer vision applications.

1. Introduction

Discrete label Markov random fields (MRFs), going back to the classic Ising and Potts models in statistical physics, offer a natural and sound probabilistic modeling framework for a host of image analysis and computer vision problems involving discrete labels, such as image segmentation, texture synthesis, and deep learning [2, 7, 10, 15, 31]. Exact probabilistic inference and maximum likelihood model parameter fitting is intractable in general MRFs defined on 2-D domains and one has to employ random sampling schemes to perform these tasks [7, 10]. Beyond its role in inference, random sampling from MRFs can be a goal in itself when the generative MRF properties are exploited, as in texture synthesis or inpainting [24, 31]. However, Markov-chain Monte-Carlo (MCMC) sampling algorithms such as Gibbs sampling can be computationally too expensive for many practical computer vision applications.

Recent powerful discrete energy minimization algorithms such as graph cuts, linear programming relaxations, or loopy belief propagation [5, 15–17] can efficiently find or well approximate the most probable (MAP) configuration for certain important classes of MRFs and have had big impact on several computer vision applications. Beyond MAP computation, energy minimization algorithms can be used for estimating model parameters using max-margin criteria [26]. However, the deterministic viewpoint on MRF modeling as energy minimization problem has important limitations as it does not provide the right conceptual framework for probabilistically characterizing the solution uncertainty or learning the model parameters from training data by maximum likelihood.

In this work we attempt to somehow bridge the gap between the probabilistic and the energy minimization approaches to MRF modeling. We propose a novel way to induce a discrete label random field model from an energy function, which amounts to locally injecting additive random noise to the continuous energy potentials, followed by finding the global (approximate) minimum configuration of the perturbed energy function. This *Perturb-and-MAP* (PM) random field is a legitimate probabilistic model which delegates the non-trivial global interactions involved in sampling to an efficient energy minimization routine, and thus allows rapid sampling from a wide range of energy functions widely used in practice.

From the probabilistic MRF perspective, the proposed technique can be seen as a one-shot approximate random sampling algorithm that completely bypasses MCMC. We study the problem of designing the perturbation process so as the Perturb-and-MAP random field be a good approximation to the corresponding Gibbs MRF. Interestingly, we identify a specific perturbation density under which the Perturb-and-MAP model is identical to its Gibbs counterpart. Although this ideal perturbation is not practically applicable since it effectively destroys the local Markov structure of the energy, it suggests low-order perturbations that only introduce noise to the unary (order-1) or a subset of the pairwise (order-2) potential tables, resulting in perturbed energies that are effectively as easy to minimize as the original unperturbed one, while producing random samples virtually indistinguishable from exact Gibbs MRF samples.

Perturb-and-MAP random fields allow qualitatively new applications of energy minimization algorithms in computer vision. First, accompanying the MAP solution with several typical posterior samples drawn from the model allows us to quantify our confidence in the solution, which can be useful in guiding the user's attention in interactive applications, propagating uncertainty in further processing steps of a more complex computer vision pipeline, or assessing the generative properties of a particular MRF model. Second, our efficient sampling algorithm allows learning of MRF or CRF parameters using the moment matching rule, in which the model parameters are updated until the generated samples reproduce the (weighted) sufficient statistics of the observed data. This approach is very popular for learning of patch-based models [10, 19], but the use of perturbed sampling instead of contrastive divergence is crucial for fast training in our applications. Similar to Gibbs MRFs, such greedy parameter update is justified because the log-likelihood of the Perturb-and-MAP model turns out to be concave. We illustrate these ideas in experiments on image segmentation and scene labeling.

Related work Our research on the Perturb-and-MAP discrete random field model has been motivated by the exact Gaussian MRF sampling algorithm popularized by [20, 24] and especially its local factor perturbation interpretation highlighted by [20]. While the underlying mathematics and methods are completely different in the discrete setup we consider here, we show that the intuition of local perturbations followed by global optimization can also lead to powerful sampling algorithms for discrete label MRFs.

Herding [29] builds a deterministic dynamical system on the model parameters designed so as to reproduce the data sufficient statistics, which is similar in spirit to the moment-matching algorithm we use for learning. However, herding is still not a probabilistic model and cannot summarize the data into a concise set of model parameters.

The limitations of MAP-based inference in discrete MRFs are nicely illustrated in [30]. They impose extra global constraints in the energy minimization problem to mitigate the tendency of MAP inference to produce singular solutions. However, they still adhere to a deterministic setting which is not suited for parameter learning. Further, optimizing the resulting modified energy functions is far more challenging than minimizing the original energy.

Averaging over multiple samples, our approach allows efficiently estimating (sum-) marginal densities and thus quantifying the per-node solution uncertainty even in graphs with loops. Max-product belief propagation [28] and dynamic graph-cuts [14] can compute max-marginals, which give some indication of the uncertainty in label assignments [14] but cannot directly estimate marginal densities.

In the context of binary image segmentation, the sampling-based marginal confidence maps we produce resemble the soft segmentation maps of the random walker model [8], although the underlying probabilistic underpinnings of the two methods are completely different.

2. Energy functions and Gibbs MRFs

Our starting point is a *deterministic* energy function

$$e(\mathbf{x};\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) \rangle \,, \tag{1}$$

where $\mathbf{x} \in \mathcal{L}^N$ is a length-N state configuration vector with entries x_i in a discrete label set $\mathcal{L}$, $\boldsymbol{\theta} \in \mathbb{R}^M$ is a real parameter vector of length M, and $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))^T$ is a vector of potentials or "sufficient statistics". We can interpret θ_j as the weight assigned to the feature $\phi_j(\mathbf{x})$: we have many different design goals or sources of information (e.g., smoothness prior, measurements), each giving rise to some features. We merge everything together into a single objective function which we want to optimize so as to recover the best/minimum energy configuration $\mathbf{x}$.

