On Learning Higher-Order Consistency Potentials for Multi-class Pixel Labeling

Kyoungup Park[1,2] and Stephen Gould[1]

[1] College of Engineering and Computer Science, Australian National University
[2] NICTA, Australia
{kyoungup.park,stephen.gould}@anu.edu.au

Abstract. Pairwise Markov random fields are an effective framework for solving many pixel labeling problems in computer vision. However, their performance is limited by their inability to capture higher-order correlations. Recently proposed higher-order models are showing superior performance to their pairwise counterparts. In this paper, we derive two variants of the higher-order lower linear envelop model and show how to perform tractable move-making inference in these models. We propose a novel use of this model for encoding consistency constraints over large sets of pixels. Importantly these pixel sets do not need to be contiguous. However, the consistency model has a large number of parameters to be tuned for good performance. We exploit the structured SVM paradigm to learn optimal parameters and show some practical techniques to overcome huge computation requirements. We evaluate our model on the problems of image denoising and semantic segmentation.

1 Introduction

Many challenging problems in computer vision, such as semantic segmentation [1,2], geometric interpretation [3] and image denoising [4], can be formulated in terms of pixel labeling. Here the goal is to assign a label to each pixel in an image from some predefined label set. Conditional random fields (CRFs) are a powerful framework for solving these problems. Usually, the CRFs encode labeling preferences via unary terms conditioned on local pixel features and a pairwise smoothness prior over adjacent pixels. However, this encoding scheme fails to model the complex structure of objects in images.

Shotton et al. [2] attempted to enforce consistency globally by learning an image-specific appearance model for each object class and encoded as hidden variables in the CRF. This approach has the benefit of encoding consistency between disconnected regions. However, the approach is necessarily iterative since appearance models need to be estimated from an initial prediction of the pixel labeling, and errors in these predictions can negatively affect the quality of the results. Moreover, the introduction of hidden variables into the model complicates parameter learning.

Many authors have demonstrated the improvement over pairwise CRFs by incorporating higher-order constraints, which encode complex relationships over

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 202–215, 2012.
© Springer-Verlag Berlin Heidelberg 2012

the image. One important class of higher-order constraint enforces consistency over contiguous regions in the image. The consistency model, for example, starts from P^n Potts model [5], which favors the same label to be assigned to all pixels within the region. However, the P^n Potts model can be quite brittle especially on poorly estimated regions. For instance, if all pixels in a clique do not take the same label, the same penalty is incurred regardless of the number of inconsistent pixels. To overcome this problem, Kohli et al. [6] proposed the Robust P^n Potts model and reported impressive segmentation results due to better modeling of superpixel regions. Their model defines a penalty proportional to the number of inconsistent labels up to some maximum penalty. This model has been generalized to an arbitrary number of linear functions that define lower and upper envelops [7]. However, model performance is still sensitive to the superpixel definition of regions. One possible remedy is provided by Ladicky et al. [8], who combined multiple contiguous regions hierarchically and included a generalization of the Robust P^n model.

With the significant improvements promised by higher-order terms, the issue of learning the model parameters efficiently and effectively has become an important research question. As the number of parameters increases by the introduction of the higher-order potential, cross-validation–typically used for pairwise models–is not effective. Szummer et al. [9] show the max-margin framework is efficient solution for learning parameters in pairwise CRF models, especially using graph cuts. Gould [10] proposed an alternative energy minimization construction for the case of binary consistency potentials and resolved the learning issue with a modified max-margin framework. On top of the standard max-margin framework, they included additional linear constraints by adding a second-order curvature constraint to ensure that the higher-order potentials remain concave functions. Komodakis [11] showed an alternative interpretation of the max-margin learning using the dual-decomposition method.

