Transformation of Markov Random Fields for Marginal Distribution Estimation

Masaki Saito Takayuki Okatani
Tohoku University, Japan
{msaito, okatani}@vision.is.tohoku.ac.jp

Abstract

This paper presents a generic method for transforming MRFs for the marginal inference problem. Its major application is to downsize MRFs to speed up the computation. Unlike the MAP inference, there are only classical algorithms for the marginal inference problem such as BP etc. that require large computational cost. Although downsizing MRFs should directly reduce the computational cost, there is no systematic way of doing this, since it is unclear how to obtain the MRF energy for the downsized MRFs and also how to translate the estimates of their marginal distributions to those of the original MRFs. The proposed method resolves these issues by a novel probabilistic formulation of MRF transformation. The key idea is to represent the joint distribution of an MRF with that of the transformed one, in which the variables of the latter are treated as latent variables. We also show that the proposed method can be applied to discretization of variable space of continuous MRFs and can be used with Markov chain Monte Carlo methods. The experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

Markov Random Fields (MRFs) have been used for a wide range of problems in computer vision, such as optical flow estimation [6, 26, 25], image restoration [17], bundle adjustment [4, 22], object segmentation [9, 13] etc. There are two types of inference problems for MRFs. One is the MAP (Maximum a Posteriori) inference and the other is the marginal inference problem. In this study we consider the latter, which is to estimate the marginal distributions of MRF variables.

As for the MAP inference problem, there exists many sophisticated algorithms such as sequential tree-reweighted message passing (TRW-S) [11] and FastPD [12]. On the other hand, there are only classical methods for the marginal inference problem, such as mean field (MF) approximation and belief propagation (BP), which usually require a large computational cost. The marginal inference problem is nevertheless important, as it needs to be solved for MPM (maximum posterior marginal) inference [15, 10, 13], learning parameters of conditional random fields (CRFs) [21], and Boltzmann machines [19, 5].

The goal of this study is to provide methods for solving the marginal inference problem more efficiently. As for the MAP inference, a mainstream approach to reduce computational cost is to transform an MRF into a smaller, simpler one. The energy function of the MRF is transformed accordingly and is minimized to find the MAP solution. This approach has been successful in practice, resulting in a number of efficient algorithms. However, the same approach cannot be directly used for the marginal inference problem. In this problem, we are interested in the probabilistic structure (given by the Boltzmann distribution) of the MRF, which needs to be preserved as much as possible before and after transforming the MRF. Otherwise, there is no guarantee that the estimates of the marginal distributions obtained for the transformed MRF well approximate those of the original MRF. Furthermore, it is even unclear how the estimates of the marginal distributions of the transformed MRF can be translated to those of the original MRF. Suppose an image segmentation problem for example. How can we obtain pixel-level marginal distributions from the estimates of the marginal distributions at superpixels? Note that these are not the case with the MAP inference, as it is basically point estimation that can be performed using the energy function alone.

To deal with these difficulties, we propose a novel generic method for transforming MRFs. The key idea is to use the variables of the transformed MRF as latent variables and then represent the joint distribution of the target MRF with them. To be specific, the representation consists of a conditional distribution of the original variables conditioned on the latent variables and their joint distribution. The former conditional distribution is determined by the selected MRF transformation. This formulation enables the direct computation of the energy function of the transformed MRF, which we call the *augmented energy*; this new energy gives the joint distribution of the transformed MRF as its Boltzmann distribution. Then, the marginal distribu-

tions of the transformed MRF are estimated from this joint distribution using any regular algorithm such as BP etc. Finally, the marginal distributions of the original MRF are directly calculated from them. This method is based on the variational principle and has a firm theoretical foundation.

This paper is organized as follows. We present our generic method for MRF transformations in Section 3. We then show three practical applications of the proposed method in Section 4, which are i) discretizing variable space of continuous MRFs, ii) grouping discrete labels of MRFs to reduce the number of labels, and iii) coarse graining of MRFs by grouping multiple sites. In Section 5, we show how some of these MRF transformations are combined to perform coarse-to-fine inference, and also how our MRF transformation approach is applied to Markov chain Monte Carlo methods. Section 6 presents experimental results.

2. Related work

Discretization of continuous MRF Continuous MRFs whose site variables are continuous have only a limited applicability, as the marginal distributions of their variables need to be represented by a limited set of pdfs (e.g., Gaussian distribution). As there is no such limitation for discrete MRFs, it is quite common to formulate problems in discrete domain, even if they are more natural to formulate in continuous domain. However, as is pointed out by Saito et al. [18], a naive discretization of variable space can cause a problem; the estimates can have errors, when the discretization is non-uniform. They extend MF and BP algorithms to be able to properly deal with this. In the present study, we reformulate the discretization as MRF transformation. This enables to deal with a wider class of algorithms, which contains practically any algorithm derived by the variational-principle such as TRW and generalized BP, and also higher-order MRFs [17, 9], both of which cannot be dealt with by their approach.

Grouping of discrete labels The number of labels in discrete MRFs directly affects computational cost. For example, in the case of second-order MRFs, the complexity of BP per one iteration is proportional to $O(KL^2)$, where K is the number of neighboring sites and L is the number of labels. Thus, if we can reduce the number of labels, so does the computational cost. The problem is how we can reduce them while minimizing the loss of accuracy. As far as the MAP inference is concerned, there exist some related studies. Veksler [23] and Wang et al. [24] both proposed heuristic algorithms for reducing the search space of variables for the problem of stereo matching. Yang et. al. [27] also proposed a sophisticated BP algorithm that makes the computational cost independent of L by selecting a few labels having small data cost. However, to the authors' knowledge, there is no study of reducing the number of labels for the marginal inference problem. The above methods cannot be directly applied to the marginal inference problem.

