Local Autoencoding for Parameter Estimation in a Hidden Potts-Markov Random Field

Citation for published version:
Song, S, Si, B, Herrmann, JM & Feng, X 2016, 'Local Autoencoding for Parameter Estimation in a Hidden Potts-Markov Random Field', *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2324-2336. https://doi.org/10.1109/TIP.2016.2545299

Digital Object Identifier (DOI):
10.1109/TIP.2016.2545299

Local Autoencoding for Parameter Estimation in a Hidden Potts-Markov Random Field

Sanming Song, Bailu Si, J. Michael Herrmann, and Xisheng Feng

***Abstract*—A local-autoencoding (LAE) method is proposed for the parameter estimation in a Hidden Potts-Markov random field (MRF) model. Due to sampling cost, Markov chain Monte Carlo (MCMC) methods are rarely used in real-time applications. Like other heuristic methods, LAE is based on a conditional independence assumption. It adapts, however, the parameters in a block-by-block style with a simple Hebbian learning rule. Experiments with given label fields show that LAE is able to converge in far less time than required for a scan. It is also possible to derive an estimate for LAE based on a Cramer-Rao bound that is similar to the classical maximum pseudo-likelihood (MPL) method. As a general algorithm, LAE can be used to estimate the parameters in anisotropic label fields. Furthermore, LAE is not limited to the classical Potts model and can be applied to other types of Potts models by simple label field transformations and straightforward learning rule extensions. Experimental results on image segmentations demonstrate the efficiency and generality of the LAE algorithm.**

***Index Terms*—Markov random field, Gibbs distribution, parameters estimation, local autoencoding, Potts model**

I. Introduction

MARKOV random fields (MRFs) are very popular statistical models [1]–[3] that are used in a wide range of computer vision problems [4]. By describing the prior knowledge that neighboring pixels are likely to be of the same class in a parameterised way, a Potts-MRF is able to include a local similarity constraint in segmenting images which are distorted by speckle noise, like underwater sonar images [5], geodesic hyperspectral images [6], synthetic aperture radar (SAR) images [7], magnetic resonance (MR) images [8] etc. In a Potts field, each unit can be in a number of Q states (classes). Statistically, when the local correlation is assumed to be Markovian, the configuration space can be modeled by a Gibbs distribution which is characterised by a vector of parameters $\boldsymbol{\beta}$ (also known as interaction parameters [5], spatial cohesion strength [9], granularity coefficient [10] or prior parameters [11]). However, with the notable exception of the special case of an isotropic Ising-MRF with circular boundary conditions [12], a direct estimation of the prior parameters is not possible, because a part of the model, the partition function, is computationally intractable [10]. This paper focuses on estimating β parameters from data.

McGrory et al. [13] have proposed a recursive method for computing the partition function analytically, which is, however, applicable only to MRFs with a small size and a weak interaction strength. Instead, two types of approximation strategies are used to find optimal parameters, namely sampling methods based on Monte-Carlo sampling, such as the Markov chain Monte-Carlo (MCMC) algorithm [14], [15], MCMC Maximum Likelihood (MCMCML) [16], and likelihood-free methods [10], [17], and heuristic methods based on the conditional independence assumption, e.g. pseudo-likelihood maximization (MPL) [18], [19], coding-based MPL [20], mean-field approximation [21], [22], least-square (LSQR) algorithm [5], [23].

MCMC methods approximate the posterior probability by Monte Carlo sampling iteratively until the generation parameters converge. They may be more precise, but are computationally more expensive [10], [16]. In engineering applications, heuristic methods may appear thus more favorable. However, even MPL [18], coding-based MPL [20], and mean-field approximations [21] have to pass through the image many times. The LSQR algorithm [23] does not only ignore a large number of blocks that are incompatible with certain mathematical constraints, but also requires complex matrix computation. In addition, LSQR appears to be feasible for the Ising field.

In this paper, we propose a new algorithm to fast estimate the interaction parameters of a Potts model that is used to describe the label field. As our algorithm uses a local autoencoding approach to capture image regularities, we will refer to it by the abbreviation LAE. Basing on the conditional independent assumption, it belongs to the class of heuristic methods related to MPL estimation. The main advantage of our method is that it is computationally very efficient and therefore interesting for applications involving limited computing times. More over, it is of a comparable accuracy to other pseudo-likelihood approaches. In detail, we will show that the LAE algorithm is superior to other methods in the following aspects: 1) LAE is very simple, since a simple local backpropagation (BP) gradient descent learning rule is able to estimate the prior parameters. 2) LAE converges rapidly, even less than a single pass. 3) Simulations show that LAE has a comparable mean-

Manuscript received July 28, 2015; revised October 9, 2015, December 7, 2015 and February 1, 2016; accepted March 12, 2016. Date of publication ***** **, 2016; date of current version March 20, 2016. This work was supported by the National Natural Science Foundation of China (No. 41506121), the China Postdoctoral Science Foundation (No. 2014M561266), Jiang Xinsong Innovation Fund (No. Y4FC012901), the State Key Laboratory of Robotics (No. Y5A1203901), the Distinguished Young Scholar Project of the Thousand Talents Program of China (Y5A1370101), the Doctoral Scientific Research Foundation of Liaoning Province (No. 201501035), and the project of "R&D Center for Underwater Construction Robotics" that was funded by the Ministry of Ocean and Fisheries (MOF) and Korea Institute of Marine Science & Technology Promotion (KIMST), Korea (No. PJT200539).

S. Song, B. Si and X. Feng are with the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China (e-mail: songsanming@sia.cn, sibailu@sia.ac.cn, fxs@sia.cn).

J.M. Herrmann is with the Institute of Perception, Action and Behavior, University of Edinburgh, Edinburgh EH8 9AB, U.K. (e-mail: michael.herrmann@ed.ac.uk).

squared error with the classical MPL method. 4) LAE makes full use of all the local blocks. 5) LAE can be used to estimate the prior parameters for arbitrary anisotropic label fields. 6) LAE can be extended to various variants of the Potts model by a simple label field transformation.

The rest of the paper is organized as follows. The Potts hidden-MRF model is simply introduced in Section II. The existing methods for the prior probability parameter estimation are briefly reviewed in Section III. The LAE algorithm is described in Section IV and experiment results are presented in Section V. We briefly conclude in Section VI. Application of the LAE algorithm to another typical Potts model, i.e. Kanter's model, is given in the Appendix.

II. Hidden-Markov random field

In Markovian image segmentation or labelling, we aim at maximizing the posterior probability $P_{L/X}(l/x)$ for the label field, where X is the original image and L is the label. According to Bayes' rule, we have

$$P_{L/X}(l/x) = \frac{P_L(l)\, P_{X/L}(x/l)}{P_X(x)}. \tag{1}$$

Now that $P_X(x)$ is fixed for a given image X, finding the label field that maximizes (1) is equivalent to finding the labels L which maximize the following energy function:

$$E = \ln P_L(l) + \ln P_{X/L}(x/l). \tag{2}$$

The choice of the conditional distribution $P_{X/L}(x/l)$ depends on the application. In many applications related to remote sensing, i.e. sonar or geodesic imaging, a Gauss or a Weibull distribution [5], [24] is often assumed. However, for the label field L, a similar assumption for the prior distribution $P_L(l)$ is hard to justify.

We will consider a Potts label field L of size $\|L\| = H \times W$ which is defined over the set $S = \{s\}_{1,\ldots,H\times W}$ of image pixels. In the model, labels L_s are described by one of Q Potts states. If L is assumed to be a Markov random field, then $P_L(l)$ can be described by a Gibbs distribution according to the Hammersley-Clifford theorem [25],

$$P_L(l) = \frac{1}{Z} e^{V_l}, \tag{3}$$

where $Z = \sum_l e^{V_l}$ is the partition function and V_l is the potential of a configuration l given by

$$V_l = \sum_s \sum_{t\in\Xi_s} \beta_{st}\delta(l_s, l_t). \tag{4}$$

The Kronecker delta function is defined by $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ otherwise. The inner sum extends over the neighborhood Ξ_s of s and the weights β_{st} are solely determined by the local site relationships, see Fig. 1 for two examples of a 2nd neighborhood system.

The solution to the model can be computed by performing the Bayesian inference using local optimization methods such as the Iterated Condition Method [5], [23] or global optimization methods such as Simulated Annealing [26]. In the inference, each unknown parameter of the model is updated iteratively, including the conditional probability parameters and the prior probability parameters. Our method will be used for updating the prior parameter β with the conditional parameters fixed.

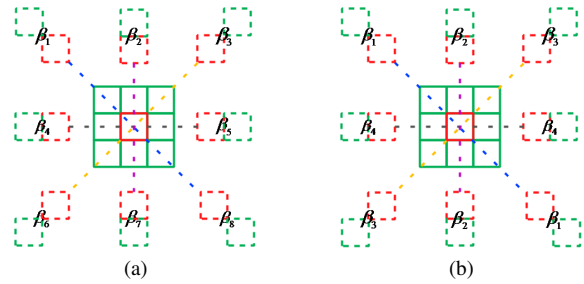

Fig. 1. Anisotropic interaction between neighboring units in a 2nd-neighborhood system, and the interaction strength is denoted by β_s. (a) is central asymmetry and (b) is central symmetry.

III. Existing methods

To simplify the descriptions, except the LSQR method, the prior parameters are isotropic in the following subsections. In this case, the blob size in the Potts field increases with β.

A. Maximum Pseudo-Likelihood

ML is a basic strategy for parameter learning, but determining the gradient of the original prior distribution (3) is forbiddingly hard as it includes calculating the partition function. To remove this problem, Besag [18] proposed the *MPL* method which makes use of a conditional independence assumption.

Setting $\mathrm{N}_{l_s} = \sum_{t\in\Xi_s} \delta(l_s, l_t)$, the conditional probability distribution for the label l_s is

$$P(l_s | l_{\Xi_s}) = \frac{e^{\beta \mathrm{N}_{l_s}}}{\sum_{l_s} e^{\beta \mathrm{N}_{l_s}}}. \tag{5}$$

With the conditional independence assumption, the conditional likelihood function, also named *pseudo-likelihood* function, can be written as

$$PL = \prod_s P(l_s | l_{\Xi_s}). \tag{6}$$

Taking the derivative of the logarithm of the pseudo-likelihood function with parameter β, the gradient descent learning rule for β becomes

$$\Delta\beta = \eta \frac{\partial}{\partial\beta} \log PL = \eta \sum_s \left\{ \mathrm{N}_{l_s} - \frac{\sum_{l_s} e^{\beta \mathrm{N}_{l_s}} \mathrm{N}_{l_s}}{\sum_{l_s} e^{\beta \mathrm{N}_{l_s}}} \right\}, \tag{7}$$

where η is a learning rate.