In this paper, we explore variants of the higher-order potential that encodes a preference for consistency over large (and possibly disjoint) regions. Building on the work of Gould [10], we generalize the binary case to the multi-class case. Our model defines a concave penalty function over the number of pixels within a predefined region (or clique) that is annotated with a given label. Importantly, we do not restrict the pixel sets to contiguous local regions. In our experiments, we show that the non-local regions are beneficial in encoding global labeling constraints. We derive α-expansion and $\alpha\beta$-swap moves that can be generalized to multi-class lower linear envelop functions. Our paper also provides efficient and effective learning methods for the large number of parameters. We show how the max-margin learning method is affected by our approximate inference to find the optimal parameters, and propose a number of heuristics to reduce the heavy computation required by the max-margin method.

Our contributions in this paper include: First, we derive two different extensions (i.e., min or *sum*) of the binary lower linear envelop function to the multi-class case with approximate move-making inference. The derivation generalizes the higher-order consistency terms with the lower linear envelop function.

Second, we explore the max-margin learning for our extended higher-order terms and reduce the training time significantly with approximate solutions for large-scale training examples. Third, we apply our model to the task of multi-class pixel labeling with non-local consistency constraints. Our approach is evaluated on the 21-class MSRC image segmentation data set.

2 Background

We begin by providing a brief background to submodular energy functions and move-making inference for multi-class conditional random fields (CRFs).[1]

Submodular Energy Functions. Consider a binary pairwise conditional Markov random field (CRF) over variables $\boldsymbol{y} = (y_1, \ldots, y_n) \in \{0,1\}^n$ and observed features $\boldsymbol{x}$. Let $\mathcal{V} = \{1, \ldots, n\}$ denote the set of *nodes* and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ denotes the set of *edges*. Then we can write the energy function for the CRF as

$$E(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i \in \mathcal{V}} \psi_i^U(y_i; \boldsymbol{x}) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}^P(y_i, y_j; \boldsymbol{x}) \tag{1}$$

where $\boldsymbol{\theta}$ are the model parameters. The terms $\psi_i^U(y_i; \boldsymbol{x})$ are known as *unary potentials* and capture the labeling preference for a single variable in the random field. The terms $\psi_{ij}^P(y_i, y_j; \boldsymbol{x})$ are known as *pairwise potentials* and define a preference over two variables. The pairwise terms are typically defined over a sparse subset $\mathcal{E}$ of all possible variables pairs (i.e. adjacent pixels in the image). If two nodes $(i,j) \in \mathcal{E}$, then the node i and the node j are said to be *neighbors.* In pixel labeling problems, it is usual to use the pairwise term to smooth via a contrast sensitive term of the form

$$\psi_{ij}(y_i, y_j; \boldsymbol{x}) = \lambda [\![y_i \neq y_j]\!] \exp\left\{ -\frac{1}{2\beta} \left\| x_i - x_j \right\|^2 \right\} \tag{2}$$

where $[\![\cdot]\!]$ is the indicator function which takes 1 when the argument is true and 0 otherwise, x_i is the color vector for pixel i, and λ and β are global and image specific constants that determine the strength of the smoothness prior.

An equivalent representation for the pairwise binary CRF is as a quadratic pseudo-Boolean function (QPBF) [12]. Here we write the energy function $E : \{0,1\}^n \rightarrow \mathbf{R}$ in *posiform* as

$$\begin{aligned} E(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\theta}) = \theta_{const}(\boldsymbol{x}) &+ \sum_{i \in \mathcal{V}} \theta_{i;0}(\boldsymbol{x}) \overline{y}_i + \theta_{i;1}(\boldsymbol{x}) y_i \\ &+ \sum_{(i,j) \in \mathcal{E}} \theta_{ij;00}(\boldsymbol{x}) \overline{y}_i \overline{y}_j + \theta_{ij;01}(\boldsymbol{x}) \overline{y}_i y_j + \theta_{ij;10}(\boldsymbol{x}) y_i \overline{y}_j + \theta_{ij;11}(\boldsymbol{x}) y_i y_j \end{aligned} \tag{3}$$

[1] We will use the terms MRF and CRF interchangeably from now on.

where $\overline{y}_i = 1 - y_i$ and all coefficients $\theta_{a;b}(\boldsymbol{x})$ are non-negative with the possible exception of the constant term θ_{const}.[2] The coefficients can be constant or some functions of the observed features $\boldsymbol{x}$. With respect to the pixel labeling problem, we have $\theta_{ij;00} = \theta_{ij;11} = 0$ and $\theta_{ij;01} = \theta_{ij;10} = \lambda \exp\left\{-\frac{1}{2\beta}\left\|x_i - x_j\right\|^2\right\}$.