Coarse-graining of MRFs As computational cost also depends on the number of sites and edges between them, it is also effective to apply coarse-graining to MRFs, i.e., transforming their graphs into smaller ones in such a way that a number of connected sites are grouped into a single site. As with the label grouping mentioned above, existing studies are limited for the MAP inference. They are targeted at specific problems such as stereo matching [6, 14, 27] and object segmentation [8]. Although Conejo et. al. [3] proposed a general method for speeding-up MRF optimization by using the coarse-graining and the label pruning methods, their method is only targeted at the MAP inference. As for the marginal inference problem, the only study the authors are aware of is that of Ferriera et al. [7], which considers only Gaussian MRFs, though. There are a few difficulties with using coarse-graining of MRFs for the marginal inference problem. One is how the marginal distributions of the original MRF can be obtained from the estimates of those of a coarse MRF. Another is how the joint distribution (or the energy function) of the coarse MRF can be obtained.

3. General-purpose method for transformation of MRFs

This section presents a general-purpose method for transforming MRFs. Its applications to specific problems will be presented in Section 4.

3.1. Preliminaries

Suppose a general MRF with N sites. Let $\mathcal{G}$ be its graph and $\mathcal{C}$ be the set of factors in $\mathcal{G}$. Each site $i \in \{1, \ldots, N\}$ has a variable x_i defined in space $\mathcal{X}_i$. The space for all the variables $\boldsymbol{x} = [x_1, \ldots, x_N]$ is expressed as $\mathcal{X} = \bigotimes_i \mathcal{X}_i$, where $\bigotimes$ is the Cartesian product. The variables may be either continuous or discrete. We will use the symbol $\sum$ to represent not only a summation over discrete variables but also an integral over continuous variables.

For brevity, we focus on MRFs in what follows. It is noted however that our method is applicable to any graphical models including directed models.

3.2. Minimization of free energy

A variety of algorithms for the estimation of marginal distribution, such as MF, BP, and TRW, can be derived by the same procedure, in which a free energy is minimized based on the variational principle. This subsection summarizes this fundamental methodology.

The probability distribution of a MRF $\mathcal{G}$ is given by

$$p_0(\boldsymbol{x}) = \frac{1}{Z_0} \prod_{c \in \mathcal{C}} \phi_c(\boldsymbol{x}_c), \tag{1}$$

where Z_0 is a normalization constant called a partition function, ϕ_c is the function of the factor c, and $\boldsymbol{x}_c$ is the site variables included in c. Letting $f_c(\boldsymbol{x}_c)$ be the negative logarithm of $\phi_c(\boldsymbol{x}_c)$, i.e., $f_c(\boldsymbol{x}_c) = -\ln \phi_c(\boldsymbol{x}_c)$, we may rewrite p_0 into

$$p_0(\boldsymbol{x}) = \frac{1}{Z_0} \exp(-E_0(\boldsymbol{x})), \tag{2}$$

$$E_0(\boldsymbol{x}) = \sum_{c \in \mathcal{C}} f_c(\boldsymbol{x}_c). \tag{3}$$

As it is generally intractable to directly compute the marginal distributions of the site variables using p_0 defined as above, an arbitrary distribution $q_0(\boldsymbol{x})$ is introduced that approximates $p_0(\boldsymbol{x})$, using which the marginal distributions are approximately computed.

The distribution $q_0(\boldsymbol{x})$ has a certain degree of freedom, within which we search for $q_0(\boldsymbol{x})$ the best approximating $p_0(\boldsymbol{x})$. This is done by minimizing the KL distance between the two:

$$\mathcal{D}[q_0 \| p_0] = \sum_{\boldsymbol{x}} q_0(\boldsymbol{x}) \ln \frac{p_0(\boldsymbol{x})}{q_0(\boldsymbol{x})}. \tag{4}$$

The substitution of Eq.(2) into Eq.(4) yields

$$\mathcal{D}[q_0 \| p_0] = \langle E_0(\boldsymbol{x}) \rangle_{q_0} - \mathcal{H}[q_0] + \ln Z_0, \tag{5}$$

where $\langle E_0(\boldsymbol{x}) \rangle_{q_0} = \sum_{\boldsymbol{x}} E_0(\boldsymbol{x}) q_0(\boldsymbol{x})$ is the expectation of the energy $E_0(\boldsymbol{x})$ with respect to $q_0(\boldsymbol{x})$, and $\mathcal{H}[q_0] = -\sum_{\boldsymbol{x}} q_0(\boldsymbol{x}) \ln q_0(\boldsymbol{x})$ is the entropy of $q_0(\boldsymbol{x})$. As the third term of Eq.(5) is independent of $q_0(\boldsymbol{x})$, the minimization of Eq.(5) is equivalent to that of the following *free energy*:

$$\mathcal{F}[q_0] = \langle E_0(\boldsymbol{x}) \rangle_{q_0} - \mathcal{H}[q_0]. \tag{6}$$

Many algorithms including MF, BP, and TRW are derived by minimizing this free energy for some selected class of q_0. For example, the generalized BP algorithm is derived when q_0 is chosen as

$$q_0(\boldsymbol{x}) = \frac{\prod_{c \in \mathcal{C}} q_c(\boldsymbol{x}_c)}{\prod_i q_i(x_i)^{\mathcal{N}(i)-1}}, \tag{7}$$

where $\mathcal{N}(i)$ is the number of clusters that include the i-th site.

3.3. MRF transformation

We now present our method for transforming MRFs. It is often the case that depending on the structure of MRFs, the algorithms of MF, BP etc. are impossible to derive, or the derived ones are computationally costly. To cope with such difficulties, we consider transforming the MRF and its associated objective function $\mathcal{F}[q_0]$ into another one, for which the resulting minimization is easier to perform.