The weights $\boldsymbol{\theta}$ are selected in a way that the model assigns low energies to desirable configurations and high energies to "everything else". When the number of parameters M is small, we can set them to reasonable values by hand. A more principled way is to learn the parameters from a labeled training set $\{\mathbf{x}_k\}_{k=1}^K$ by discriminative criteria such as structured max-margin [15, 19, 26, 27]. Computationally, one typically ends up with efficient iterative algorithms that require MAP inference at each parameter update step.

The Gibbs distribution is the standard way to induce a *probabilistic* model from the energy function $e(\mathbf{x};\boldsymbol{\theta})$. It defines a Markov random field whose probability mass function has the exponential family form

$$f_G(\mathbf{x};\boldsymbol{\theta}) = Z^{-1}(\boldsymbol{\theta}) \exp\left(-e(\mathbf{x};\boldsymbol{\theta})\right) , \tag{2}$$

where $Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}} \exp\left(-e(\mathbf{x};\boldsymbol{\theta})\right)$ is the partition function.

In the probabilistic setting, maximum (penalized) likelihood (ML) is the natural criterion for learning the weights. Given the labeled training set $\{\mathbf{x}_k\}_{k=1}^K$, we fit the parameters $\boldsymbol{\theta}$ by maximizing the log-likelihood function $L_G(\boldsymbol{\theta}) = -\log Z(\boldsymbol{\theta}) - (1/K)\sum_{k=1}^K e(\mathbf{x}_k;\boldsymbol{\theta})$, possibly also including an extra penalty term regularizing the weights. For fully observed models and energies of the form (1) the log-likelihood is a concave function of the weights $\boldsymbol{\theta}$ and thus the global maximum can be found by gradient ascent

[11, 15, 31]. The gradient is $\partial L_G/\partial\theta_j = E^G_{\boldsymbol\theta}\{\phi_j(\mathbf{x})\} - E_D\{\phi_j(\mathbf{x})\}$. Here $E^G_{\boldsymbol\theta}\{\phi_j(\mathbf{x})\} \triangleq \sum_{\mathbf{x}} f_G(\mathbf{x};\boldsymbol\theta)\phi_j(\mathbf{x}) = -\partial(\log Z)/\partial\theta_j$ and $E_D\{\phi_j(\mathbf{x})\} \triangleq (1/K)\sum_{k=1}^K \phi_j(\mathbf{x}_k)$ are, respectively, the sufficient statistics under the Gibbs model and the data. Upon convergence, $E^G_{\boldsymbol\theta}\{\phi_j(\mathbf{x})\} = E_D\{\phi_j(\mathbf{x})\}$. Thus, ML estimation of the Gibbs model can be thought of as moment matching: random samples drawn from the trained model reproduce the sufficient statistics observed in the training data.

The chief computational challenge in ML parameter learning of the Gibbs model lies in estimating the model sufficient statistics $E^G_{\boldsymbol\theta}\{\phi_j(\mathbf{x})\}$. Note that this inference step needs to be repeated at each parameter update step. The model sufficient statistics can be computed exactly in tree-structured (and low tree-width) graphs, but in general graphs one needs to resort to MCMC techniques for approximating them [10, 11, 31], an avenue considered too costly for many computer vision applications. Deterministic approximations such as variational techniques or loopy sum-product belief propagation do exist, but often are not accurate enough. Simplified criteria such as pseudo-likelihood [3] have been applied as substitutes to ML, but they can sometimes give results grossly different to ML.

Beyond model training, random sampling is very useful in itself, to reveal what are typical instances of the model – what the model has in its "mind" – and in applications such as texture synthesis [31]. Further, we might be interested not only in the global minimum energy configuration, but in the marginal densities or posterior means as well [24]. In loopy graphs these quantities are typically intractable to compute, the only viable way being through sampling. Our Perturb-and-MAP random field model is designed specifically so as to be amenable to rapid sampling.

3. Perturb-and-MAP random fields

We propose a novel way to induce a probabilistic model from an energy function:

Definition. *The* Perturb-and-MAP *random field is defined by* $\mathbf{x}(\boldsymbol\epsilon) = \operatorname{argmin}_{\mathbf{q}} e(\mathbf{q};\boldsymbol\theta+\boldsymbol\epsilon)$*, where* $\boldsymbol\epsilon$ *is a random real-valued additive parameter perturbation vector.*

In other words, we inject noise to the model parameters $\tilde{\boldsymbol\theta} = \boldsymbol\theta + \boldsymbol\epsilon$, followed by finding the least energy configuration $\mathbf{x}(\boldsymbol\epsilon)$ of the perturbed energy function. While Perturb-and-MAP random fields can also be built on energies defined over continuous labels [20], our focus in this paper will be on random fields over discrete labels.

The main motivation for defining a probabilistic model in such a way is that for certain energy functions $e(\mathbf{x};\boldsymbol\theta)$ there exist powerful algorithms which can find the MAP state efficiently. Thus, by construction, we can efficiently draw *exact* one-shot samples from the Perturb-and-MAP model without resorting to expensive MCMC techniques.

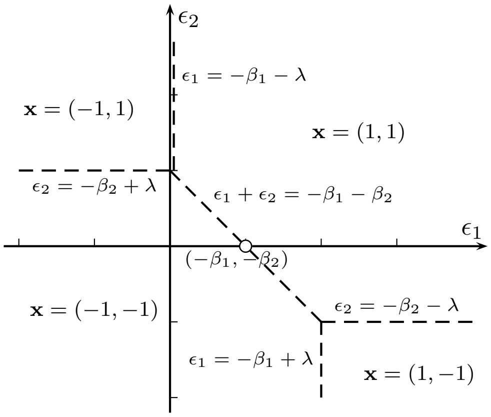
ϵ_2
$\epsilon_1 = -\beta_1 - \lambda$
$\mathbf{x} = (-1,1)$
$\mathbf{x} = (1,1)$
$\epsilon_2 = -\beta_2 + \lambda$
$\epsilon_1 + \epsilon_2 = -\beta_1 - \beta_2$
ϵ_1
$(-\beta_1, -\beta_2)$
$\mathbf{x} = (-1,-1)$
$\epsilon_2 = -\beta_2 - \lambda$
$\epsilon_1 = -\beta_1 + \lambda$
$\mathbf{x} = (1,-1)$

Figure 1. Perturb-and-MAP geometry under the Ising energy with $N = 2$ nodes and perturbations only in the unary terms, $\tilde\beta_i = \beta_i + \epsilon_i$, for parameter values $\beta_1 = -1$, $\beta_2 = 0$, and $\lambda = 1$. The $\boldsymbol\epsilon$-space is split into four polyhedra, with $\mathbf{x}(\boldsymbol\epsilon) = \mathbf{x}$ iff $\boldsymbol\epsilon \in \mathcal{P}_{\mathbf{x}} - \boldsymbol\theta$.