A pseudo-Boolean function is called submodular if and only if $f(\boldsymbol{u}) + f(\boldsymbol{v}) \geq f(\boldsymbol{u} \vee \boldsymbol{v}) + f(\boldsymbol{u} \wedge \boldsymbol{v})$ for all binary vectors $\boldsymbol{u}, \boldsymbol{v} \in \{0,1\}^n$. An equivalent condition for the binary pairwise CRFs (i.e., quadratic pseudo-Boolean function) can be written in posiform notation with $\theta_{ij;00} = \theta_{ij;11} = 0$ for all variable pairs $(i,j) \in \mathcal{E}$. Note that the contrast sensitive smoothness prior (Eq. (2)) is submodular. In the computer vision literature submodularity is sometimes referred to as regularity [13].

The goal of inference is to find the assignment $\hat{\boldsymbol{y}}$ with minimum energy. Message-passing algorithms can be suitable for the objective, but it is well known that for submodular pairwise energy functions this can be done efficiently by finding the minimum-cut in a suitably constructed graph [14,15,16]. Unfortunately, in general for multi-label CRFs (or indeed, non-submodular binary CRFs), inference is intractable and we need to resort to approximate routines.

Move-Making Inference. Generally, most energy minimization problems are NP-hard [13]. However, there are good approximate solutions available. For example, some move-making algorithms reduce the energy minimization to sequence of smaller problems which are submodular. Here, each move restricts the label space of variables to at most two values from the label set. The algorithm starts from an initial labeling. Then, it searches move spaces to minimize the energy relative to the previous assignment. If no improvement is found after searching over the restricted label space, the solution is considered to converge to a local minimum. The algorithm can be formalized as follows: Consider the current assignment $\boldsymbol{y}^{prev} \in \mathcal{L}^n$, where $\mathcal{L}$ is the set of possible labels to each variable. If the energy $E(\boldsymbol{y}_t)$ from $\boldsymbol{y}_t \in \{\boldsymbol{y}^{prev}\} \cup \mathcal{Y}^t$ is less than $E(\boldsymbol{y}^{prev})$, the assignment $\boldsymbol{y}_t$ is updated as $\boldsymbol{y}^{next}$ at the iteration t, where $\mathcal{Y}^t$ is a possible movement at the iteration t. Otherwise, $\boldsymbol{y}^{prev}$ is kept as $\boldsymbol{y}^{next}$.

An early example of move-making algorithms is Iterated Conditional Modes (ICM) [17]. For a variable, it finds the optimal solution conditioned on all other variables. However, the update of a single variable makes its convergence slow and can easily get stuck in poor local optima. The more advanced examples of move-making algorithms are α-expansion and $\alpha\beta$-swap [15]. In α-expansion, a label from $\mathcal{L}$ is chosen iteratively. Each variable can switch to the chosen label α or keep the current label, which expands the current label α to other regions as long as the energy is reduced. It continues to iterate through the label set until the energy reduces no more. Similarly, during $\alpha\beta$-swap, two labels from $\mathcal{L}$ are iteratively selected. Then it makes moves by swapping only between variables with either of the two labels and retains all other variable assignments. In summary, the three algorithms are characterized as follows:

[2] Note that this representation is not unique.

- **ICM**: for a given variable $i \in \mathcal{V}$, we choose the $y_i^{next} \in \mathcal{L}$ that minimizes the energy with $y_i^{next} = y_j^{prev}$ for all $j \neq i$.
- α**-expansion**: for all i, we choose the $y_i^{next} \in \{y_i^{prev}, \alpha\}$ that jointly minimize the energy.
- $\alpha\beta$**-swap**: for all i such that $y_i^{prev} \in \{\alpha, \beta\}$, we choose the $y_i^{next} \in \{\alpha, \beta\}$ that jointly minimize the energy and $y_j^{next} = y_j^{prev}$ for all other variables.