Toward this end, introducing a new variable $\boldsymbol{z}_1$, we consider an approximate distribution $q_0(\boldsymbol{x})$ defined in the form of

$$q_0(\boldsymbol{x}) = \sum_{\boldsymbol{z}_1} q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1) q_1(\boldsymbol{z}_1), \tag{8}$$

where $q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)$ is a conditional distribution that we arbitrarily choose for our purpose and $q_1(\boldsymbol{z}_1)$ is a unknown distribution that we are to determine. By using Eq.(8) we wish to transform the optimization of $q_0(\boldsymbol{x})$ into that of $q_1(\boldsymbol{z}_1)$ that will be easier to perform. For example, it is often effective to use $\boldsymbol{z}_1$ having a lower-dimensionality than $\boldsymbol{x}$, or to use discrete $\boldsymbol{z}_1$ when $\boldsymbol{x}$ is continuous. An obvious issue is how to choose $q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)$. We choose it differently for different purposes, which will be described in the subsequent sections.

Using Eq.(8), the free energy of q_0 given in Eq.(6) is rewritten as follows:

$$\mathcal{F}[q_0] = \langle E_1(\boldsymbol{z}_1) \rangle_{q_1} - \mathcal{H}[q_1] + \langle S_1(\boldsymbol{x}) \rangle_{q_0(\boldsymbol{x})}, \tag{9}$$

where $E_1(\boldsymbol{z}_1)$ and $S_1(\boldsymbol{x})$ are defined as follows:

$$E_1(\boldsymbol{z}_1) = \sum_{\boldsymbol{x}} q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1) \left\{ E_0(\boldsymbol{x}) + \ln q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1) \right\}, \tag{10}$$

$$S_1(\boldsymbol{x}) = -\sum_{\boldsymbol{z}} q_{0,1}(\boldsymbol{z}_1|\boldsymbol{x}) \ln q_{0,1}(\boldsymbol{z}_1|\boldsymbol{x}). \tag{11}$$

In the above, we used $q_{0,1}(\boldsymbol{z}_1|\boldsymbol{x}) = q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1) q_1(\boldsymbol{z}_1) / q_0(\boldsymbol{x})$. The right hand side of Eq.(9) has a similar form to a free energy (defined as in Eq.(6) for q_0) except for the third term. To be specific, if we neglect the third term, we may think of Eq.(9) as the free energy of $q_1(\boldsymbol{z}_1)$ for the MRF whose energy is given by Eq.(10).

The third term of Eq.(9) does vanish when a condition is met as follows.

Lemma 3.1. (Erasure of S_1) *Let $\delta(\boldsymbol{z}_1)$ be the delta function. It holds that $S_1(\boldsymbol{x}) = 0$ if there exists a unique mapping function $\zeta_1 : \mathcal{X} \mapsto \mathcal{Z}_1$ that satisfies*

$$q_{0,1}(\boldsymbol{z}_1|\boldsymbol{x}) = \delta(\zeta_1(\boldsymbol{x}) - \boldsymbol{z}_1), \tag{12}$$

for any $\boldsymbol{x} \in \mathcal{X}$ and for any distribution $q_1(\boldsymbol{z}_1)$.

Thus, under the condition of this lemma, we can regard Eq.(9) as the free energy of the MRF model with a new energy $E_1(\boldsymbol{z}_1)$. As this energy includes the original energy $E_0(\boldsymbol{x})$ as well as additional terms as in Eq.(10), we call this the *augmented energy*. The results are summarized as follows:

Theorem 3.2. (MRF transformation) *Suppose a MRF specified by the distribution $p_0(\boldsymbol{x})$. When its approximation $q_0(\boldsymbol{x})$ is specified by Eq.(8) with $q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)$ satisfying the*

condition of Lemma 3.1, the variational solution to the marginal inference problem with this MRF (which searches for $q_0(\boldsymbol{x})$ that minimizes $D[q_0\|p_0]$) reduces to that with the MRF specified by $p_1(\boldsymbol{z}_1)$ defined as

$$p_1(\boldsymbol{z}_1) = \frac{1}{Z_1}\exp(-E_1(\boldsymbol{z}_1)), \tag{13}$$

where $E_1(\boldsymbol{z}_1)$ is the augmented energy defined by Eq.(10).

When the marginal inference problem with a MRF is intractable or computationally costly (even with the variational approach), we may transform the MRF into another one using the above method. As the transformed MRF is a regular MRF, many existing algorithms including MF, BP, and TRW can be used for its marginal inference. The outline of the proposed method is summarized as follows.

1. Choose $q_{0,1}(\boldsymbol{x}|\boldsymbol{z})$ that implements the target transformation of the MRF.
2. Compute the augmented energy $E_1(\boldsymbol{z}_1)$ as in Eq.(10).
3. Compute the marginal distributions for the transformed MRF (having $E_1(\boldsymbol{z}_1)$ as the energy) by using a selected algorithm (e.g., BP, TRW, etc.).

The marginal distributions of $q_0(\boldsymbol{x})$ may sometimes be necessary. In that case, they are to be computed from those of $q_1(\boldsymbol{z}_1)$. Although there is no automatic method, it will be easy to do so in some cases, as will be shown in the next section.

4. Applications

This section shows how the above method for MRF transformation can be applied to real problems. We consider three problems, the discretization of variable space, the grouping of discrete labels, and the coarse graining of MRFs.

4.1. Discretization of variable space

As described earlier, the discrete formulation of MRFs has a wider applicability than the continuous formulation. Thus, it is a common approach to discretize the variable space of a continuous problem and then apply some algorithm designed for discrete variables. However, as was pointed out in [18], if the discretization is non-uniform, the regular algorithms that do not consider the non-uniformity could yield inaccurate results. The method presented in the last section can derive algorithms that better handle such non-uniformity.