B. Coding

The *coding* scheme was proposed also by Besag [25] in order to better accommodate the conditional independence assumption. Coding has a similar learning rule as MPL (7), with the only difference that the lattice is partitioned into several disjoint sets. A minor problem is how to combine the results from each set, but usually a simple average is feasible.

C. Mean field approximation

Mean field approximation can be considered as a gradient-descent method that has the function of "*smoothing*". The learning rule is

$$\Delta\beta = \eta \sum_{s} \sum_{z_s} P\left(z_s \left| x_s\right.\right) \left[\mathrm{R}_{\mathrm{z_s}} - \frac{\sum\limits_{z_s=1}^{Q} e^{\beta \mathrm{R}_{\mathrm{z_s}}} \mathrm{R}_{\mathrm{z_s}}}{\sum\limits_{z_s=1}^{Q} e^{\beta \mathrm{R}_{\mathrm{z_s}}}} \right], \quad (8)$$

where z_s is the vector description for the scalar label l_s, and $P\left(z_s \left| x_s\right.\right)$ is the conditional expectation under the mean-field assumption. $\mathrm{R}_{\mathrm{z_s}} = \sum\limits_{t \in \Xi_s} z_s \left\langle z_t \right\rangle$, where $\langle \cdot \rangle$ is the mean-field approximation.

The mean-field solution is an heuristic parameter estimation step in the expectation-maximization (EM) procedure [21], [22]. Essentially, it uses the gradient descent and resembles the MPL method, with the exception of replaceing the homogeneous clique number with the approximative local interaction energies $\mathrm{R}_{\mathrm{z_s}}$, and weighting them with the conditional probability.

D. MCMC

Besides the heuristic methods that accelerate the estimation with the cost of accuracy, there is another kind of methods that pursue the precision at all costs, like Monte Carlo sampling. There is a 1-to-1 correspondence between β and the expectation of the homogeneous cliques number $\mathrm{N_l} = \sum\limits_{s} \mathrm{N_{l_s}}$ [16], which demonstrates that we can write a learning rule for β using $\mathrm{N_l}$ according to the Monte-Carlo sampling [14], [15],

$$\Delta\beta_t = \eta \left(\hat{E}_t \left(\mathrm{N_l}\right) - \mathrm{N_{l0}} \right), \quad (9)$$

with the expectation

$$\hat{E}_t \left(\mathrm{N_l}\right) = \frac{1}{M} \sum_{m=1}^{M} \mathrm{N_l} \left(L^{m,t}\right). \quad (10)$$

The new set $\left\{L^{m,t+1}\right\}_{m=1,\ldots,M}$ are sampled in the burn-in period by Gibbs sampling with current parameter β_t. $\hat{E}_t \left(\mathrm{N_l}\right)$ is the expectation of samples, and $\mathrm{N_{l0}}$ is calculated from the label field to be estimated. To reduce the computation cost, $L^{0,t+1} = L^{M,t}$ is used. Even then, the computation cost is huge because MCMC generates multiple label field samples by Monte-Carlo sampling.

E. MCMCML

MCMCML has the same learning rule, i.e. Eq. 9, as the MCMC. However, instead of the average sampling, MCMCML uses the importance sampling,

$$\hat{E}_t \left(\mathrm{N_l}\right) = \frac{\sum\limits_{m=1}^{M} \mathrm{N_l} \left(L^{m,t}\right) e^{(\beta-\psi)\mathrm{N_l}\left(L^{m,t}\right)}}{\sum\limits_{m=1}^{M} e^{(\beta-\psi)\mathrm{N_l}\left(L^{m,t}\right)}} \quad (11)$$

The expectation is a weighted summation of samples, with the similar samples dominating the gradient. The samples may be calculated off-line and stored in a dataset [16], but the Gibbs sampling scale is still large.

F. Likelihood-free methods

Recently, Pereyra et al. [10], [27] proposed a likelihood-free method based on the approximative Bayesian computation [28] and the likelihood-free Metropolis-Hastings sampling [17], [29]. Instead of the gradient descent rules that are derived from the likelihood or pseudo-likelihood function, it jointly estimates the label, the prior parameter and other auxiliary variables by the hybrid Gibbs sampling.

Likelihood-free methods are only exact when perfect sampling [29], [30] is used. Taking the time-criticality into consideration, we will not discuss it in the latter experiments.

G. Least Squares

The LSQR method, proposed by S. Geman and D. Geman in 1984 [1], is also a heuristic method derived from the conditional probability.

The ratio of the conditional probabilities (5) is

$$\frac{P\left(l_s = q_1 \left| l_{\Xi_s}\right.\right)}{P\left(l_s = q_2 \left| l_{\Xi_s}\right.\right)} = \frac{e^{\sum\limits_{t \in \Xi_s} \beta_t \delta(q_1, l_t)}}{e^{\sum\limits_{t \in \Xi_s} \beta_t \delta(q_2, l_t)}} = e^{\sum\limits_{t \in \Xi_s} \beta_t \left[\delta(q_1, l_t) - \delta(q_2, l_t)\right]}. \quad (12)$$

On other hand, the conditional probability can be approximated by the histogram statistics,

$$\frac{P\left(l_s = q_1 \left| l_{\Xi_s}\right.\right)}{P\left(l_s = q_2 \left| l_{\Xi_s}\right.\right)} \approx \frac{N_{l_s, l_{\Xi_s}}\left(q_1, \Gamma_s\right)}{N_{l_s, l_{\Xi_s}}\left(q_2, \Gamma_s\right)} \quad (13)$$

where $N_{l_s, l_{\Xi_s}}\left(q_i, \Gamma_s\right)$ is the number of blocks that label q_i is surrounded by Γ_s which is the neighborhood label configuration of site s.

From Eqs. 12 and 13, we have

$$\sum_{t \in \Xi_s} \beta_t \left[\delta\left(q_1, l_t\right) - \delta\left(q_2, l_t\right)\right] = \ln \left[\frac{N_{l_s, l_{\Xi_s}}\left(q_1, \Gamma_s\right)}{N_{l_s, l_{\Xi_s}}\left(q_2, \Gamma_s\right)} \right] \quad (14)$$

For 8 neighbors, there are 2^8 configurations Γ implying 2^8 linear equations which can be solved by a Least Squares method. However, due to the mathematical constraints, many equations should be deleted [11]. Therefore, a large number of image blocks are ignored. Moreover, it is only available for Ising model because there are $Q^8 C_Q^2$ equations with Q Potts states. The resulting coefficient matrix not only is extremely sparse, but also needs huge memory to store.

In summary, the LSQR method appears to be suitable for Ising fields only. The MCMC methods are more precise than the heuristic algorithms. However, with iterative Monte-Carlo sampling, it has a huge computation cost. Even with the heuristic methods, the whole image needs to be passed several times. Therefore, it is necessary to design an algorithm that is available for the time-critical applications.

IV. Local autoencoding

In this section, a local-autoencoding (LAE) scheme will be proposed to estimate the prior parameters. The LAE algorithm is intuitively designed for the standard Potts model, its extension to another Potts model is described in the Appendix.

The LAE method is an attempt to use the block information in the label field efficiently, thus avoiding the redundancy

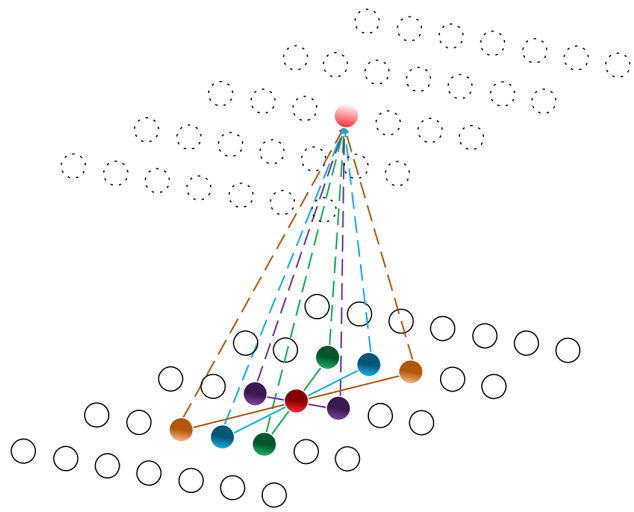

Fig. 2. Local autoencoding. The label of a pixel (shown in red) is predicted by the labels of its eight neighbors. The solid connections represent the filter $\boldsymbol{\beta}$. The predictive nature of the LAE is visualized as a feedforward network shown as the dashed lines and circles which do not need to be specifically represented in the algorithm.

in the LSQR method. It exploits the decomposition of the potential function V_l (4). Like other heuristic methods [18], [20]–[23], LAE is also based on the maximization of an estimate of the pseudo-likelihood (6). We rewrite Eq. 6 as

$$PL(l) = \prod_s P(l_s | l_{\Xi_s}). \tag{15}$$

where the conditional probability is

$$\begin{aligned} P(l_s | l_{\Xi_s}) &= \frac{e^{V(l_s, l_{\Xi_s})}}{\sum_{q=1}^{Q} e^{V(q, l_{\Xi_s})}} && (16)\\ &= \frac{1}{1 + \sum_{q \neq l_s} e^{[V(q, l_{\Xi_s}) - V(l_s, l_{\Xi_s})]}} && (17)\\ &= \frac{1}{1 + \sum_{q \neq l_s} e^{\sum_{t \in \Xi_s} \beta_{st}[\delta(q, l_t) - \delta(l_s, l_t)]}}. && (18) \end{aligned}$$