For α-expansion and $\alpha\beta$-swap, an efficient graph-cut based algorithm has been proposed to minimize the energy functions composed of pairwise potential functions [16,13].

3 Higher-Order Consistency Potentials

The usual form of the energy function for a CRF model is composed of the sum of unary and pairwise potentials. The unary potential ψ_i is represented as the negative log of the likelihood of a label assigned to a node i while the pairwise potential ψ_{ij} encodes the interaction between the neighborhood nodes. The Potts model is an example of the pairwise potential to impose image smoothness. In spite of its improvement in removing noise, the pairwise terms can over smooth resulting in poor object boundaries. Due to the limitations of the pairwise potential model, the development of more sophisticated models with higher-order terms have been developed (see Eq. (4)). These higher degree potentials ψ_c^H are capable of capturing more powerful statistics over images.

$$E(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i \in \mathcal{V}} \psi_i^U(y_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}^P(y_i, y_j) + \sum_{c \in \mathcal{C}} \psi_c^H(\boldsymbol{y}_c) \tag{4}$$

In this section, we describe two variants of lower linear envelop functions introduced by [7]. These can encode arbitrary concave functions over the number of variables taking a given assignment. Gould [10] showed that the potentials can be represented by pairwise submodular energy functions for the case of binary MRFs. We extend the binary lower linear envelope functions to multi-class inference problem using α-expansion and $\alpha\beta$-swap move algorithms.

Generalized Multi-class Representation. Here we extend and generalize the binary lower linear envelope potential functions described in Gould [10] to a multi-class problem. Consider an arbitrary set of multi-class variables $\boldsymbol{y}_c = \{y_i \mid i \in c\}$ where $c \subseteq \{1, \ldots, n\}$. We define our multi-class consistency potential as

$$\psi_c^H(\boldsymbol{y}_c) = \bigoplus_{l \in \mathcal{L}} \min_{k=1,\ldots,K_l} \{a_k^l \sum_{i \in c} \omega_i [\![y_i = l]\!] + b_k^l\} \tag{5}$$

where (a_k^l, b_k^l) are parameter pairs of each k-th linear envelop function for label l. We assume that the parameters are sorted by a_k^l in decreasing order for each label, which means that $a_k^l > a_{k+1}^l$ and $b_k^l < b_{k+1}^l$ for each label l. Thus, a set of K_l linear functions composes piecewise linear envelop functions to assign a penalty for each label l being inconsistent over the subset $\boldsymbol{y}_c$. The aggregation $\bigoplus$ can represent $\{\min \text{ or } \sum\}$ to measure various penalties over the cliques. We consider three cases for the same objective:

1. $\psi_c^H(\boldsymbol{y}_c) = \min_l \min_k \{b_k^l + a_k^l \sum_{i \in c} [\![y_i = l]\!]\}$
2. $\psi_c^H(\boldsymbol{y}_c) = \sum_l \min_k \{b_k^l + a_k^l \sum_{i \in c} [\![y_i = l]\!]\}$
3. $\psi_c^H(\boldsymbol{y}_c) = \min_{k'} \{b_{k'} + a_{k'} \sum_{i \in c} [\![y_i = l_{k'}]\!]\}$

Case 1 finds the least potential function among all label set with K_l functions and is identical to Case 3 just with different parameterizations over all sorted functions. Case 2 is the extended form of the binary case [10], which minimizes the sum of the penalties over each lower linear envelop function. Note that the per-variable weight ω_i is set as one for convenience.