To do so, the method transforms the target MRF in the following way. Suppose an MRF having N sites with continuous variables $\boldsymbol{x} = [x_1, \ldots, x_N]$. We define $\boldsymbol{z}_1 = [z_1, \ldots, z_N]$, where z_i is the discrete variable of the i-th site that takes one of S_i discrete values, i.e., $z_i \in \mathcal{Z}_i \equiv \{1, \cdots, S_i\}$. We then choose $q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)$ of Eq.(8) as

$$q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1) = \prod_{i=1}^{N} q(x_i|z_i), \tag{14}$$

where $q(x_i|z_i)$ is a rectangular density such that the position of the rectangle varies depending on z_i. To be specific, when z_i takes a discrete value $s \in \mathcal{Z}_i$, it is given as

$$q(x_i|z_i = s) \equiv h_i^s(x_i), \tag{15}$$

where $h_i^s(x_i)$ is defined to be

$$h_i^s(x_i) = \begin{cases} 1/\mathcal{V}_i^s & \text{if } x_i \in \mathcal{X}_i^s \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

where $\mathcal{X}_i^s$ is the support of $h_i^s(x_i)$ in $\mathcal{X}$ and $\mathcal{V}_i^s$ is its volume; see Fig.1. By choosing $\mathcal{X}_i^s$ appropriately, the requirement of the proposed method is met.

Proposition 4.1. *If $\mathcal{X}_i^s \cap \mathcal{X}_i^t = \emptyset$ for any $s \neq t$, then $q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)$ of Eq.(14) satisfies the condition of Lemma 3.1.*

The augmented energy $E_1(\boldsymbol{z}_1)$ is calculated in a straightforward manner. Let $\mathcal{X}(\boldsymbol{z}_c) = \bigotimes_{i\in c} \mathcal{X}_i^{z_i}$ and $\mathcal{V}(\boldsymbol{z}_c)$ be the volume of $\mathcal{X}(\boldsymbol{z}_c)$. (Recall c is a factor of the graph.) From Eqs.(10), (14), and (15), $E_1(\boldsymbol{z}_1)$ is calculated as follows:

$$E_1(\boldsymbol{z}_1) = \sum_{c\in\mathcal{C}} g_c(\boldsymbol{z}_c) - \sum_{i=1}^{N} \ln \mathcal{V}_i^{z_i}, \tag{17}$$

where $g_c(\boldsymbol{z}_c)$ is given by

$$g_c(\boldsymbol{z}_c) = \frac{1}{\mathcal{V}(\boldsymbol{z}_c)} \sum_{\boldsymbol{x}_c \in \mathcal{X}(\boldsymbol{z}_c)} f_c(\boldsymbol{x}_c). \tag{18}$$

Note that the first term in the augmented energy is the regular energy of discrete MRFs. The second term is the additional term that accounts for the non-uniform discretization. In fact, when the discretization is uniform, $\mathcal{X}_i^{z_i}$'s will have the same shape and thus $\mathcal{V}_i^{z_i}$'s will be constant for different z_i's. Then we may neglect the term $-\ln \mathcal{V}_i^{z_i}$, resulting in the regular energy. If the discretization is non-uniform, we need to consider the second term. We can use any discrete algorithm for the marginal inference of the transformed MRF. We have only to replace the regular energy with the augmented energy derived as above.

4.2. Grouping of discrete labels

A similar method to the above one for dividing continuous variable space $\mathcal{X}_i$ into a discrete set of $\mathcal{X}_i^s$'s can be used to dividing discrete variable space, by which we can reduce the number of labels. To be specific, we divide the discrete variable space $\mathcal{X}_i$ into several subsets $\mathcal{X}_i^s \subset \mathcal{X}_i$ such

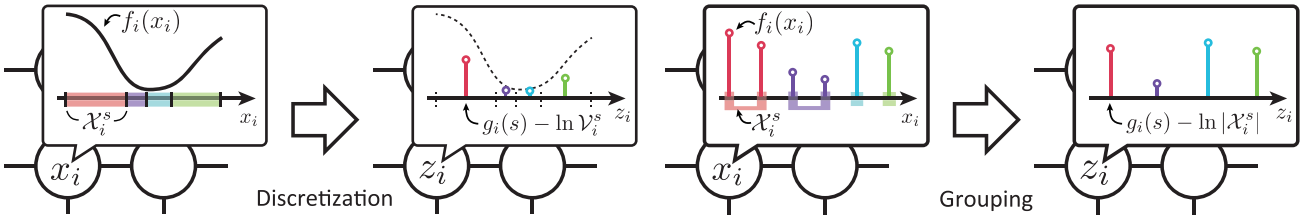


Figure 1. Left: Discretization of variable space. Right: Grouping of discrete labels. $f_i(x_i)$ is the unary term in the site i. $\mathcal{X}_i^s$ is the support of a label and is a set of labels to be grouped into a label.

that $\mathcal{X}_i^s \bigcap \mathcal{X}_i^t = \emptyset$; see Fig.1. This grouping of the labels is represented by making a few modifications to the above continuous-discrete transformation. We replace $h_i^s(x_i)$ of Eq.(16) with

$$h_i^s(x_i) = \begin{cases} 1/|\mathcal{X}_i^s| & \text{if } x_i \in \mathcal{X}_i^s \\ 0 & \text{otherwise,} \end{cases} \tag{19}$$

where $|\mathcal{X}_i^{z_i}|$ is the number of elements in $\mathcal{X}_i^{z_i}$. Then the augmented energy will be

$$E_1(\boldsymbol{z}_1) = \sum_{c \in \mathcal{C}} g_c(\boldsymbol{z}_c) - \sum_{i=1}^{N} \ln |\mathcal{X}_i^{z_i}|, \tag{20}$$

where $g_c(\boldsymbol{z}_c)$ is equivalent to the one in Eq.(18) except that $\mathcal{V}(\boldsymbol{z}_c)$ is replaced with $|\mathcal{X}(\boldsymbol{z}_c)|$.