Eq. 18 was obtained by substituting Eq. 4 into Eq. 17. With

$$\delta(q, l_t) + \delta(l_s, l_t)|_{q \neq l_s} = \begin{cases} 1 & \text{if } l_t = q \parallel l_t = l_s \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

the conditional probability (18) becomes

$$\begin{aligned} & P(l_s | l_{\Xi_s}) \\ &= \left\{ 1 + \sum_{q \neq l_s} e^{\sum_{l_t = q | l_s} \beta_{st}[-2\delta(l_s, l_t) + 1] + \sum_{l_t \neq q \& l_t \neq l_s} \beta_{st}[-2\delta(l_s, l_t)]} \right\}^{-1}_{t \in \Xi_s} \\ &= \left\{ 1 + \sum_{q \neq l_s} e^{\sum_{l_t = q \& l_t \neq l_s} \beta_{st} - \sum_{l_t \neq q \& l_t = l_s} \beta_{st}} \right\}^{-1}_{t \in \Xi_s} \\ &= \left\{ 1 + \sum_{q \neq l_s} e^{\sum_{l_t = q} \beta_{st} - \sum_{l_t = l_s} \beta_{st}} \right\}^{-1}_{t \in \Xi_s} \end{aligned} \tag{20}$$

Algorithm 1 Local autoencoding for Potts model

Initialize β_i by a small value, $\beta_i = 0.01 * U(0, 1)$;
Set the initial learning rate $\eta_0 = 0.05$;
for $t = 1$; $t < T$; $t++$ **do**
  Slowly decreasing the learning rate by $\eta = \eta_0 / t$;
  for $r = 2$; $r < H$; $r++$ **do**
    for $c = 2$; $c < W$; $c++$ **do**
      Get the 3*3 block l_{Ξ_s} that centers at $s = (r, c)$;
      for $q = 1$; $q < Q$; $q++$ **do**
        if $q \neq l_s$ **then**
          Transform the label from $\{1, \ldots, Q\}$ *to* $\{-1, 0, 1\}$ *by Eq. 28;*
          Calculate the output of pixel s by the cross-correlation between l'_{Ξ_s} and filter $\boldsymbol{\beta}$ using Eq. 25;
          Adjust $\boldsymbol{\beta}$ by Eq. 27;
          if $\|\boldsymbol{\beta}\| = 4$ **then**
            $\beta_i = \frac{\beta_i + \beta_{9-i}}{2}$;
          end if
        end if
      end for
    end for
  end for
end for
Output $\hat{\boldsymbol{\beta}}$ by averaging the estimations in the last τ steps;

If β_{st} is coincidentally very large when $l_s = l_t$ and β_{st} is very small (or negative) when $l_s \neq l_t$, then $P(l_s | l_{\Xi_s})$ and the pseudo-likelihood function Eq. 15 would be very large. In other words, when maximizing $PL(l)$, β_{st} should increase when $l_s = l_t$ and decrease when $l_s \neq l_t$. So, interactions between the pixels of the same class should be strengthened. In verse, interactions between different classes should be shrunken. *Such a rule is similar to a Hebbian learning rule of a binary Hopfield network.* The learning rules for the prior parameter, starting with the Ising model and then the Potts model, will be developed in the following *subsections*.

1) Ising model: If the label values in $l \in \{-1, 1\}$, Eq. 4 can be rewritten as

$$\begin{aligned} V_l &= \sum_s \sum_{t \in \Xi_s} \beta_{st} \left(\delta(l_s, l_t) - \frac{1}{2} + \frac{1}{2} \right) && (21)\\ &= \frac{1}{2} \sum_s \sum_{t \in \Xi_s} \beta_{st} l_s l_t + c, && (22) \end{aligned}$$

where $c = \frac{1}{2} \sum_s \sum_{t \in \Xi_s} \beta_{st}$ is a constant. The first term turns out to be the Hamiltonian of a Hopfield network, but with local spatially invariant connections. Such a transformation shows again that we can estimate the prior parameters $\{\beta_{st}\}$ by Hebbian learning.

The following filter parameters are assumed

$$\boldsymbol{\beta} = \{\beta_{st}\} = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 \\ \beta_4 & 0 & \beta_5 \\ \beta_6 & \beta_7 & \beta_8 \end{bmatrix}. \tag{23}$$

The Hamiltonian term in Eq. 21 can be further rewritten as

$$H_l = -\sum_s l_s H_{l_s} = -\sum_s l_s (l_{\Xi_s} \cdot \boldsymbol{\beta}), \tag{24}$$

which shows that the local field potential of site s, H_{l_s}, is the cross-correlation of each local block with the same filter $\boldsymbol{\beta}$. Thus, the remaining task is to learn the filter $\boldsymbol{\beta}$.

Based on the transformation, we propose an estimation algorithm for the filter coefficients by local-autoencoding. The central idea of LAE is to *encode the label of each pixel by the labels of its neighboring pixels*, see Fig. 2. It should be noted that LAE approximates the label field by itself, so the two layers are identical. In the experiments, we only have to store one label field. The top layer in Fig. 2 is plotted merely for visualization.

If the output of pixel s is

$$\ell_s = f\left(l_{\Xi_s} \cdot \boldsymbol{\beta}\right), \tag{25}$$

where f is a sigmoid function,

$$f\left(x\right) = \frac{2}{1+e^{-x}} - 1. \tag{26}$$

Defining the error energy $e_s = \frac{1}{2}(\ell_s - l_s)^2$, we derive a learning rule according to the BP algorithm

$$\Delta\beta_i = -\,\eta\left(\ell_s - l_s\right)\frac{1}{2}\left(1-\ell_s^2\right) l_{s_i}. \tag{27}$$

where η is the learning rate.

2) Potts model: The Potts state q is a symbol rather than a real value. Thus, the problem arises how to transform the labels such that they become suitable for Hebbian learning. Eq. 20 implies that we should employ $Q-1$ transformations (except $q = l_t$). For each vertex t in the neighborhood of the current vertex s, three cases need to be considered, $l_t = q$, $l_t = l_s$ and the remaining ones. When encoding the center vertex s with the vertexes in the local 2nd-order neighborhood, only vertexes that have the same label contribute positively to the center vertex.

Based on these considerations, we designed the following transformation strategies: (1) The center label is fixed to 1; (2) All neighboring labels that are the same as the center label are transformed to 1, while those labels that are equal to the current transformation index q turn to -1. (3) The neighboring label that is neither equivalent to the center label nor equivalent to current q has nothing to do with the maximization process, we set it to 0 directly. Formally, we transform the block labels from $\{1, ..., Q\}$ to $\{-1, 0, 1\}$ by

$$\begin{cases} l'_s & = 1 \\ l'_t & = \begin{cases} 1 & \text{if } l_t = l_s \\ -1 & \text{if } l_t = q \\ 0 & \text{otherwise} \end{cases} \end{cases} \tag{28}$$

for all $q \neq l_s$. Note that the 0-state does not exist in the Ising field. Simply, each block is transformed to $Q-1$ binary blocks with each label being in $\{-1, 0, 1\}$.

A transformation example is given in below ($Q = 4$),

$$\begin{bmatrix} 2 & 2 & 1 \\ 1 & 3_s & 1 \\ 4 & 4 & 3 \end{bmatrix} \left\langle \begin{array}{l} q = 1 \\ q = 2 \\ q \neq 3_s \\ q = 4 \end{array} \right. \begin{array}{c} \begin{bmatrix} 0 & 0 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \\ \begin{bmatrix} -1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \\ \begin{bmatrix} 0 & 0 & -1 \\ -1 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \tag{29}$$

Now, the learning rule, Eqs. 25 and 27, for the Ising case is available also for the Potts model.

The detailed algorithm is shown as Algorithm 1. Other learning strategies, like *conjugate gradient* method or *Newtonian descent* algorithm [31], may converge more quickly, but for simplicity, only a primitive BP learning rule is used here.

It should be noted that the output of the LAE algorithm is the average of the last τ steps. When $T > 1$, we set τ to the size of the image, $\tau_1 = (H-2)(W-2)$. When $T = 1$, we choose $\tau_2 = \tau_1/3$. As it will become clear in the experiments below, a smaller τ would be adequate for practical applications.

V. Experiments on Parameter Estimation

In this section, we use four experiments to test the performance of the proposed LAE method. The former three experiments are conducted on the label fields, which are generated by the Gibbs sampling with known prior parameters (completely observable data). The convergence speed and the Cramer-Rao bound for isotropic case are presented in Sections V-A and V-B respectively. In section V-C, we show that LAE can also be used to estimate the anisotropic interaction parameters. In the last experiment (Section V-D), we try to segment the real sonar image by incorporating LAE into the ICM process.

A. Convergence

In the first experiment, we use different algorithms listed in Section III to estimate the prior parameters of a known label field. Fig. 3(a) and (b) show two samples ($Q = 2$, size 256×256) generated by the Gibbs Sampling [26].

The existing methods, including MCMC methods and the heuristic methods, have to scan the whole field for multiple times. As it can be seen in Fig. 3(e)–(g), the scan times needed for MPL, Coding, MCMC, MCMCML are about 40,100,5 and 5 respectively. However, the proposed LAE method quickly converge to the vicinity of the true value even only ~ 1000 vertexes (about $\frac{1}{60}$ field) have been visited. Note that MCMC methods have to sample at least a label field in each run.

The estimation results for Fig. 3(a) are MPL ($\hat{\beta} = 0.7001$), Coding ($\hat{\beta} = (0.7120 + 0.6950 + 0.6901 + 0.7037)/4 = 0.7002$, 4 sets), MCMC ($\hat{\beta} = 0.7024$), MCMCML ($\hat{\beta} = 0.6988$) and LAE ($\hat{\beta} = 0.7003$). The results for Fig. 3(b) are MPL ($\hat{\beta} = 1.0147$), Coding ($\hat{\beta} = (1.0551 + 1.0165 + 1.0001 + 0.9910)/4 = 1.0157$, 4 sets),