3.1. Weight space geometry

A particular state $\mathbf{x} \in \mathcal{L}^N$ will be minimizing the deterministic energy (1) if, and only if, $e(\mathbf{x};\boldsymbol\theta) \le e(\mathbf{q};\boldsymbol\theta), \forall \mathbf{q} \in \mathcal{L}^N$. This set of $|\mathcal{L}|^N$ linear inequalities defines a polyhedron $\mathcal{P}_{\mathbf{x}}$ in the weight space

$$\mathcal{P}_{\mathbf{x}} = \{\boldsymbol\theta \in \mathbb{R}^M : \langle \boldsymbol\theta, \boldsymbol\phi(\mathbf{x}) - \boldsymbol\phi(\mathbf{q})\rangle \le 0, \forall \mathbf{q} \in \mathcal{L}^N\}\,. \quad (3)$$

Actually, $\mathcal{P}_{\mathbf{x}}$ is a polyhedral cone [4], since $\boldsymbol\theta \in \mathcal{P}_{\mathbf{x}}$ implies $\alpha\boldsymbol\theta \in \mathcal{P}_{\mathbf{x}}$, for all $\alpha \ge 0$. The polyhedra $\mathcal{P}_{\mathbf{x}}$ split the weight space $\mathbb{R}^M$ into regions of influence of each discrete state $\mathbf{x} \in \mathcal{L}^N$. Under the Perturb-and-MAP model, $\mathbf{x}(\boldsymbol\epsilon)$ will be assigned to a particular state $\mathbf{x}$ if, and only if, $\boldsymbol\theta + \boldsymbol\epsilon \in \mathcal{P}_{\mathbf{x}}$ or, equivalently, $\boldsymbol\epsilon \in \mathcal{P}_{\mathbf{x}} - \boldsymbol\theta \triangleq \{\boldsymbol\epsilon \in \mathbb{R}^M : \boldsymbol\theta + \boldsymbol\epsilon \in \mathcal{P}_{\mathbf{x}}\}$. In other words, if a specific instantiation of the perturbation $\boldsymbol\epsilon$ falls in the shifted polyhedron $\mathcal{P}_{\mathbf{x}} - \boldsymbol\theta$, then the Perturb-and-MAP model generates $\mathbf{x}$ as sample.

We assume that perturbations are drawn from a density $f_{\boldsymbol\epsilon}(\boldsymbol\epsilon)$ which does not depend on the parameters $\boldsymbol\theta$. The probability mass of a state $\mathbf{x}$ under the Perturb-and-MAP model is then the weighted volume of the corresponding shifted polyhedron under the perturbation measure

$$f_{PM}(\mathbf{x};\boldsymbol\theta) = \int_{\mathcal{P}_{\mathbf{x}}-\boldsymbol\theta} f_{\boldsymbol\epsilon}(\boldsymbol\epsilon)d\boldsymbol\epsilon\,, \quad (4)$$

which is the counterpart of the Gibbs density in Eq. (2). It is intractable (NP-hard) to compute the volume of general polyhedra in a high-dimensional space; see, e.g., [1, p. 29]. However, for the class of perturbed energy functions which can be globally minimized efficiently, we can readily draw exact samples from the Perturb-and-MAP model, without ever explicitly evaluating the integrals in Eq. (4).

Example: Perturb-and-MAP Ising model Let us consider the Ising energy $e(\mathbf{x};\boldsymbol\theta) = \frac{-1}{2}\sum_{i=1}^N \big(\beta_i x_i +$

$\sum_{i'=i+1}^{N} \lambda_{ii'} x_i x_{i'}\big)$ over the discrete "spins" $x_i \in \{-1, 1\}$; here β_i is the external field strength ($\beta_i > 0$ favors $x_i = 1$) and $\lambda_{ii'}$ is the coupling strength (attractive coupling $\lambda_{ii'} > 0$ favors the same spin for x_i and $x_{i'}$). This energy function can be written in the standard inner product form (1) with $\boldsymbol{\theta} = (\{\beta_i\}, \{\lambda_{ii'}\})^T$ and $\boldsymbol{\phi}(\mathbf{x}) = \frac{-1}{2}(\{x_i\}, \{x_i x_{i'}\})^T$. The MRF defined by (2) is the Ising Gibbs random field.

Defining a Perturb-and-MAP Ising random field requires specifying the perturbation density. In this example, we leave the binary term parameters $\lambda_{ii'}$ intact and only perturb the unary term parameters β_i. In particular, for each unary factor, we set $\tilde{\beta}_i = \beta_i + \epsilon_i$, with ϵ_i IID samples from the logistic distribution with density $l(z) = \frac{1}{4}\operatorname{sech}^2(\frac{z}{2})$. This corresponds to the order-1 Gumbel perturbation we discuss in Sec. 4 and ensures that if a particular node x_i is completely isolated, it will then follow the same Bernoulli distribution $\Pr\{x_i = 1\} = 1/(1 + e^{-\beta_i})$ as in the Gibbs case. The $\boldsymbol{\epsilon}$-space geometry in the case of two labels ($N = 2$) under the Ising energy $e(\mathbf{x}; \boldsymbol{\theta}) = -0.5(\beta_1 x_1 + \beta_2 x_2 + \lambda x_1 x_2)$ for a specific value of the parameters $\boldsymbol{\theta}$ and perturbations only to unary terms is depicted in Fig. 1. We show in Fig. 2 some statistics comparing the Gibbs and Perturb-and-MAP random fields for a toy Ising energy involving 9 variables and randomly generated parameters. The probability landscape under the two models looks quite similar.