As with the above continuous-discrete transformation, the additional term $-\ln |\mathcal{X}_i^{z_i}|$ compensates for the non-uniformity of the grouping of labels. Its effect will be large when each group $\mathcal{X}_i^{z_i}$ contains a different number of labels.

4.3. Coarse graining of MRFs

The proposed method can also be applied to coarse graining of MRFs. After downsizing the graph of an MRF, it is then required to transform the energy $E_0(\boldsymbol{x})$ accordingly. Our method provides a systematic way for this transformation, which was missing in the literature.

Our method assumes that it is already determined how to modify the graph. Suppose that N sites of the graph are grouped into K blocks ($K < N$). Each block becomes a single site of the new graph. Let $\mathcal{C}(k)$ be the set of the sites grouped into the k-th block ($k = 1, \ldots, K$), such that $\mathcal{C}(k) \neq \emptyset$ for any k and also $\mathcal{C}(k) \cap \mathcal{C}(k') = \emptyset$ for any $k \neq k'$. We then consider a new variable z_k for each block k, which shares the same variable space as x_i; thus, if x_i is discrete, so is z_i.

We choose $q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)$ of Eq.(8) as

$$q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1) = \prod_{k=1}^{M} q(\boldsymbol{x}_k|z_k), \tag{21}$$

where $\boldsymbol{x}_k$ indicates a vector containing all the site variables of the k-th block, and further choose $q(\boldsymbol{x}_k|z_k)$ as

$$q(\boldsymbol{x}_k|z_k) = \prod_{i \in \mathcal{C}(k)} \delta(x_i - z_k), \tag{22}$$

where $\delta(x)$ is Dirac's delta function if the site x_i is continuous and is Kronecker's delta function if x_i is discrete. Although there are other possibilities, the above choice of $q_{0,1}(\boldsymbol{x}|\boldsymbol{z})$ is natural, as it enforces that the sites of the original MRF belonging to each group will have the same value as the corresponding site of the coarse grained MRF. It also satisfies the requirement of the proposed method.

Proposition 4.2. *The conditional distribution $q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)$ defined by Eqs.(21) and (22) satisfies the condition of Lemma 3.1.*

The augmented energy $E_1(\boldsymbol{z}_1)$ can be calculated as above, but unlike earlier MRF transformations, the results will vary depending on the structure of MRFs. For lack of space, we show here only the derivation for second-order MRFs. The energy $E_0(\boldsymbol{x})$ of a second-order MRF is given as

$$E_0(\boldsymbol{x}) = \sum_i f_i(x_i) + \sum_{(i,j) \in \mathcal{E}} f(x_i, x_j), \tag{23}$$

where $f_i(x_i)$ and $f_{ij}(x_i, x_j)$ are the unary and the pairwise terms, respectively; $\mathcal{E}$ is the set of edges in $\mathcal{G}$. Using Eq.(10) and Eqs. (21) - (23), $E_1(\boldsymbol{z}_1)$ is calculated as

$$E_1(\boldsymbol{z}_1) = \sum_k \left(\sum_{i \in \mathcal{C}(k)} f_i(z_k) + \sum_{(i,j) \in \mathrm{In}(k)} f_{ij}(z_k, z_k) \right) + \sum_{(k,l) \in \mathcal{E}_{\mathrm{Ex}}} \sum_{(i,j) \in \mathrm{Ex}(k,l)} f_{ij}(z_k, z_l), \tag{24}$$

where $\mathrm{In}(k)$ indicates the set of the edges contained in the k-th block (i.e., the edges between any pair of the sites in the k-th block); $\mathcal{E}_{\mathrm{Ex}}$ is the set of pairs of any neighboring blocks; $\mathrm{Ex}(k,l)$ indicates the set of the edges crossing the boundary between the neighboring (k-th and l-th) blocks.

For notational simplicity, we rewrite Eq.(24) as

$$E_1(\boldsymbol{z}_1) = \sum_k g_k(z_k) + \sum_{(k,l) \in \mathcal{E}_{\mathrm{Ex}}} g_{kl}(z_k, z_l), \tag{25}$$

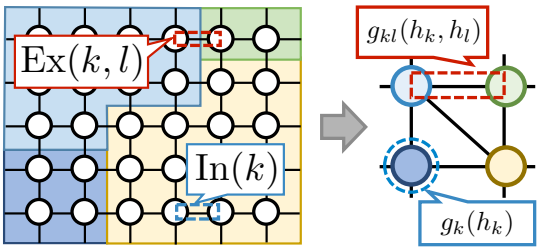


Figure 2. Illustration of coarse graining of an MRF graph and how the interactions between the sites in the original graph are transformed to unary and pairwise terms of the coarse-grained MRF.

where

$$g_k(z_k) = \sum_{i\in\mathcal{C}(k)} f_i(z_k) + \sum_{(i,j)\in\mathrm{In}(k)} f_{ij}(z_k,z_k), \quad (26a)$$

$$g_{kl}(z_k,z_l) = \sum_{(i,j)\in\mathrm{Ex}(k,l)} f_{ij}(z_k,z_l). \quad (26b)$$

The second term of (26a) expresses the interaction occurring within each block, and constitutes the unary term of the augmented energy. The term $g_{kl}(z_k,z_l)$ of (26b) expresses the interaction between the blocks and serves as the pairwise term. These are illustrated in Fig.2. As in the earlier MRF transformations, we may use any algorithm for the transformed, coarse grained MRF. One can use the derived augmented energy as if it is a regular energy of a regular MRF.

Although it is omitted here, higher-order MRFs can be treated in a similar way, and the results are similar, too. For any energy term having only the site variables contained in a single block, it reduces to the form of (26a). For any term having site variables split to different blocks, it will reduce to the form of (26b).