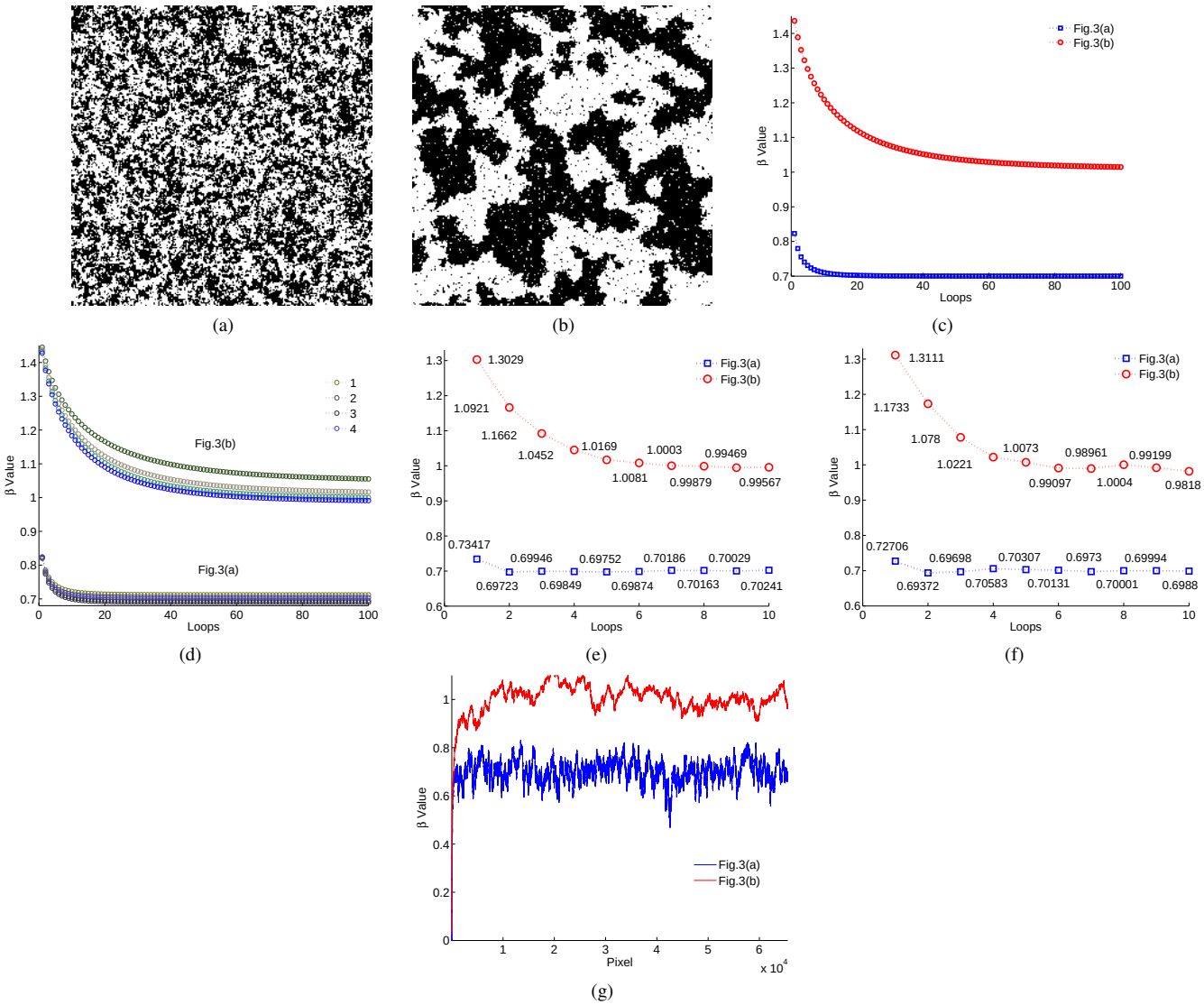

Fig. 3. Convergence speed. (a) and (b) are label field samples generated with $\beta = 0.7$ and $\beta = 1.0$, respectively, by Gibbs sampling; (c)–(g) are the prior parameter estimation process of MPL, Coding, MCMC, MCMCML and LAE, respectively.

MCMC ($\hat{\beta} = 0.9957$), MCMCML ($\hat{\beta} = 0.9818$) and LAE ($\hat{\beta} = 1.0026$). Both show that the LAE method has comparable performance with the existing methods. τ_2 is used to obtain the final output in LAE here. The results demonstrate that the proposed LAE algorithm is superior in the convergence speed and, at the same time, produces accurate estimates.

B. Cramer-Rao Bound for isotropic β

In order to assess the performance of the LAE and the other estimators, it is interesting to compare the mean-squared-error (MSE) of each estimator with the CRB for β, which provides a lower bound on the MSE of the ML estimator.

1) CRB vs. iterative methods: The theoretical calculation is available only for the Ising field ($Q = 2$) with circular boundary conditions [12], see the solid-black curve in Fig. 4(b). However, Pereyra et al. proposed that the computationally intractable CRB for a Potts model can be addressed by a Monte-Carlo integration [32]. That is,

$$\mathrm{CRB}_{\mathrm{approx}} = -\log\left(\mathrm{Var}\left(\tilde{\mathrm{N}}\mathrm{l}\right)\right), \tag{30}$$

where $\tilde{\mathrm{N}}\mathrm{l}$ is a vector with length $50,000$. 'Var' stands for the variance. In Fig. 4(b), $\mathrm{CRB}_{\mathrm{approx}}$ curves are plotted in dotted-cross. To assess the performance, we calculate the logarithm of the MSE for each estimator,

$$\mathrm{logMSE}_{\mathrm{estimator}} = \log\left\{\mathrm{E}\left[\left(\tilde{\hat{\beta}} - \beta_{\mathrm{generate}}\right)^2\right]\right\}, \tag{31}$$

where $\hat{\beta}$ is the estimation value.

Results in Fig. 4 compare the $\mathrm{logMSE}_{\mathrm{estimators}}$ with the theoretical value and the $\mathrm{CRB}_{\mathrm{approx}}$. For each $\beta_{generate}$, $50,000$ samples are generated along the Markov chains with a Gibbs sampler. Similar to [32], the samples are defined on a toroidal graph of size $H \times W = 32 \times 32$. For each sample, we count its homogeneous clique number N_1, and estimate the prior parameter $\hat{\beta}$ with different methods. For a comparison, we also show the results when $Q = 3$ in Fig. 5.

In Fig. 4(b), $\mathrm{CRB}_{\mathrm{approx}}$ varies largely around $\beta = 0.9$ for the Ising field, which coincides well with the theoretical phase-transition temperature $\beta_c = \log(1 + \sqrt{Q}) = 0.8814$. This also

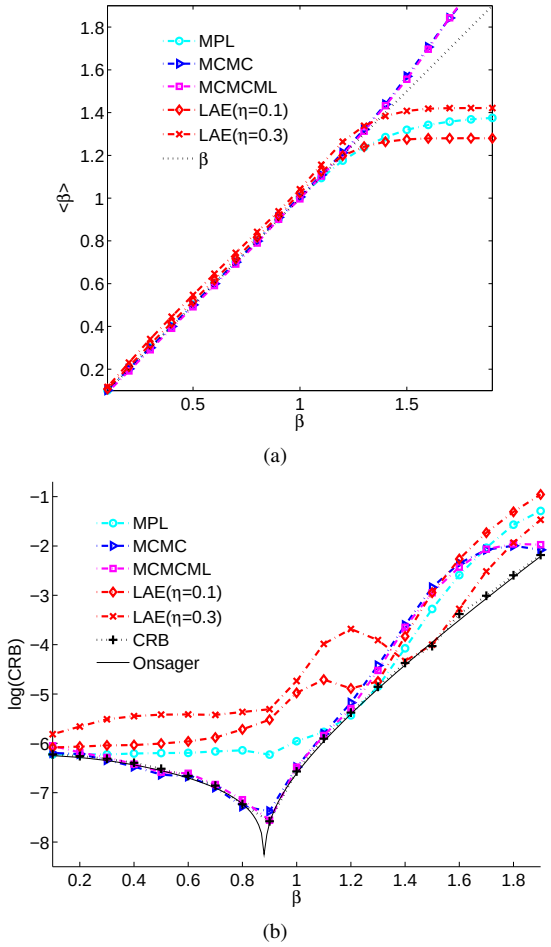

Fig. 4. CRB for an Ising model ($Q = 2$, size=32×32). "Onsager" denotes the theoretical value provided in [12], 'CRB' curve is the $\mathrm{CRB}_{\mathrm{approx}}$ calculated by Eq. 30. Other values are $\mathrm{logMSE}_{\mathrm{estimators}}$ calculated by Eq. 31.

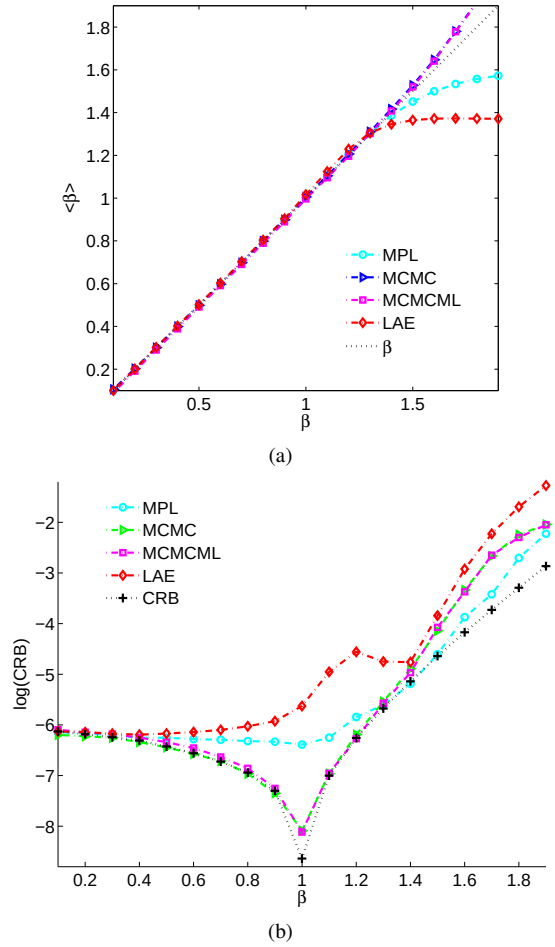

Fig. 5. CRB for a 3-state Potts model (size=32×32).

holds for Fig. 5(b). Five other observations can be drawn from the results: Firstly, the mean of LAE (red dotted-diamond in Fig. 5(a)) saturates a little earlier than other methods when $\eta = 0.1$. The corresponding $\mathrm{logMSE}_{\mathrm{estimator}}$ begins to increase quickly then, i.e. around $\beta = 1.2$. Secondly, the LAE method has comparable performance with other algorithms when β is not very large, which demonstrates that LAE performs well in $[0, 1]$ when $Q = 2, 3$ and size$= 32 \times 32$. Thirdly, a larger learning rate i.e. $\eta = 0.3$ (red cross in Fig. 5(a)), obtains a larger MSE. According to our experiments, $\eta = 0.1$ is selected in the following simulations. Fourthly, the logMSE of the LAE method is a little larger than the MPL algorithm when $\beta > 0.7$. However, they have comparable performance when $\beta \leq 0.7$. After all, they are both originated from the conditional independence assumption. Lastly, the sampling methods have better performance when $\beta < 1.2$. However, all methods degenerate with the same tendency since then, because of the the homogeneous label fields.