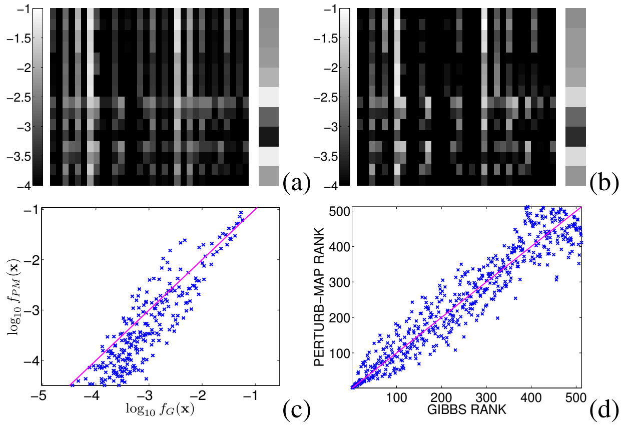

Figure 2. Ising energy on 3×3 grid, with β_i and $\lambda_{ii'}$ IID from $\mathcal{N}(0, 1)$. We compare the Gibbs (exact computation) and PM (10^6 Monte-Carlo runs) random fields. (a) $\log_{10} f_G(\mathbf{x})$ and $f_G(x_i = 1)$. (b) $\log_{10} f_{PM}(\mathbf{x})$ and $f_{PM}(x_i = 1)$. Scatter-plot of state log probabilities (c) and state ranking (d) under the two models.

3.2. Parameter estimation

We would like to estimate the parameters $\boldsymbol{\theta}$ of the Perturb-and-MAP model from a labeled training set $\{\mathbf{x}_k\}_{k=1}^{K}$ by maximizing the log-likelihood

$$L_{PM}(\boldsymbol{\theta}) = (1/K) \sum\nolimits_{k=1}^{K} \log f_{PM}(\mathbf{x}_k; \boldsymbol{\theta}) \,. \qquad (5)$$

Although we will not explore this further, we can also perform parameter estimation from partially observed data using expectation maximization, as in standard Gibbs MRFs [15], using Perturb-and-MAP sampling at the E-step.

We can design the perturbations so as the Perturb-and-MAP log-likelihood L_{PM} is a concave function of $\boldsymbol{\theta}$. This ensures that the likelihood landscape is well-behaved and allows the use of local search techniques for parameter estimation, exactly as in the Gibbs case. Specifically (see supplementary material for all proofs in the paper):

Proposition 1. *If the perturbations $\boldsymbol{\epsilon}$ are drawn from a log-concave density $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$, the log-likelihood $L_{PM}(\boldsymbol{\theta})$ is a concave function of the energy parameters $\boldsymbol{\theta}$.*

The family of log-concave distributions [4], i.e., $\log f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ is a concave function of $\boldsymbol{\epsilon}$, includes the Gaussian, the logistic, and other commonly used distributions.

The gradient of $L_{PM}(\boldsymbol{\theta})$ is in general hard to compute. Motivated by the parameter update formula in the Gibbs case, we opt for the moment matching learning rule, $\theta_j(t+1) = \theta_j(t) + r(t)\Delta\theta_j$, where

$$\Delta\theta_j = E_{\boldsymbol{\theta}}^{PM}\{\phi_j(\mathbf{x})\} - E_D\{\phi_j(\mathbf{x})\} \,. \qquad (6)$$

Here $E_{\boldsymbol{\theta}}^{PM}\{\phi_j(\mathbf{x})\} \triangleq \sum_{\mathbf{x}} f_{PM}(\mathbf{x}; \boldsymbol{\theta})\phi_j(\mathbf{x})$ is the expected sufficient statistic under the Perturb-and-MAP model for the current parameter values $\boldsymbol{\theta}$, which we can efficiently estimate by drawing exact samples from it. We typically adjust the learning rate by a Robbins-Monro type schedule, e.g., $r(t) = r_1/(r_2 + t)$. Figure 5 illustrates parameter learning by moment matching in a spatially homogeneous Perturb-and-MAP Ising model.

Changing the parameters $\boldsymbol{\theta}$ under the moment matching rule (6) indeed reduces the discrepancy between the model and data sufficient statistics. Specifically:

Proposition 2. *If $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ differ only in the j-element, with $\theta'_j > \theta_j$, then $E_{\boldsymbol{\theta}'}^{PM}\{\phi_j(\mathbf{x})\} \leq E_{\boldsymbol{\theta}}^{PM}\{\phi_j(\mathbf{x})\}$.*

The inequality in Proposition 2 will be strict if the perturbation density satisfies some mild conditions – see supplementary material. To see the effect of parameter update in the Perturb-and-MAP Ising model of Fig. 1, assume that $E_{\boldsymbol{\theta}}^{PM}\{\phi_3(\mathbf{x})\} = E_{\boldsymbol{\theta}}^{PM}\{-\frac{1}{2}x_1 x_2\}$ is larger than $E_D\{\phi_3(\mathbf{x})\}$. Under (6), we increase the coupling strength $\theta_3 = \lambda$; we see from Fig. 1 that the polyhedra of states $\mathbf{x} = (1, 1)$ and $\mathbf{x} = (-1, -1)$ expand over those of $\mathbf{x} = (1, -1)$ and $\mathbf{x} = (-1, 1)$, thus decreasing $E_{\boldsymbol{\theta}}^{PM}\{\phi_3(\mathbf{x})\}$.