5. Other applications

5.1. Coarse-to-fine inference

We have shown how the proposed method is used to transform MRFs for different purposes. Although it is not explicitly mentioned so, the discussion so far mostly considers MRF transformations in the direction of downsizing them. This is the case with the grouping of discrete labels and the MRF coarse graining. However, the proposed method can be used to "upsizing" MRFs, i.e., transforming MRFs into those having more sites or more labels. This is useful when we employ the coarse-to-fine strategy for the inference with large-size MRFs.

Such coarse-to-fine inference can be implemented as follows. For a given MRF, we first transform it into a smaller one by one (or a combination) of the above techniques and perform the marginal inference with the transformed MRF. We then consider another transformation of the original MRF that has an intermediate size between the first and the original MRFs. Let the approximate distributions for the first and second MRFs be $q(\boldsymbol{x}) = \sum_{\boldsymbol{z}_1} q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)q_1(\boldsymbol{z}_1)$ and $q'(\boldsymbol{x}) = \sum_{\boldsymbol{z}_2} q_{0,2}(\boldsymbol{x}|\boldsymbol{z}_2)q_2(\boldsymbol{z}_2)$, respectively. By appropriately designing the second transformation such that the space of $q'(\boldsymbol{x})$ include that of $q(\boldsymbol{x})$, there always exists $q_2(\boldsymbol{z}_2)$ such that $q(\boldsymbol{x}) = q'(\boldsymbol{x})$. Therefore, we can transfer the result obtained with the first MRF (i.e., $q_1(\boldsymbol{z}_1)$) to the second MRF, which gives an estimate of $q_2(\boldsymbol{z}_2)$. Using this as an initial value, we perform the marginal inference with the second MRF, which is expected to yield more accurate estimate of $p_0(\boldsymbol{x})$ due to the increased degrees of freedom. We may iterate this process until we reach the original MRF.

As good initial values are given at each step, the inference in this coarse-to-fine manner is expected to reduce the total computational cost as compared with performing the marginal inference with the original MRF once. The proposed method provides a smooth connection between two MRFs in consecutive steps. Thus, it is also possible to employ the coarse graining and the label grouping at the same time at each step.

5.2. Markov chain Monte Carlo (MCMC)

As mentioned above, the proposed method can be used with any algorithm derived from the variational principle, such as MF, BP, TRW etc. The method can also be used with MCMC-based algorithms such as Gibbs Sampling and Slice Sampling. It is similarly expected to reduce computational cost by downsizing MRFs.

MCMC-based methods estimate marginal distributions by generating a lot of samples from the target distribution $p_0(\boldsymbol{x})$. From the viewpoint of the variational principle, it is equivalent to defining the approximate density $q_0(\boldsymbol{x})$ as

$$q_0(\boldsymbol{x}) = \frac{1}{M}\sum_m \delta(\boldsymbol{x} - \boldsymbol{x}^m), \quad (27)$$

where M is the number of samples, $\boldsymbol{x}^m$ is the sample from the distribution $p_0(\boldsymbol{x})$, and δ is the delta function. It is easy to calculate (the estimate of) the marginal distribution of x_i from $q_0(\boldsymbol{x})$, which is merely the histogram of the generated samples, i.e., $(\sum_m \delta(x_i - x_i^m))/M$.

An advantage of using MCMC methods for marginal inference is that the estimates can be more accurate than those of MF, BP etc., provided that we can generate a large number of samples. However, this prohibitively increases computational cost in most cases, which is the reason why MF, BP etc. are preferred. The computational cost of MCMC methods depend on the size of the MRF, rigorously, the number of sites and either the dimensionality of the variable space in continuous cases or the number of labels in discrete cases. Therefore, it is attractive to downsize the MRF and reduce the computational cost by the proposed method.

To do so, we transform the target MRF with $p_0(\boldsymbol{x})$ into a smaller one with $p_1(\boldsymbol{z}_1)$ by one or a combination of the

individual methods described in Section 4. We then apply a regular MCMC method to the transformed MRF, generating samples $z_1^1, \ldots, z_1^M$ from $p_1(\boldsymbol{z}_1)$. (Note that this is expected to be computationally less costly than sampling $p_0(\boldsymbol{x})$.) The approximate distribution of $p_0(\boldsymbol{x})$ is given from these samples as

$$q_0'(\boldsymbol{x}) = \frac{1}{M}\sum_m \sum_{\boldsymbol{z}_1} q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1)\delta(\boldsymbol{z}_1 - \boldsymbol{z}_1^m) = \frac{1}{M}\sum_m q_{0,1}(\boldsymbol{x}|\boldsymbol{z}_1^m). \quad (28)$$

It differs from the original $q_0(\boldsymbol{x})$ of Eq.(27) in that it consists of a set of *distributions* $q_{0,1}(\boldsymbol{x}|\boldsymbol{z}^m)$ not of *samples* x^m. It is nevertheless still easy to calculate the marginal densities of $\boldsymbol{x}$ using $q_0'(\boldsymbol{x})$.

A caveat is that unlike $q_0(\boldsymbol{x})$, $q_0'(\boldsymbol{x})$ will never coincide with the true density $p_0(\boldsymbol{x})$ even if we generate an infinite number of samples from $p_1(\boldsymbol{z}_1)$. Our experiments show that this might not be a serious issue in reality, although this is not a rigorous proof. Even when it is really a problem, the above approach will still be useful when used with the coarse-to-fine strategy, in which starting with a small-size MRF, we gradually increase the MRF size until reaching the original MRF. In that case, Eq.(28) gives a smooth connection in the transition from an MRF to another.

6. Experimental results

6.1. Discretization of variable space

If the variable space of a MRF is discretized in a uniform manner and nevertheless an ordinary algorithms is naively used for it, the results will be inaccurate. This was first pointed out in [18], in which only MF and BP are considered. The proposed method can handle any algorithm derived from the variational principle as well as methods of MCMC, yielding their extensions that can properly deal with non-uniform discretization. To demonstrate these, we show here the results for TRW and Gibbs sampling. We used OpenGM [1] for their implementation.