2) CRB vs. label field size: The estimation process of LAE makes full use of all kinds of the block configuration information in the label field. It means that a larger label field, which has a larger probability containing all the necessary blocks with the given β, may lead to a smaller estimation bias. Consequently, a smaller MSE would be obtained. To test the hypothesis, we plot the mean and the logMSE of the proposed LAE estimator for different label field sizes in Fig. 6 ($Q = 3$).

Indeed, Fig. 6(a) demonstrates that the bias decreases gradually when the label field size grows. The reason is very simple. When β becomes very large, each sample becomes an almost homogeneous label field. The difference in β can only be reflected by some minor specific configurations. However, these configurations appear by probabilities, which is induced by Gibbs sampling. A larger label field size means an increase of the chance of those specific configurations. Therefore, the estimation approaches the true parameter better.

Consequently, $\mathrm{logMSE}_{\mathrm{estimators}}$ (markers filled with white in Fig. 6(b)) decrease gradually when $\beta < 1$. However, they grow almost linearly with the same gradient when β becomes larger, because N_1 no longer grows. For the same reason, $\mathrm{CRB}_{\mathrm{approx}}$ (markers filled with solid black in Fig. 6(b)) grows gradually.

3) CRB vs. Potts number: In the last test (see Section V-B2), we found that the performance of the proposed LAE estimator is dominated by the minor "specific" configurations and the number of homogeneous cliques when β becomes very large. Both of them grow with the Potts number Q, it is expected that the mean of the LAE estimator will increase with Q, while the $\mathrm{CRB}_{\mathrm{approx}}$ will decrease with Q.

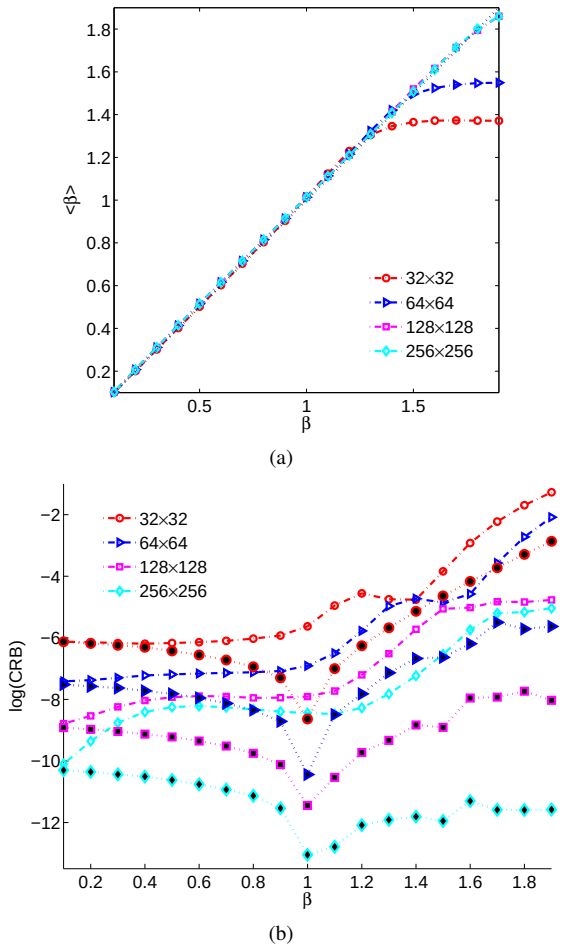

Fig. 6. The relationship between the CRB and the sample scales ($Q = 3$). In (b), all markers filled with solid black are the approximative CRB, others are the logMSE of each estimator.

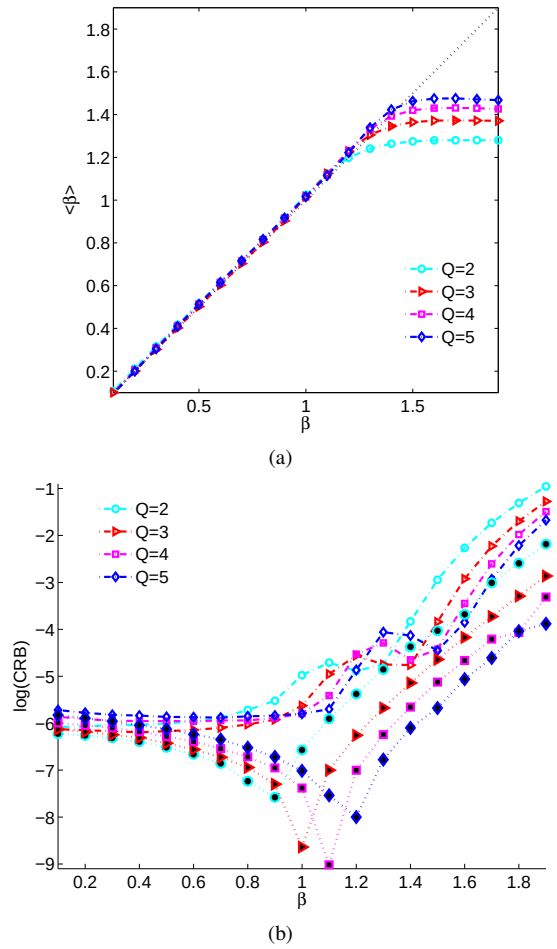

Fig. 7. The relationship between the CRB and the number of Potts states (size=32×32).

The dependence relationship between the mean, the CRB, the MSE and the Potts number are plotted in Fig. 7. It can be seen from Fig. 7(a) that the mean indeed approaches the true generation parameter when Q increases. The reason is that the label field becomes more and more *rugged*. Therefore, the same graph is more likely to have more "specific" information. However, the limited size, 32×32 is insufficient to provide all the "specific" configurations, leading to an estimate with relatively larger bias. Therefore, $\text{logMSE}_{\text{estimator}}$ decreases with Q, and grows linearly when β is very large.

TABLE I
ESTIMATION RESULT OF ANISOTROPIC PRIOR PARAMETERS WITH CENTRAL SYMMETRY

Fig. 8	Estimation	β_1	β_2	β_3	β_4
	True β	-1	1	1	2
(a)	$\hat{\boldsymbol{\beta}}$(LAE1)	-0.9215	0.8709	1.1140	2.0021
	$\hat{\boldsymbol{\beta}}$(LAE10)	-1.1204	1.0423	1.0320	2.1079
(c)	$\hat{\boldsymbol{\beta}}$(LAE1)	-0.9655	0.9830	0.9931	1.9400
	$\hat{\boldsymbol{\beta}}$(LAE10)	-1.0514	1.0395	0.9973	1.9780
(e)	$\hat{\boldsymbol{\beta}}$(LAE1)	-1.0155	0.9731	1.0610	1.9758
	$\hat{\boldsymbol{\beta}}$(LAE10)	-1.0317	1.0134	1.0029	2.0096
(g)	$\hat{\boldsymbol{\beta}}$(LAE1)	-0.8926	0.9412	1.0214	1.9152
	$\hat{\boldsymbol{\beta}}$(LAE10)	-0.9376	0.9691	1.0034	1.9280

C. Prior parameter estimation when β is anisotropic

We use the LAE method to estimate the anisotropic prior parameters. The performance can be measured by the difference between the estimated parameters $\hat{\boldsymbol{\beta}}$ and the generation parameters $\boldsymbol{\beta}$. To maintain consistency with the literature [23], [24], [33], we first apply the method to estimate the sample generated by anisotropic parameters with central symmetry. To highlight the robustness, accuracy and generality, we design another experiment in which the central-symmetry condition is dropped and the samples are generated by the Gibbs sampler [26].

1) When β is anisotropic but central-symmetric: When β is anisotropic but central-symmetric, there are four interaction parameters to be considered. The four samples shown in Fig. 8(a), (c), (e), (g) are generated with the same prior parameters, but with a number of Potts states increasing from $Q = 2$ to $Q = 5$. The corresponding parameter estimation processes are plotted for $T = 10$ in Fig. 8(b), (d), (f), (h).

In addition, we also analyze the effects of the number of iterations by changing the parameter T in the LAE algorithm. LAE with only one pass ($T = 1$) is termed LAE1, while LAE10 implies 10 passes. The results for LAE1 and LAE10

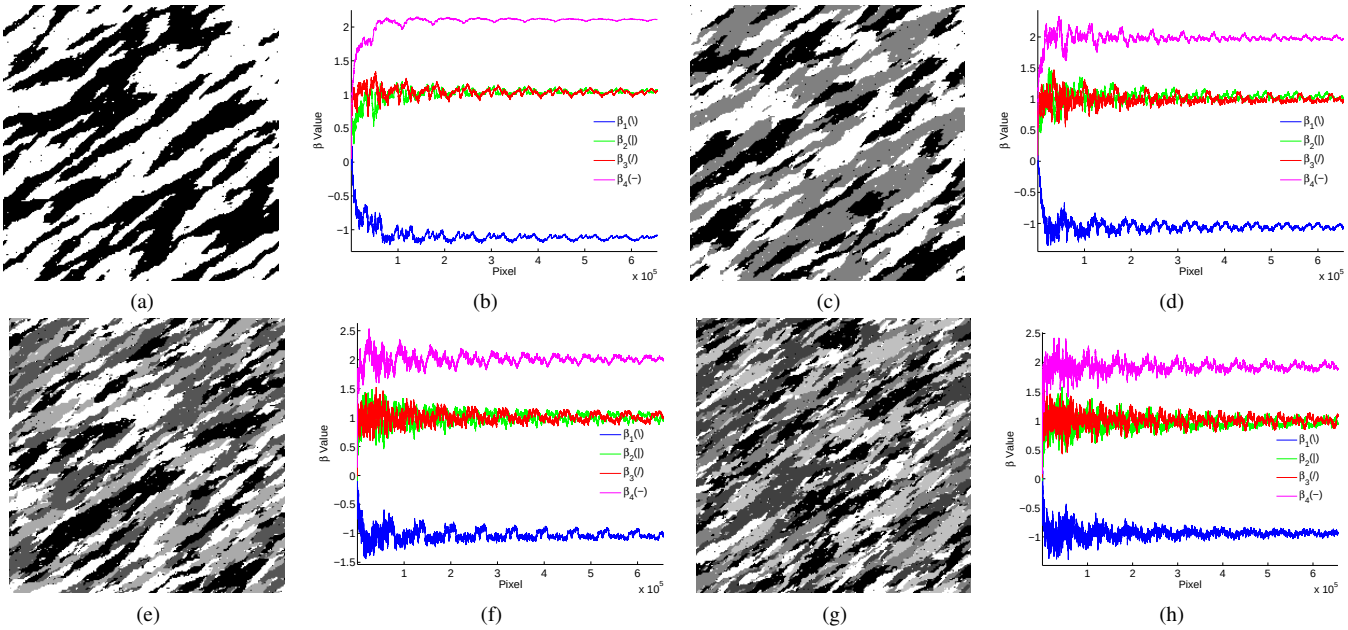

Fig. 8. Estimates of the anisotropic prior parameters with central symmetry by local autoencoding. The Potts states of the samples shown in (a), (c), (e), (g) are $Q = 2, 3, 4, 5$ respectively. They are generated by the same 4 interaction parameters, and the parameters are shown in Table I; (b), (d), (f), (h) shows the estimation process of the prior parameters ($T = 10$). (\), (|), (/) and (−) shows the direction of each interaction strength.

are listed in Table I.