Unlike the Gibbs case, the fixed points of the Perturb-and-MAP moment matching criterion do not need to be exact minima of the log-likelihood (5). However, some reassurance is provided by the fact that the M-projection of $f_{PM}(\boldsymbol{\theta}_{MM})$ (Perturb-and-MAP model trained by moment matching) is $f_G(\boldsymbol{\theta}_{ML})$ (Gibbs model trained by ML/MM) [15, Th. 8.6]. Specifically, $\mathcal{D}\left(f_{PM}(\boldsymbol{\theta}_{MM}) \,\|\, f_G(\boldsymbol{\theta}_{ML})\right) \leq \mathcal{D}\left(f_{PM}(\boldsymbol{\theta}_{MM}) \,\|\, f_G(\boldsymbol{\theta})\right), \forall \boldsymbol{\theta} \in \mathbb{R}^M$, where $\mathcal{D}(\cdot \,\|\, \cdot)$ is the Kullback-Leibler divergence between two distributions.

4. Perturb-and-MAP perturbation design

Although any perturbation density induces a legitimate Perturb-and-MAP model, it is desirable to carefully design it so as the Perturb-and-MAP model approximates as closely as possible the corresponding Gibbs MRF. The Gibbs MRF has important structural properties that are not automatically satisfied by the Perturb-and-MAP model under arbitrary perturbations: (a) Unlike the Gibbs MRF, the Perturb-and-MAP model is not guaranteed to respect the state ranking induced by the energy, i.e., $e(\mathbf{x}) \leq e(\mathbf{x}')$ does not necessarily imply $f_{PM}(\mathbf{x}) \geq f_{PM}(\mathbf{x}')$, see Fig. 2(d). (b) The Markov dependence structure of the Gibbs MRF follows directly from the support of the potentials $\phi_j(\mathbf{x})$, while the Perturb-and-MAP might give rise to longer-range probabilistic dependencies. (c) The maximum entropy distribution under moment constraints $E\{\phi_j(\mathbf{x})\} = \bar{\phi}_j$ has the Gibbs form; the Perturb-and-MAP model trained by moment matching can reproduce these moments but will in general have smaller entropy than its Gibbs counterpart.

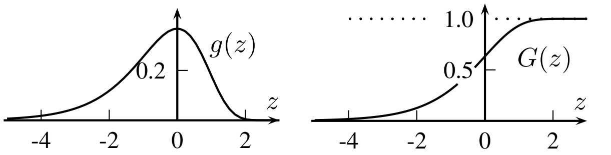

Figure 3. Gumbel probability density and cumulative distribution.

The *Gumbel* distribution arising in extreme value theory [25] turns out to play an important role in our effort to design a perturbation mechanism that yields a Perturb-and-MAP model closely resembling the Gibbs MRF. It is a continuous univariate distribution with log-concave density $g(z) = \exp(-(-z + e^z))$, plotted in Fig. 3. We can efficiently draw independent Gumbel variates by transforming standard uniform samples by $u \to \log(-\log(u))$. The Gumbel density naturally fits into the Perturb-and-MAP model, thanks to the following key Lemma – c.f. [18]:

Lemma 1. *Let $(\theta_1, \ldots, \theta_m)$, with $\theta_n \in \mathbb{R}$. We additively perturb them by $\tilde{\theta}_n = \theta_n + \epsilon_n$, with ϵ_n IID Gumbel samples. Then the probability that $\tilde{\theta}_n$ attains the minimum value is* $\Pr\{\mathrm{argmin}(\tilde{\theta}_1, \ldots, \tilde{\theta}_m) = n\} = e^{-\theta_n} / \sum_{n'=1}^{m} e^{-\theta_{n'}}$.

Gumbel perturbation on fully-expanded potential table The Gibbs random field on N sites x_i, $i = 1, \ldots, N$, each allowed to take a value from the discrete label set $\mathcal{L}$ can be considered as a discrete distribution with $|\mathcal{L}|^N$ states. This can be made explicit if we enumerate $\{\mathbf{x}_j, j = 1, \ldots, \bar{M} = |\mathcal{L}|^N\}$ all the states and consider the maximal equivalent re-parameterization of Eq. (1)

$$\bar{e}(\mathbf{x}; \bar{\boldsymbol{\theta}}) \triangleq \langle \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\phi}}(\mathbf{x}) \rangle = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}) \rangle \,, \qquad (7)$$

where $\bar{\theta}_j = e(\mathbf{x}_j; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\mathbf{x}_j) \rangle$, $j = 1, \ldots, \bar{M}$, is the *fully-expanded* potential table and $\bar{\phi}_j(\mathbf{x})$ is the indicator function of the state $\mathbf{x}_j$ (i.e., equals 1, if $\mathbf{x} = \mathbf{x}_j$ and 0 otherwise). Using Lemma 1 we can show:

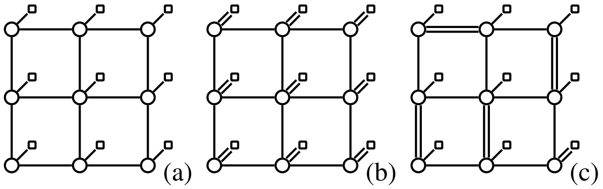

Figure 4. Reduced-order Gumbel perturbation. Perturbed potentials are denoted with double line. (a) Graph of the original energy involving unary and pairwise potentials on a 4-neighborhood graph. (b) Order-1 perturbation. (c) Order-2 perturbation.

Proposition 3. *If we perturb each entry of the fully expanded $\mathcal{L}^N$ potential table with IID Gumbel noise samples $\epsilon_j, j = 1, \ldots, \bar{M}$, then the Perturb-and-MAP and Gibbs models coincide, i.e., $f_{PM}(\mathbf{x}; \boldsymbol{\theta}) = f_G(\mathbf{x}; \boldsymbol{\theta})$.*

This order-N perturbation is not practically applicable when N is large since it independently perturbs all $\bar{M} = |\mathcal{L}|^N$ entries of the fully expanded potential table and effectively destroys the local Markov structure of the energy function, rendering it too hard to minimize. Nevertheless, it shows that it is possible to design a Perturb-and-MAP model that exactly replicates the Gibbs MRF and paves the way for the design of reduced-order Gumbel perturbations.