$\boldsymbol{\alpha}$-Expansion Move. From the generalized higher-order potential function, we derive the α-expansion move for approximate minimization of the linear envelop potential functions. Let $\boldsymbol{y}^{prev} \in \boldsymbol{\mathcal{L}}^n$ be the current best assignment of labels and $S_l = \{i \mid y_i^{prev} = l\}$ be the subset whose variables are assigned to label l. For α-expansion, we constrain the moves to $y_i^{next} \in \{y_i^{prev}, \alpha\}$. To encode the expansion moves, the binary transfer vector $\boldsymbol{t}$ is defined as

$$y_i^{next} = \begin{cases} y_i^{prev} \ (\neq \alpha) & \text{if } t_i = 0 \\ \alpha & \text{if } t_i = 1 \ . \end{cases} \tag{6}$$

Then, we can rewrite the restricted potential function with the new variables;

$$\begin{aligned} \psi_c^\alpha(\boldsymbol{t}) &= \bigoplus_{l \in \mathcal{L}} \min_k \left\{ a_k^l \sum_{i \in S_l} \{t_i [\![l = \alpha]\!] + \bar{t}_i [\![l \neq \alpha]\!]\} + b_k^l \right\} \\ &= \left(\bigoplus_{l \neq \alpha} \min_k \{a_k^l \sum_{i \in S_l} \bar{t}_i + b_k^l\} \right) \oplus \left(\min_k \{a_k^\alpha (\sum_{i \notin S_\alpha} t_i + N_\alpha) + b_k^\alpha\} \right) \end{aligned} \tag{7}$$

where $N_l = |S_l|$. Replacing $\bar{t}_i = 1 - t_i$ and substituting each coefficient and subset with

$$\tilde{a}_k^l = \begin{cases} -a_k^l & \text{for } l \neq \alpha \\ a_k^\alpha & \text{for } l = \alpha \end{cases}, \quad \tilde{b}_k^l = b_k^l + a_k^l N_l, \text{ and } \tilde{S}_l = \begin{cases} S_l & \text{for } l \neq \alpha \\ \overline{S}_\alpha & \text{for } l = \alpha \end{cases}, \tag{8}$$

we have the general form of a lower linear envelop function over binary variables

$$\psi_c^\alpha(\boldsymbol{t}) = \bigoplus_{l \in \mathcal{L}} \min_k \{\tilde{a}_k^l \sum_{i \in \tilde{S}_l} t_i + \tilde{b}_k^l\} \ . \tag{9}$$

$\boldsymbol{\alpha\beta}$-Swap Move. Similarly defining the transfer vector $\boldsymbol{t}$ as

$$y_i^{next} = \begin{cases} \alpha & \text{if } t_i = 0 \\ \beta & \text{if } t_i = 1 \end{cases} \quad \forall i \in S_\alpha \cup S_\beta \ , \tag{10}$$

we arrive at

$$\psi_c^{\alpha\beta}(\boldsymbol{t}) = \left(\bigoplus_{l=\alpha,\beta} \min_k \{\tilde{a}_k^l \sum_{i \in S_\alpha \cup S_\beta} t_i + \tilde{b}_k^l\} \right) \bigoplus_{l \neq \alpha,\beta} C_l \tag{11}$$

where $C_l = \min_k\{a_k^l N_l + b_k^l\}$ is a constant to account for all variables excluded from the move. Again, this is a lower linear envelop function over binary variables.

Applying Auxiliary Variables. In addition to the transfer vector $\boldsymbol{t}$ above, we need to transform the higher-order potential functions to the form of quadratic pseudo-Boolean function for energy minimization to be tractable. For simplicity, we assume that each class has K linear functions and there are $|\mathcal{L}|$ different classes. However, our method extends to the general case of $K_l \neq K_{l'}$. In [10], an alternative way to minimize over the piecewise linear functions is introduced using binary auxiliary variables. The minimization over K linear functions can be encoded by introducing $K-1$ binary auxiliary variables. Extended to the multi-class case, the required binary variables z_k^l are $|\mathcal{L}|(K-1)$ for the set of all labels. Here we take an example of replacing the aggregation with '$\sum$'.[3] Then, we can derive the QPBF representation for α-expansion as