For the sake of comparison, we use the same experimental setting as [18]. That is, we consider a simple Gaussian MRF of a 5×5 grid graph with pairwise 4-neighbor connections:

$$E(\boldsymbol{x}) = \sum_i x_i^2 + \sum_{(i,j)\in\mathcal{E}} (x_i - x_j)^2. \quad (29)$$

For this MRF, we divide the variable space in an asymmetric way that the negative and positive parts in the range $[-2:2]$ are discretized by 64 and 16 points, respectively, as shown in Fig.3. We then applied the ordinary and extended versions of TRW and Gibbs sampling to this MRF. For Gibbs sampling, we generated 10^7 samples, from which we calculate marginal distributions either naively (by Eq.(17) without the second term) or by our method. We set the burn-in period to 1000 steps.

Figure 3 shows the results. They are the estimates of the marginal distribution of the site at the upper-left corner of the 5×5 graph. The white dots are the results of the naive TRW and Gibbs sampling, whereas the blue histograms are those of their extended counterparts that are obtained by the proposed method. Note that the former are purely discrete distributions and we adjusted the vertical scale properly for comparison. The red curves are the exact distributions. (As it is a Gaussian MRF, its marginal distributions can be computed analytically in the continuous domain.) It is observed that while the distributions estimated by the naive methods have some biases, those of the extended methods do not. Although they are less significant, the variances are more accurate for the extended methods, too.

6.2. Downsizing CRFs

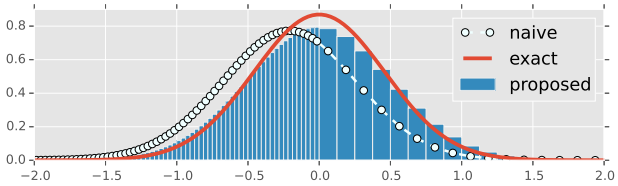
We next examine how the proposed method works for downsizing a discrete CRF. As an example problem, we chose a CRF-based formulation of semantic labeling. To be specific, we consider its learning step for determining CRF parameters, to which we applied coarse graining and label grouping. Owing to its theoretical foundation, the proposed method is expected to minimize inaccuracy caused by the downsizing. Therefore we evaluated computational efficiency as well as estimation accuracy. We used the MSRC-21 dataset [20] for the experiments. It consists of images of 320×213 pixels, each of which is given one of 21 discrete object labels. We used the "accurate ground truth" introduced by [13] for the evaluation of results.

We consider a grid CRF whose energy is given by

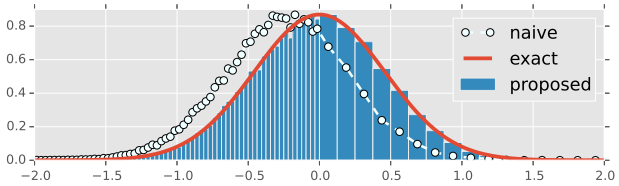
$$E(\boldsymbol{x}|\mathcal{I};\theta) = \sum_i f_i(x_i|\mathcal{I}) + \sum_{(i,j)\in\mathcal{E}}\sum_{s,t}\theta_{st}\delta(x_i - s)\delta(x_j - t), \quad (30)$$

where $\mathcal{I}$ is the input image; x_i is the variable of the i-th site taking one of the 21 labels; $f_i(x_i|\mathcal{I})$ is the unary term; and θ_{st} is the parameter representing the interaction between the label s and t. In the learning step, θ_{st} is determined from the training data consisting of the pairs of an image and its true label. This is performed by maximizing the likelihood calculated from the (estimates of) marginal distributions at the sites. Their estimation requires to use BP or similar methods, which is the bottleneck in the entire process of learning. This can be resolved or mitigated by downsizing the MRF.

We used the following two methods for the downsizing. The first is grouping the discrete labels, where we reduced



TRW



Gibbs Sampling

Figure 3. Results for non-uniformly discretized variable space. The marginal distribution of the site at the upper-left corner of a 5×5 grid is estimated by the naive and extended versions of TRW and Gibbs sampling. See text for details.

Table 1. Quantitative results on the MSRC-21 dataset.

	time [h]	speedup	disparity	accuracy
full MRF	9.5	-	0.0	81.6
2 labels	0.89	10.6×	0.01235	77.8
3 labels	0.95	10.0×	0.00526	80.7
4 labels	1.0	9.2×	0.00496	81.3
5 labels	1.1	9.0×	0.00473	81.3
4×4 grid	0.66	14.4×	0.215	81.5
3×3 grid	1.1	8.5×	0.236	81.6
2×2 grid	2.5	3.9×	0.237	81.7

the number of labels to K for each pixel. (We fixed it throughout the learning.) To be specific, we selected $K-1$ labels having the smallest values of the unary terms and grouped the other labels into one label. Note that the selection was performed independently at each pixel and thus the resulting grouping may be different for different pixels. The second is coarse graining of the MRF, where we downsized the original grid MRF by grouping the pixels in $b \times b$ square blocks into a single "superpixel." Note that in spite of the downsizing, we do estimate the marginal distributions of the *original* MRF. They are used to calculate the likelihood, which is to be minimized.

We divided the MSRC-21 dataset into 276, 59, and 256 images for training, validation, and test, which is the same as [13, 20]. We used BP [16] with the damping factor 0.5 and 50 iteration counts for estimating marginal distributions for each MRF. We multiplied the unary term of [13] by $1/10$ to stabilize the computation. We employ the stochastic gradient descent (SGD) method for maximizing the likelihood to determine θ_{st}'s. We set the learning rate to 1.5×10^{-5}, the batch size to 8, and the number of epochs to 5. In the testing step, we used the α-expansion [2] to obtain the MAP estimates for the MRF, which was used to measure the accuracy of the learned parameters. We used OpenGM [1] for the implementation on a PC with Intel Core i7-2600 having eight CPU cores clocked at 3.40GHz.