Table I shows that there is no obvious difference between the results of LAE1 and LAE10. We define a measure that is similar to the logMSE in Eq.31 to evaluate the performance of the algorithms,

$$\Lambda = \log\left[\frac{1}{\|\beta\|}\sum_i\left(\hat{\beta}_i - \beta_i\right)^2\right], \tag{32}$$

where $\|\boldsymbol{\beta}\|$ is the length of filter $\boldsymbol{\beta}$. The value for LAE1 and LAE10 is respectively, $\Lambda_{LAE1} = -6.8344$ and $\Lambda_{LAE10} = -7.2605$. It demonstrates again that, only within one pass, the proposed LAE method is able to obtain an accurate estimation. We conclude that LAE1 is available for the practical applications, though multiple passes will improve the precision.

2) Fully anisotropic interactions: The LAE algorithm does not require central symmetry, which suggests that LAE may also be used to estimate the samples that are generated by fully anisotropic interaction parameters.

In the experiment, we first use LAE1 and LAE10 with eight parameters to estimate the central-symmetric samples in Fig. 8. This does not only help to evaluate the performance of the proposed method, but also allows us to examine whether the interaction parameters of these samples are indeed strictly central-symmetric. Results are given in Table II. Then, we change the generation parameters (show red in Table III), and use the new anisotropic parameters without central symmetry to generate new samples. Fig. 9 shows the fully anisotropic samples, and Table III lists the estimation results.

Two observations can be drawn from the estimation results: Firstly, even with central-symmetric label fields, like Fig. 8, there exists bias between parameters in the central-symmetric positions, i.e. $\zeta_i = |\beta_i - \beta_{9-i}| \neq 0$, which demonstrates that samples generated by central-symmetric interaction parameters are not strictly regular across the whole image. The larger of ζ_i, the more irregular of the direction of i. Therefore, the bias can be used to measure the regularity in each direction.

Secondly, it is important that for samples generated by fully anisotropic parameters (Fig. 9), the estimates provided by LAE1 and LAE10 are approximately central-symmetric, which shows that *central-symmetry is a feasible assumption for most label fields*. This also demonstrates that samples generated by central-asymmetric parameters are unsuitable to be described by the original generation parameters. In other words, each sample can be generated by more than one set of prior parameters, which is expected to be due to the mixing effects of central-asymmetric forces.

D. Segmentation of incompletely observable data

In the end, we apply the LAE method to incompletely observable data, i.e. sonar images by integrating it into the ICM method. ICM is selected for its low computation cost [34]. Being a local optimization method, it may converge to a local extreme. Better optimization can be expected when global methods are applied, such as Graph Cuts, Simulated Annealing, Genetic algorithms [26]. The ICM method starts from an initialized label field, then loops between parameter estimation and label updating until a balanced Markov chain is obtained. In each iteration, LAE and ML are used to estimate the prior parameters and conditional parameters respectively, and the Gibbs sampling moves the label field towards the balance state.

An important task in sonar image segmentation is the extraction of the shadow areas. The first column of Fig. 10 shows three pictures that are taken by a forward-looking sonar (Fig. 10(a)), a side-scan sonar (Fig. 10(b)) and a multi-beam high-resolution sonar (Fig. 10(c)). The remaining columns of

TABLE II
ESTIMATE CENTRIC-SYMMETRIC LABEL FIELD SAMPLES IN FIG. 8 WITH 8 PARAMETERS WITH LAE10

Estimation	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
True β	-1	1	1	2	2	1	1	-1
$\hat{\beta}$(a)	-1.1354	1.0729	1.0373	2.1266	2.1774	1.0491	1.0695	-1.1768
$\hat{\beta}$(c)	-1.0943	1.0318	1.0085	2.0262	1.9652	1.0058	1.0656	-1.0306
$\hat{\beta}$(e)	-1.0386	1.0303	1.0066	2.0108	2.0333	1.0176	1.0099	-1.0400
$\hat{\beta}$(g)	-0.9544	0.9509	1.0144	1.9552	1.9281	1.0130	0.9986	-0.9321

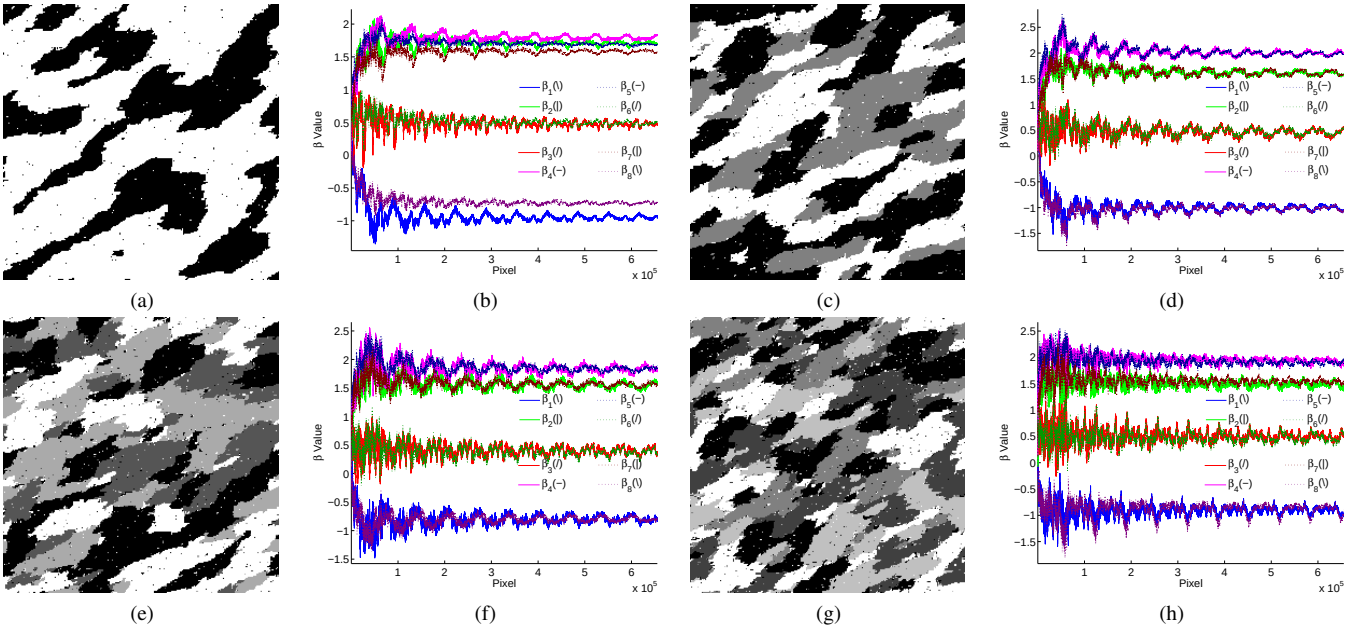

Fig. 9. Fully anisotropic prior parameters estimation by local autoencoding. The Potts states of the samples shown in (a), (c), (e), (g) are $Q = 2, 3, 4, 5$ respectively. They are generated by the same 8 interaction parameters, and the parameters are shown in Table III; (b), (d), (f), (h) shows the estimation process of the prior parameters ($T = 10$).

TABLE III
ESTIMATION RESULTS FOR FULLY ANISOTROPIC PRIOR PARAMETERS (THE RED TRACE HIGHLIGHTS THE CHANGES COMPARED TO THE CENTRIC-SYMMETRIC CASE.)

Fig. 9	Estimation	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
	True β	-1	1	1	2	2	1	-1	1
(a)	$\hat{\beta}$(LAE1)	-1.0181	1.7366	0.5269	1.9241	1.8344	0.6372	1.5091	-0.6504
	$\hat{\beta}$(LAE10)	-0.9501	1.7049	0.4741	1.7872	1.6881	0.4854	1.5840	-0.7290
(c)	$\hat{\beta}$(LAE1)	-1.1792	1.7024	0.6341	2.2634	2.3668	0.7214	1.7127	-1.2571
	$\hat{\beta}$(LAE10)	-1.0002	1.6082	0.4533	1.9899	1.9858	0.4634	1.6087	-1.0085
(e)	$\hat{\beta}$(LAE1)	-0.9458	1.8845	0.4816	1.9552	2.0906	0.5390	1.7335	-0.8670
	$\hat{\beta}$(LAE10)	-0.7944	1.5394	0.4001	1.8273	1.8288	0.3736	1.5590	-0.7948
(g)	$\hat{\beta}$(LAE1)	-1.0013	1.5831	0.6120	2.0285	2.1246	0.6439	1.6932	-1.1361
	$\hat{\beta}$(LAE10)	-0.9006	1.4897	0.4979	1.9358	1.8995	0.4724	1.5413	-0.8857

Fig. 10 show the segmentation results by different methods. The raw images are first segmented by the ICM procedure with different prior parameter estimation methods. In the post-processing stage, morphological operators, such as image erosion and dilation, are adopted sequentially to get rid of the pepper-and-salt noise. In the results, candidate areas larger than 100 pixels are displayed with different colors.

We compare the results with the frequently used LSQR method [5], [23], [24], [33], [35]. In Fig. 10, "LAE10_8" discriminates background speckles from foreground objects more precisely, while "LSQR_8" is unsuitable for real image segmentation. Further, "LAE1_4", "LAE1_8", "LAE10_4" provide very similar segmentation results, which not only proves that the LAE algorithm with $T = 1$ and four prior parameters is already sufficient for image segmentation, but also highlights the computation superiority of LAE. The most important is that, LAE can be used to estimate arbitrary anisotropic samples.