Reduced-order Gumbel perturbation In practice, we employ low-order Gumbel perturbations, typically only perturbing the unary (order-1) or a subset of the pairwise (order-2) potential tables. This yields perturbed energies effectively as easy to minimize as the original unperturbed one, while producing random samples closely resembling Gibbs MRF samples. We emphasize that even the order-1 Perturb-and-MAP model is able to reproduce the sufficient statistics of the data and is thus far more accurate than a mean-field approximation of the Gibbs MRF. Thanks to the log-concavity of the Gumbel density, the log-likelihood of the Perturb-and-MAP model remains concave for Gumbel perturbations of any order, as follows from Proposition 1.

To be more specific, consider the second-order energy

$$e(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{N} \Big(V_i(x_i) + \sum_{i' \in \mathcal{N}(i)} V_{ii'}(x_i, x_{i'}) \Big) \qquad (8)$$

where each site x_i can take a discrete label in $\mathcal{L}$. This is a generalization of the Ising model considered in Sec. 3.1, where $|\mathcal{L}| = 2$. Each V_i is a $|\mathcal{L}| \times 1$ unary potential table and each $V_{ii'}$ is a $|\mathcal{L}| \times |\mathcal{L}|$ pairwise potential table.

The order-1 perturbation, illustrated in Fig. 4(b), amounts to adding IID Gumbel noise to each entry of every unary potential table V_i. This requires generating $|\mathcal{L}|N$

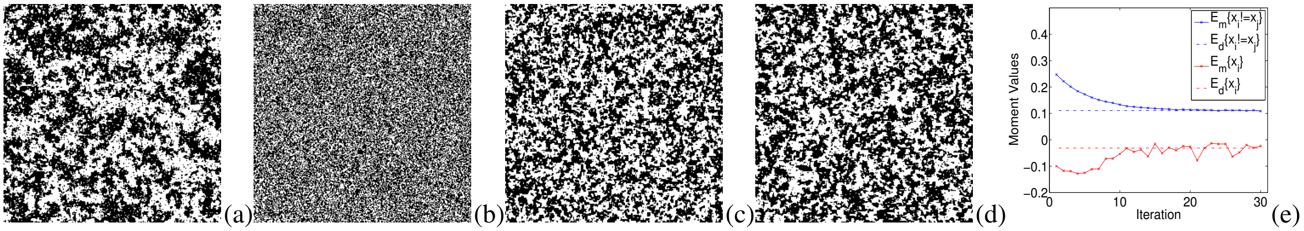

Figure 5. Low-order Perturb-and-MAP Ising random field parameter learning. The two model parameters, the global coupling strength λ and field strength β are fitted by moment matching. (a) One of the 10 Gibbs Ising model samples just below the critical temperature ($\lambda = 0.88$, $\beta = 0$, 256×256 grid) that we used as training data. (b) Perturb-and-MAP Ising sample at initial parameter values. (c) Order-1 Perturb-and-MAP sample at fitted parameter values. (d) Order-2 Perturb-and-MAP sample at fitted parameter values. (e) Model moments of the order-2 Perturb-and-MAP model as they progress towards the training data moments during moment matching learning.

IID Gumbel samples. Note that in the special case of the Ising model, the order-1 Gumbel perturbation is equivalent to adding logistic noise to the unary factor parameter β_i, since the difference of two IID Gumbel samples follows a logistic distribution [25].

In the order-2 perturbation, illustrated in Fig. 4(c), we add IID Gumbel noise to each entry of a subset of the pairwise potential tables $V_{ii'}$. We make sure that at most one of the pairwise potentials adjacent to any node is perturbed. If none of the pairwise potentials adjacent to a node can be perturbed, then we perturb its associated unary potential. In total, the perturbation process requires generating at most $(N/2)|\mathcal{L}|^2$ IID Gumbel samples. Higher-order Gumbel perturbations involving clusters of 3 or more variables can be similarly defined.

It is desirable to select the strongest among the pairwise potentials adjacent to each node for order-2 perturbation. For energies defined on 4-connected planar graphs, we globally find an optimal subset of strongest links by solving a stable marriage (also called stable matching) problem on the corresponding Red-Black bipartite graph using the Gale-Shapley algorithm. See [13] and particularly [9] for a description of the Gale-Shapley algorithm as it applies to sets of men/women of unequal size, as can happen in our case. To indicate the mating preferences of each node x_i, we rank its neighbors in decreasing order of pairwise mutual information $I_{ii'} = \sum_{x_i, x_{i'}} p_{ii'}(x_i, x_{i'}) \log \frac{p_{ii'}(x_i, x_{i'})}{p_i(x_i) p_{i'}(x_{i'})}$, with $p_{ii'}(x_i, x_{i'}) \propto \exp\left(-V_i(x_i) - V_{i'}(x_{i'}) - V_{ii'}(x_i, x_{i'})\right)$ and $p_i(x_i) = \sum_{x_{i'}} p_{ii'}(x_i, x_{i'})$. For the Ising model, $I_{ii'}$ increases with the edge strength $|\lambda_{ii'}|$. When producing multiple samples, we perform link selection only once. The computational cost is around 0.1 sec for 300×300 images with our implementation of the Gale-Shapley algorithm.