Table 1 shows quantitative results. The "disparity" column shows the mean differences of the parameter θ_{st} between the full MRF and its downsized versions. The "accuracy" column shows the percentage of correctly labeled pixels. The rows of "L labels" show the results of different label grouping, and those of "$b \times b$ grid" show the results of differently coarse-grained MRFs. It is observed that both methods for downsizing achieve significant speed ups at a small expense of inaccuracy. An exception is two-label grouping, which shows considerably lower accuracy. This indicates that the reduction from 21 to only two labels is excessive. An interesting remark is that the label grouping yields much smaller disparity than the coarse graining, and nevertheless their labeling accuracy are almost the same or the latter is even slightly better. An implication of this is that label grouping is more "accurate" in the sense that it is more close to the results of full MRFs. However, there is no guarantee that full MRFs are better at learning better parameters. The coarse grained MRFs could avoid local maxima.

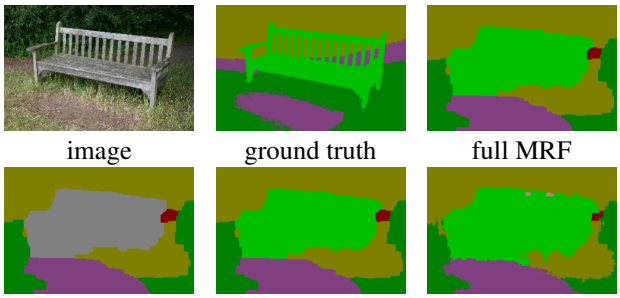


Figure 4. Qualitative results on the MRRC-21 dataset.

Figure 4 shows a qualitative comparison of the results. It is observed that the two-label grouping yields very inaccurate labeling and the four-label grouping and the 4×4 coarse graining both yield similar results to the original MRF. We have checked that this is the case with the other images.

7. Summary

We have described a novel generic method for transforming MRFs and its applications to several practical problems. We have also described that these MRF transformations are combined to perform coarse-to-fine inference, and can be used with MCMC methods. The experimental results demonstrate the effectiveness of the proposed method.

Acknowledgement This work was supported by CREST, JST and JSPS KAKENHI Grant Number 25135701.

References

[1] B. Andres, T. Beier, and J. H. Kappes. OpenGM: A C++ Library for Discrete Graphical Models. *CoRR*, abs/1206.0, 2012. 7, 8

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *PAMI*, 23(11), 2001. 8

[3] B. Conejo, N. Komodakis, S. Leprince, and J. P. Avouac. Speeding-up Graphical Model Optimization via a Coarse-to-fine Cascade of Pruning Classifiers. In *NIPS*, 2014. 2

[4] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *CVPR*, 2011. 1

[5] S. M. A. Eslami, N. Heess, and J. Winn. The Shape Boltzmann Machine : a Strong Model of Object Shape. In *CVPR*, 2012. 1

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Belief Propagation for Early Vision. *IJCV*, 70(1):41–54, 2006. 1, 2

[7] M. A. R. Ferreira and H. K. H. Lee. *Multiscale Modeling: A Bayesian Perspective*. Springer, 2007. 2

[8] T. Kim, S. Nowozin, P. Kohli, and C. D. Yoo. Variable Grouping for Energy Minimization. In *CVPR*, 2011. 2

[9] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust Higher Order Potentials for Enforcing Label Consistency. *IJCV*, 82(3), 2009. 1, 2

[10] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009. 1

[11] V. Kolmogorov. Convergent Tree-Reweighted Message Passing for Energy Minimization. *PAMI*, 28(10):1568–1583, 2006. 1

[12] N. Komodakis and G. Tziritas. Approximate Labeling via Graph Cuts Based on Linear Programming. *PAMI*, 29(8):1436–1453, 2007. 1

[13] P. Krähenbühl and V. Koltun. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *NIPS*. 2011. 1, 7, 8

[14] C. Lei and Y.-H. Yang. Optical Flow Estimation on Coarse-to-Fine Region-Trees using Discrete Optimization. In *ICCV*, 2009. 2

[15] S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer Publishing Company, Incorporated, 3rd edition, 2009. 1

[16] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy Belief Propagation for Approximate Inference: An Empirical Study. In *Uncertainty in Artificial Intelligence*, pages 467–475, 1999. 8

[17] S. Roth and M. J. Black. Fields of Experts. *IJCV*, 82(2):205–229, 2009. 1, 2

[18] M. Saito, T. Okatani, and K. Deguchi. Discrete MRF Inference of Marginal Densities for Non-uniformly Discretized Variable Space. In *CVPR*, 2013. 2, 4, 7

[19] R. Salakhutdinov and G. Hinton. Deep Boltzmann Machines. In *AISTATS*, 2009. 1

[20] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *IJCV*, 81(1), 2009. 7, 8

[21] C. Sutton and A. McCallum. *An Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2007. 1

[22] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2011. 1

[23] O. Veksler. Reducing Search Space for Stereo Correspondence with Graph Cuts. In *BMVC*, 2006. 2

[24] L. Wang, H. Jin, and R. Yang. Search Space Reduction for MRF Stereo. In *ECCV*, 2008. 2

[25] L. Wang, G. Zhao, L. Cheng, and M. Pietikainen. *Machine Learning for Vision-Based Motion Analysis*. Springer, 2011. 1

[26] L. Xu, J. Jia, and Y. Matsushita. Motion Detail Preserving Optical Flow Estimation. In *CVPR*, 2010. 1

[27] Q. Yang, L. Wang, and N. Ahuja. A constant-space belief propagation algorithm for stereo matching. In *CVPR*, 2010. 2