The segmentation results of the EM procedure with mean-field approximation (EMMF for short, Section III-C) [22] (also for $Q = 2$) are shown in the last column of Fig. 10. With the forward-looking and the side-scan sonar images, the shadow areas mix with the the background areas. This implies that "smoothing" is undesirable for the segmentation of sonar

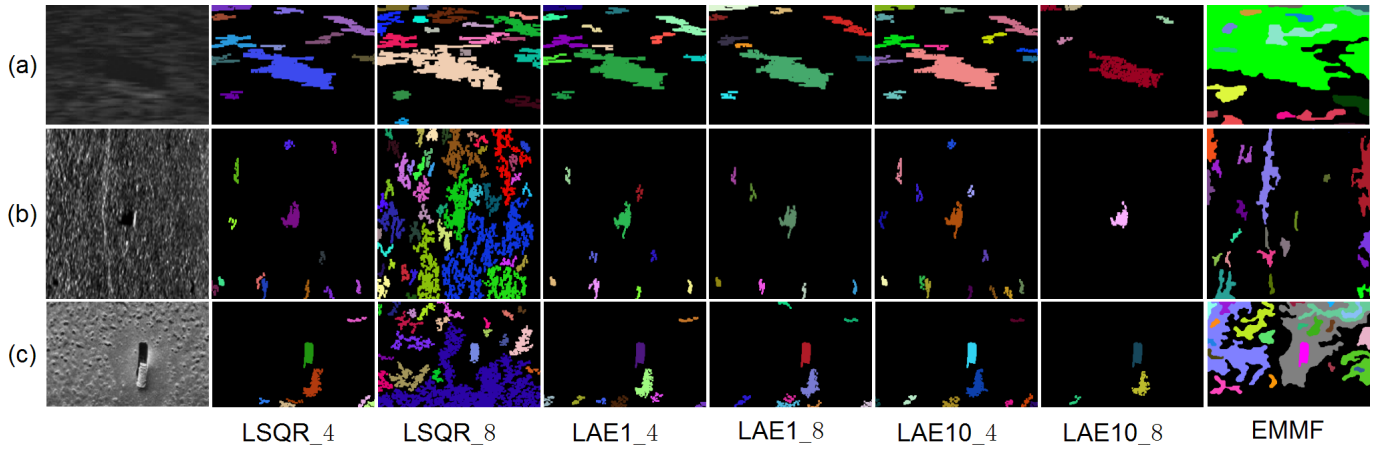

Fig. 10. Candidate shadow areas from sonar image extracted by the ICM method. Columns correspond to variants of the the LSQR algorithm and the LAE method used for estimate parameters, and the EMMF method which uses the mean field approximation algorithm. The variants are defined by the number of neighborhood parameters and the number of loops, e.g. "LAE1_4" indicates $T = 1$ and $\|\boldsymbol{\beta}\| = 4$ for the LAE algorithm.

images.

VI. Conclusions and Discussions

In this paper, a novel prior parameters estimation method, named local autoencoding (LAE), for MRF model is developed. LAE makes full use of the local block information in the image and learns the interaction parameters block by block with simple BP learning rules.

The experimental results illustrate that it converges rapidly to the true parameter, while provides an accurate estimate at the same time, which are favorable for time-critical practices. Besides, LAE has low computation and memory costs. In addition, LAE is attractive in several other aspects:

Firstly, LAE is a local-learning algorithm, which means that it is capable of estimating the local interaction parameters in any given specified area. In other words, LAE is a local self-adaptive algorithm, and it can be used to measure the differences of interaction parameters across the whole image.

Secondly, LAE is an incremental-learning algorithm. The estimated values from the current iteration can be used as the initial values for the parameters in the next iteration. In the ICM iteration, only those pixels that have label switchings in Gibbs sampling process should be chosen to improve the filter by incremental learning.

Thirdly, LAE is not limited to the specific neighborhood system that we have used here. Higher-order neighborhoods help to preserve finer structures [36], [37]. In that case, we only need change the size of the filter, and LAE algorithm is readily applicable then.

Lastly, a major advantage of the LAE algorithm is that it presents a general framework for prior parameter estimation. It can be used to estimate the interaction parameters for other Potts models. For example, the prior parameters of the above Potts model are only determined by the relative site relationships, while Kanter's model [38] parameters in Appendix A are also determined by the labels between neighboring pixels. Descombes' Potts model [16] is an exceptional example, where interaction parameters, βs, are no longer local spatial-invariant (e.g. homogeneous), they are determined by a set of variables (e.g. inhomogeneous) [16]. In this case, we may use the same transformation rule in Eq.(28), and then take the derivative of the error energy with respect to each variable. Result shows that the LAE algorithm is feasible for Descombes' Potts model (data is not shown). There are also other kinds of Potts models, like [39], [40], their prior parameters are homogeneous and can be estimated by similar extensions. Although these models deserve a more detailed consideration and a comparative evaluation, the full treatment is beyond the scope of this paper.

Appendix A
Extension to Kanter's Potts model

In Kanter's model [38], each Potts unit can be in a number Q of distinct active states with equal probability $1/Q$. The Hamiltonian with only local interactions can be written as

$$H = -\frac{1}{2}\sum_{s}\sum_{t\in\Xi_s}\sum_{v,p=1}^{Q}\beta_{st}^{vp}u_{l_s v}u_{l_t p} \tag{33}$$

where $u_{l_s v} = \delta_{l_s v} - \frac{1}{Q}$ is the interaction operator between states l_s and v. There are $8Q^2$ interaction parameters to be estimated. To further reduce the computation and ease the consideration, we take two simplified strategies:

1) β_{st}^{vp} is isotropic, $\beta_{st}^{vp} \equiv \beta^{vp}$;
2) Omit the constant $-\frac{1}{Q}$ in the interaction operator.

Similary, when maximizing the conditional probability

$$P\left(l_s\,|l_{\Xi_s}\right) = \left\{1 + \sum_{q\neq l_s} e^{\sum\limits_{l_t=p\&q=v}\beta^{vp} - \sum\limits_{l_t=p\&l_s=v}\beta^{vp}}\right\}_{t\in\Xi_s}^{-1}, \tag{34}$$

β^{vp} should increase when $l_s = v$ and $l_t = p$, and decrease otherwise. Note that the first term in the exponent has nothing to do with β^{vp}. It is also similar to the Hebbian learning and the LAE algorithm is applicable now. Then, the transformation

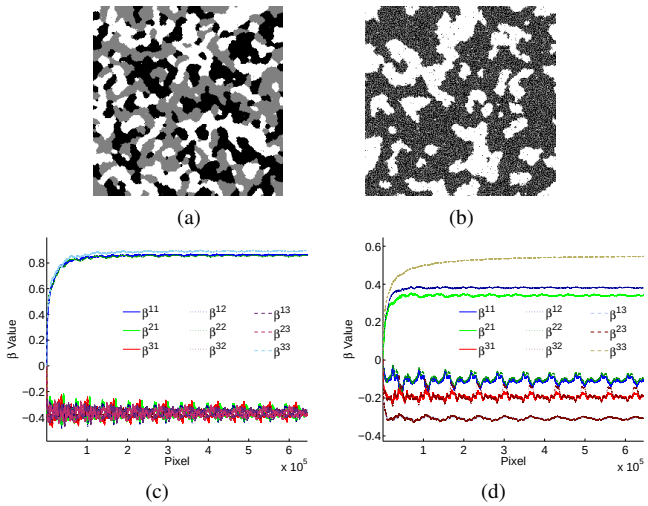

Fig. 11. Prior parameter estimation for Kanter's Potts model ($Q = 3$). (a) and (b) are samples, (c) and (d) are the corresponding parameter estimation process using the LAE algorithm ($T = 10$).

rule from Potts states to $\{-1, 1\}$ is

```
for s
  l_s == p?l'_s = 1 : l'_s = -1;
for t ∈ Ξ_s
  l_t == q?l'_t = 1 : l'_t = -1;
(Optional)
if l'_s + Σ_{t∈Ξ_s} l'_t = -9
  discard the block;
end
```

A transformation example ($Q = 3$) is shown below,

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 3 \end{bmatrix} \begin{matrix} \nearrow \\ \searrow \end{matrix} \begin{matrix} \beta^{23} \begin{bmatrix} -1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix} \surd \\ \beta^{32} \begin{bmatrix} -1 & -1 & -1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{bmatrix} \bigcirc \end{matrix} \tag{35}$$

The same block is necessary in estimating β^{23}, but is optional in estimating β^{32}.

Similar to Eq.24, the output of pixel s when we estimate $\boldsymbol{\beta}^{pq}$ is

$$\ell_s^{vp} = f\left(\sum_{t\in\Xi_s} \beta_{st}^{vp} l'_s l'_t\right) \tag{36}$$

The error energy $e_s^{vp} = \frac{1}{2}(\ell_s^{vp} - l_s^{vp})^2$. The gradient descent rule for β_i^{vp} is

$$\Delta\beta_i^{vp} = -\eta\,(\ell_s^{vp} - l_s^{vp})\,\frac{1}{2}(1 - l_s^{vp})^2 \sum_{t\in\Xi_s} l'_s l'_t \tag{37}$$

Metropolis sampler is adopted for the label field generation with the single-site energy being $\mathcal{E}_s = \sum_{t\in\Xi_s}\sum_{v,p} \beta_{st}^{vp} u_{l_s v} u_{l_t p}$. Two samples generated by the simplified Kanter's Potts model and their prior parameter estimation process is shown in Fig.11. The detailed generation parameters and estimation value are listed in Table IV. The results show that the LAE algorithm can be used to estimate Kanter's Potts model by simple label transformations.