While the order-1 perturbation preserves submodularity [17], order-2 perturbation can yield non-submodular functions even when the original energy is submodular. For the Ising model we can compute in closed form the probability that a single pairwise link of strength λ will be submodular after order-2 Gumbel perturbation $\Pr\{\tilde{\lambda} \geq 0\} = e^{2\lambda}(e^{2\lambda} - 2\lambda - 1)/(e^{2\lambda} - 1)^2$; e.g., for $\lambda = 4$, $\Pr\{\tilde{\lambda} \geq 0\} \approx 0.998$. Thus, if the links selected for perturbation are sufficiently strong (and the link selection process described in the previous paragraph contributes to this goal), then most of the perturbed pairwise potentials will remain submodular and the perturbed energy can efficiently be minimized with techniques such as QPBO [16] which gracefully handle the few non-submodular links. This is the approach we follow in the interactive image segmentation application. Otherwise, for weak links the order-1 perturbation should be preferred, which is anyway accurate enough in this case.

In Fig. 5 we juxtapose Perturb-and-MAP samples produced by order-1 and order-2 Gumbel perturbations with a Gibbs MRF sample from the Ising model, produced with the Propp-Wilson exact sampling algorithm [21]. We have fitted the parameters of the Perturb-and-MAP models by moment matching so that they reproduce the first and second order statistics of the Gibbs sample. We see that even the order-1 Gumbel perturbation captures quite well the overall appearance of the exact Gibbs sample. The order-2 sample further improves the approximation quality, better capturing the appearance of same-spin clusters in the Gibbs sample.

5. Applications and experiments

We present experiments with the Perturb-and-MAP model applied to image segmentation and scene labeling. Further results are included in the supplementary material. Software is available from the first author's web home page.

5.1. Interactive image segmentation

We first report interactive segmentation experiments, performed on the Grabcut dataset which includes human annotated ground truth segmentations [22]. The task is to segment a foreground object, given a relatively tight tri-map imitating user input obtained by a lasso or pen tool.

This is a relatively small dataset (50 images) not split into training and test sets and carefully optimized techniques which exploit the regularities of the dataset are achieving extremely low pixel misclassification results (around 4.5% using adaptive thresholding on the output of the random walker model [8]) – see [23] for a recent review.

In our implementation we closely follow the CRF formulation of [23], using the same parameters for defining the image-based CRF terms and considering pixel interactions in a 8-neighborhood. We used our Perturb-and-MAP sampling algorithm with order-2 Gumbel perturbation and QPBO optimization [16] to learn the weights of the potentials – 5 weights in total, one for the unary and one for each of the 4 pairwise connections of the center pixel with its S, E, NE, SE neighbors. Using these parameters, we obtained a classification error rate of 5.6% with the global MAP decision rule. This is similar to the best results attainable with the particular CRF model and hand-tuned weights.

In Fig. 6 we illustrate the ability of the Perturb-and-MAP model to produce soft segmentation maps. The soft segmentation map (average over 20 posterior samples) gives a qualitatively accurate estimate of the segmentation uncertainty, which could potentially be useful in guiding user interaction in an interactive segmentation application.

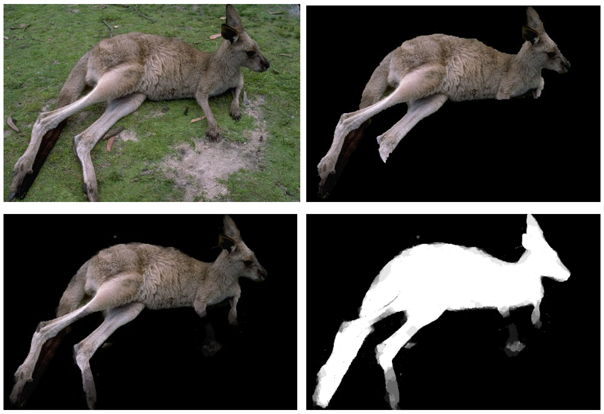

Figure 6. Interactive image segmentation results on the Grabcut dataset. Parameters learned by PM moment matching. Top: the original image and the least energy MAP solution. Bottom: soft Perturb-and-MAP segmentation and the corresponding mask.

5.2. Scene layout labeling

We next consider an application of Perturb-and-MAP random fields in scene layout labeling [12]. We use the tiered layout model of [6], which allows exact global inference by efficient dynamic programming [6]. The model has a relatively large number of parameters, making it difficult to hand tune. Training them with the proposed techniques illustrates our ability to effectively learn model parameters from labeled data.

We closely follow the evaluation approach of [6] in setting up the experiment: We use the dataset of 300 outdoor images (and the standard cross-validation splits into training/test sets) with ground truth from [12] for our experiments. Similarly to [6], we use five labels: T (sky), B (ground), and three labels for the middle region, L (facing left), R (facing right), C (front facing). We also do not include in our label set the classes “porous” and “solid”. The per-pixel class confidences used as unary terms are produced using classifiers that we trained using the dataset and software provided by [12] following the standard five-fold cross-validation protocol. The small difference between the baseline confidence-only classification results reported by [6] and our baseline result should be attributed to our use of the newer version of Hoiem’s software.

We first fit the tiered scene model parameters (pairwise compatibility tables between the different classes) on the training data using Perturb-and-MAP moment matching (order-1 Gumbel perturbation). Weights are initialized as Potts CRF potentials and refined by moment matching rule; we separated the training set in batches of 10 images each and stopped after 50 epochs over the training set.

The following tables report row-normalized confusion matrices for MAP (least energy configuration) and marginal MODE (i.e., assign each pixel to the label that appears most frequently in 20 random Perturb-And-Map conditional samples from the model); in both cases the learned weights are used. Our results are better than the confidence-only baseline mean accuracy of 82.1% [12], and the MAP and MODE results of 82.1% and 81.8%, respectively, that we obtained with the hand-set weights of [6].

MAP (acc 82.7%)

	B	L	C	R	T
B	95.3	0.5	3.3	0.5	0.5
L	22.9	46.7	21.9	6.6	1.9
C	24.3	6.7	52.5	11.4	5.0
R	16.0	4.4	24.8	49.4	5.4
T	1.1	0.6	3.1	0.7	94.4

Marginal MODE (acc 82.6%)

	B	L	C	R	T
B	95.3	0.5	3.2	0.5	0.5
L	23.0	46.9	21.7	6.6	1.9
C	24.6	6.9	51.5	11.7	5.3
R	16.5	4.5	24.2	49.1	5.7
T	1.0	0.6	3.0	0.8	94.7

Table 1. *Tiered labeling confusion matrices (learned weights).*

In Fig. 7 we show some indicative examples of different scene layout labelings obtained by the confidence-only, the MAP, and the Perturb-and-MAP model. The uncertainty of the solution is indicated by entropy maps.

6. Perspective

The work of Geman and Geman [7] showed that sampling coupled with artificial temperature annealing can be used as a general purpose method for finding the least energy configuration in discrete label MRFs. The advent of much faster deterministic energy minimization techniques has decreased interest in sampling as an intermediary for MAP computation. Interestingly, the Perturb-and-MAP model works in the opposite direction to simulated annealing, allowing powerful algorithms for MAP computation to act as intermediaries for MRF sampling. We hope that this research will help establish discrete optimization techniques as tools for probabilistic modeling in computer vision.

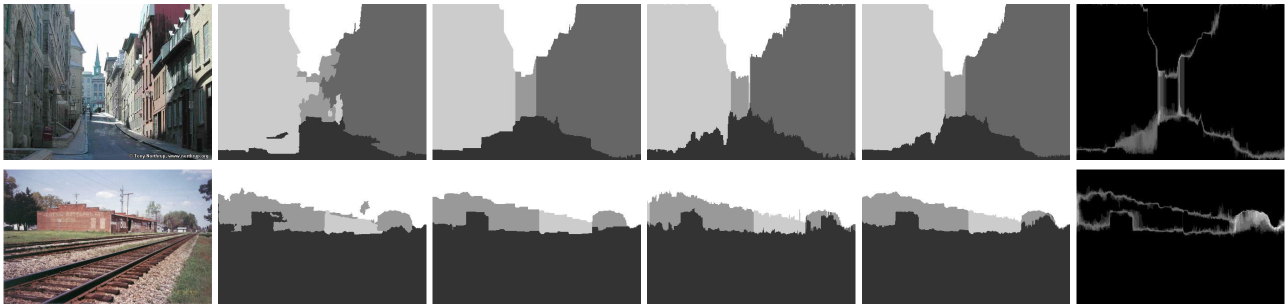

Figure 7. Tiered scene labeling results with pairwise potentials learned by our Perturb-and-MAP moment matching algorithm. Left to right: image; confidence-only result; least energy MAP solution; single Perturb-and-MAP sample; PM marginal mode; PM marginal entropy.

Acknowledgments

This work was supported by the U.S. Office of Naval Research under the MURI grant N000141010933 and by the Korean Ministry of Education, Science, and Technology, under the National Research Foundation WCU program R31-10008. We thank X. He, I. Kokkinos, and M. Raptis for their feedback and help at various stages of this project.

References

[1] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Univ. Press, 2009.

[2] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *JRSS (B)*, 36(2):192–236, 1974.

[3] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.

[6] P. Felzenszwalb and O. Veksler. Tiered scene labeling with dynamic programming. In *Proc. CVPR*, 2010.

[7] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. PAMI*, 6(6):721–741, 1984.

[8] L. Grady. Random walks for image segmentation. *IEEE Trans. PAMI*, 28(11):1768–1783, 2006.

[9] D. Gusfield and R. Irving. *The Stable Marriage Problem*. MIT Press, 1989.

[10] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neur. Comp.*, 14(8):1771–1800, 2002.

[11] G. Hinton and T. Sejnowski. Optimal perceptual inference. In *Proc. CVPR*, pages 448–453, 1983.

[12] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007.

[13] J. Kleinberg and E. Tardos. *Algorithm Design*. Addison-Wesley, 2006.

[14] P. Kohli and P. Torr. Measuring uncertainty in graph cut solutions. *Comp. Vision Image Underst.*, 112(1):30–38, 2008.

[15] D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT Press, 2009.

[16] V. Kolmogorov and C. Rother. Minimizing non-submodular functions with graph cuts – a review. *IEEE Trans. PAMI*, 29(7):1274–1279, July 2007.

[17] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. PAMI*, 26(2):147–159, 2004.

[18] D. Kuzmin and M. K. Warmuth. Optimum follow the leader algorithm. In *Proc. COLT*, pages 684–686, 2005.

[19] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F.-J. Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2007.

[20] G. Papandreou and A. Yuille. Gaussian sampling by local perturbations. In *Proc. NIPS*, 2010.

[21] J. Propp and D. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struc. Algor.*, 9(1):223–252, 1996.

[22] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *Proc. SIGGRAPH*, pages 309–314, 2004.

[23] C. Rother, V. Kolmogorov, Y. Boykov, and A. Blake. Interactive foreground extraction using graph cut. In *Advances in Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

[24] U. Schmidt, Q. Gao, and S. Roth. A generative perspective on MRFs in low-level vision. In *Proc. CVPR*, 2010.

[25] F. Steutel and K. Van Harn. *Infinite divisibility of probability distributions on the real line*. Dekker, 2004.

[26] M. Szummer, P. Kohli, and D. Hoiem. Learning CRFs using graph cuts. In *Proc. ECCV*, pages 582–595, 2008.

[27] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Proc. NIPS*, 2003.

[28] M. Wainwright, T. Jaakkola, and A. Willsky. MAP estimation via agreement on trees: Message-passing and linear programming. *IEEE Trans. Inf. Theory*, 51(11):3697–3717, 2005.

[29] M. Welling. Herding dynamical weights to learn. In *Proc. ICML*, pages 1121–1128, 2009.

[30] O. Woodford, C. Rother, and V. Kolmogorov. A global perspective on MAP inference for low-level vision. In *Proc. ICCV*, pages 2319–2326, 2009.

[31] S.-C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *IJCV*, 27(2):107–126, 1998.