TABLE IV
ESTIMATION RESULTS OF THE KANTER'S POTTS MODEL IN FIG.11

Fig.11	$\boldsymbol{\beta}^{vp}$	$\hat{\boldsymbol{\beta}}^{vp}$
(a)	$\begin{bmatrix} 0.90 & -0.30 & -0.30 \\ -0.60 & 0.70 & -0.60 \\ -0.30 & -0.30 & 0.90 \end{bmatrix}$	$\begin{bmatrix} 0.84 & -0.36 & -0.36 \\ -0.37 & 0.85 & -0.36 \\ -0.37 & -0.36 & 0.84 \end{bmatrix}$
(b)	$\begin{bmatrix} -0.10 & 0.50 & -0.35 \\ 0.10 & -0.05 & -0.40 \\ -0.20 & -0.15 & 0.50 \end{bmatrix}$	$\begin{bmatrix} -0.11 & 0.38 & -0.31 \\ 0.34 & -0.10 & -0.31 \\ -0.19 & -0.18 & 0.55 \end{bmatrix}$

ACKNOWLEDGMENT

The authors are grateful to *Marcelo Pereyra* for discussions about the estimation of parameters of Markov random fields, to *Prof. Yandong Tang* for advices on improving the paper, and to the anonymous reviewers for valuable comments on the previous versions of the paper.

REFERENCES

[1] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 721–741, 1984.
[2] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001.
[3] N. Komodakis, N. Paragios, and G. Tziritas, "MRF optimization via dual decomposition: Message-passing revisited," in *Computer Vision, IEEE 11th International Conference on*, 2007, pp. 1–8.
[4] C. Wang, N. Komodakis, and N. Paragios, "Markov random field modeling, inference & learning in computer vision & image understanding: A survey," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1610–1627, 2013.
[5] M. Mignotte, C. Collet, P. Perez, and P. Bouthemy, "Sonar image segmentation using an unsupervised hierarchical MRF model," *Image Processing, IEEE Transactions on*, vol. 9, no. 7, pp. 1216–1231, 2000.
[6] O. Eches, J. A. Benediktsson, N. Dobigeon, and J.-Y. Tourneret, "Adaptive markov random fields for joint unmixing and segmentation of hyperspectral images," *Image Processing, IEEE Transactions on*, vol. 22, no. 1, pp. 5–16, 2013.
[7] M. Gong, L. Su, M. Jia, and W. Chen, "Fuzzy clustering with a modified mrf energy function for change detection in synthetic aperture radar images," *Fuzzy Systems, IEEE Transactions on*, vol. 22, no. 1, pp. 98–109, 2014.
[8] Y. Zhang, B. Michael, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 1, pp. 45–57, 2001.
[9] M. T. Moores, A. N. Pettitt, and K. Mengersen, "Scalable Bayesian inference for the inverse temperature of a hidden Potts model," pp. 1–30, 2015.
[10] M. Pereyra, N. Dobigeon, H. Batatia, and J.-Y. Tourneret, "Estimating the granularity coefficient of a Potts-Markov random field within a Markov chain Monte Carlo algorithm," *Image Processing, IEEE Transactions on*, vol. 22, no. 6, pp. 2385–2397, 2013.
[11] S. Song, B. Si, X. Feng, and J. M. Herrmann, "Prior parameter estimation for Ising- MRF-based sonar image segmentation by local center-encoding," in *MTS|IEEE Oceans*, 2015, pp. 1–5, doi:10.1109/OCEANS-Genova.2015.7271429.
[12] J.-F. Giovannelli, "Estimation of the Ising field parameter thanks to the exact partition function," in *ICIP*, 2010, pp. 1441–1444.
[13] C. A. McGrory, D. M. Titterington, R. Reeves, and A. N. Pettitt, "Variational Bayes for estimating the parameters of a hidden Potts model," *Statistics and Computing*, vol. 19, no. 3, pp. 329–340, 2009.
[14] L. Younes, "Maximum likelihood estimation for Gibbsian fields," *Lecture Notes-Monograph Series*, pp. 403–426, 1991.

[15] C. J. Geyer, "Practical Markov chain Monte Carlo," *Statistical Science*, pp. 473–483, 1992.

[16] X. Descombes, R. D. Morris, J. Zerubia, and M. Berthod, "Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood," *Image Processing, IEEE Transactions on*, vol. 8, no. 7, pp. 954–963, 1999.

[17] J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen, "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants," *Biometrika*, vol. 93, no. 2, pp. 451–458, 2006.

[18] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236, 1974.

[19] T. Rydén and D. Titterington, "Computational bayesian analysis of hidden markov models," *Journal of Computational and Graphical Statistics*, vol. 7, no. 2, pp. 194–211, 1998.

[20] J. Besag, "Efficiency of pseudolikelihood estimation for simple Gaussian fields," *Biometrika*, pp. 616–618, 1977.

[21] J. Zhang, "The mean field theory in EM procedures for blind Markov random field image restoration," *Image Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 27–40, 1993.

[22] G. Celeux, F. Forbes, and N. Peyrard, "EM procedures using mean field-like approximations for Markov model-based image segmentation," *Pattern recognition*, vol. 36, no. 1, pp. 131–144, 2003.

[23] H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using Gibbs random fields," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 1, pp. 39–55, 1987.

[24] M. Mignotte, C. Collet, P. Pérez, and P. Bouthemy, "Three-class Markovian segmentation of high-resolution sonar images," *Computer Vision and Image Understanding*, vol. 76, no. 3, pp. 191–204, 1999.

[25] J. Besag, "Statistical analysis of non-lattice data," *The Statistician*, pp. 179–195, 1975.

[26] S. Z. Li and S. Singh, *Markov random field modeling in image analysis*. Springer, 2009, vol. 26.

[27] M. Pereyra, N. Whiteley, C. Andrieu, and J.-Y. Tourneret, "Maximum marginal likelihood estimation of the granularity coefficient of a Potts-Markov random field within an MCMC algorithm," in *IEEE Workshops on Statistical Signal Processing (SSP)*, 2014, pp. 121–124.

[28] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, "Approximate Bayesian computational methods," *Statistics and Computing*, vol. 22, no. 6, pp. 1167–1180, 2012.

[29] I. Murray, Z. Ghahramani, and D. MacKay, "Mcmc for doubly-intractable distributions," in *In Proc. $22^n d$ Conf. UAI, Arlington*. AUAI Press, 2006, pp. VA: 359–366.

[30] J. G. Propp and D. B. Wilson, "Exact sampling with coupled markov chains and applications to statistical mechanics," *Random structures and Algorithms*, vol. 9, no. 1-2, pp. 223–252, 1996.

[31] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Inc., New Jersey, 1999.

[32] M. Pereyra, N. Dobigeon, H. Batatia, and J.-Y. Tourneret, "Computing the Cramer–Rao bound of Markov random field parameters: Application to the Ising and the Potts models," *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 47–50, 2014.

[33] S. Reed, Y. Petillot, and J. Bell, "An automatic approach to the detection and extraction of mine features in sidescan sonar," *Oceanic Engineering, IEEE Journal of*, vol. 28, no. 1, pp. 90–105, 2003.

[34] C. Dumontier, F. Luthon, and J.-P. Charras, "Real-time DSP implementation for MRF-based video motion detection," *Image Processing, IEEE Transactions on*, vol. 8, no. 10, pp. 1341–1347, 1999.

[35] Y. Petillot, Y. Pailhas, J. Sawas, N. Valeyrie, and J. Bell, "Target recognition in synthetic aperture and high resolution side-scan sonar," in *European Conference on Underwater Acoustics, ECUA*, vol. 10, 2010.

[36] N. Komodakis and N. Paragios, "Beyond pairwise energies: Efficient optimization for higher-order MRFs," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2009, pp. 2985–2992.

[37] P. Kohli, A. Osokin, and S. Jegelka, "A principled deep random field model for image segmentation," in *Computer Vision and Pattern Recognition, IEEE Conference on*, 2013, pp. 1971–1978.

[38] I. Kanter, "Potts-glass models of neural networks," *Phys. Rev. A*, vol. 37, pp. 2739–2742, 1988.

[39] F. Gerl, K. Bauer, and U. Krey, "Learning with Q-state clock neurons," *Zeitschrift für Physik B Condensed Matter*, vol. 88, no. 3, pp. 339–347, 1992.

[40] S. Song, H. Yao, and A. Y. Simonov, "Latching chains in k-nearest-neighbor and modular small-world networks," *Network: Computation in Neural Systems*, vol. 26, no. 1, pp. 1–24, 2015.

Sanming Song received the M.S. degree from Guangdong University for Technology, China in 2008 and the Ph.D. degree from Harbin Institute of Technology, China in 2013, both in computer science and technology. From 2014-2015, he took a postdoctoral research position in Shenyang Institute of Automation, Chinese Academy of Sciences, where he is currently an associate researcher in the underwater robots laboratory. His interests focus on robotic intelligence, sonar image processing and computational neuroscience.

Bailu Si pursued a PhD degree in theoretical neurophysics in the University of Bremen, Germany in 2007. From 2008-2013, he held postdoctoral researcher positions first in the International School for Advanced Studies in Italy, and later in the Weizmann Institute of Science in Israel. Since December 2013, he is a group leader in the Shenyang Institute of Automation, Chinese Academy of Sciences. His main research interests are spatial cognition, active perception and intelligent robotics.

J. Michael Herrmann received a Doctorate (1993) from the University of Leipzig. His PhD focused on mathematical aspects of artificial neural networks. He was working as a research assistant from 1992 and as a postdoc in Denmark and Japan. In 2008, after a temporary position as an assistant professor at the University of Goettingen, he was appointed to the post of Lecturer in Robotics at the School of Informatics at the University of Edinburgh. His research interests are in self-organization, robot learning, neural avalanches, auditory system, bio-medical data analysis, metaheuristic optimization, sensor fusion and information theory.

Xisheng Feng was born in Haicheng, Liaoning province of China, in 1941. He graduated from the Department of Industrial Enterprises Electrification and Automation in Harbin Institute of Technology, China in 1965. Since 1973, he has been a senior researcher with the underwater robots laboratory at the Shenyang Institute of Automation, Chinese Academy of Sciences.

As a pioneer of the underwater robots in China, he began to design and develop the underwater vehicles in 1982. He was the main designer of "HR-01" (1985), which was the first 200-meter remotely operated vehicle (ROV) in China. Later, he was in charge of China's first 1000-meter autonomous underwater vehicle (AUV), "Explorer" (1994). Prof. Feng was the deputy general designer (technical director) of China's first 6000-meter AUV, "CR-01" (1995), which won the Special Award in Chinese Academy of Sciences. The project of "Research, development and application of the untethered underwater robots" won the First Class Prize of The State Scientific and Technological Progress Award. In 1999, he was elected as an Academician of Chinese Academy of Engineering (CAE).

His main research interests include marine robotics, ROV and AUV, acoustic image processing and intelligent control.