$$\hat{\psi}_c^\alpha(\boldsymbol{t}, \boldsymbol{z}) = \min_{\boldsymbol{z}} \tilde{E}_c^\alpha(\boldsymbol{t}, \boldsymbol{z}^\alpha) + \sum_{l \neq \alpha} \tilde{E}_c^l(\boldsymbol{t}, \boldsymbol{z}^l) \tag{12}$$

where

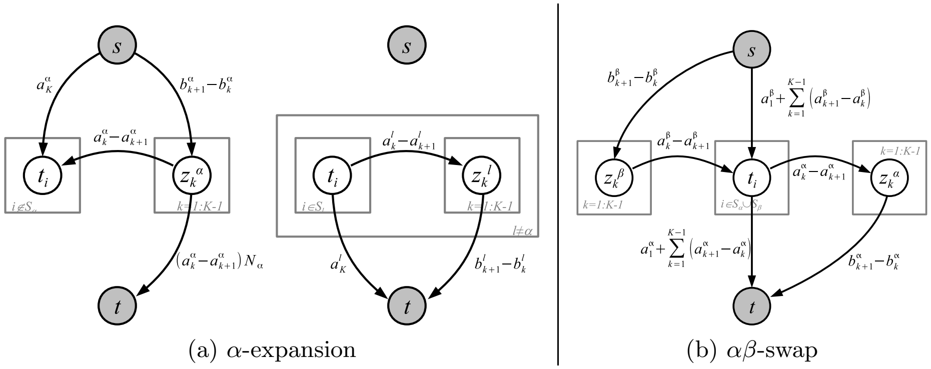
$$\begin{aligned} \tilde{E}_c^\alpha(\boldsymbol{t}) = a_K^\alpha \sum_{i \notin S_\alpha} t_i + a_K^\alpha N_\alpha + b_1^\alpha + \sum_{k=1}^{K-1} \overline{z}_k^\alpha (a_k^\alpha - a_{k+1}^\alpha) \sum_{i \notin S_\alpha} t_i \\ + \sum_{k=1}^{K-1} \overline{z}_k^\alpha (a_k^\alpha - a_{k+1}^\alpha) N_\alpha + \sum_{k=1}^{K-1} z_k^\alpha (b_{k+1}^\alpha - b_k^\alpha) \end{aligned} \tag{13}$$

and

$$\tilde{E}_c^l(\boldsymbol{t}) = b_1^l + a_K^l \sum_{i \in S_l} \overline{t}_i + \sum_{k=1}^{K-1} z_k^l (a_k^l - a_{k+1}^l) \sum_{i \in S_l} \overline{t}_i + \sum_{k=1}^{K-1} \overline{z}_k^l (b_{k+1}^l - b_k^l) \ . \tag{14}$$

Note that contrasting the description in [10], the explicit constraints to enforce $z_k^l \geq z_{k+1}^l$ is not required any more because the constraints are implicitly sufficient to get the minimum energy. Now, we have the restricted submodular potential functions, which can be minimized in time polynomial in the number of variables using the graph cuts algorithm. Figure 1 illustrate the st-graph constructions for both the move-making algorithms.

[3] Unlike the $\alpha\beta$-swap move, the α-expansion should be derived for '$\sum$' because of the submodularity condition.


s
a_K^α
$b_{k+1}^\alpha - b_k^\alpha$
$a_k^\alpha - a_{k+1}^\alpha$
t_i
z_k^α
$i \notin S_\alpha$
k=1:K-1
$(a_k^\alpha - a_{k+1}^\alpha) N_\alpha$
t
s
$a_k^l - a_{k+1}^l$
t_i
z_k^l
$i \in S_l$
k=1:K-1
$l \neq \alpha$
a_K^l
$b_{k+1}^l - b_k^l$
t
s
$b_{k+1}^\beta - b_k^\beta$
$a_1^\beta + \sum_{k=1}^{K-1} (a_{k+1}^\beta - a_k^\beta)$
$a_k^\beta - a_{k+1}^\beta$
z_k^β
t_i
$a_k^\alpha - a_{k+1}^\alpha$
z_k^α
k=1:K-1
$i \in S_\alpha \cup S_\beta$
k=1:K-1
$a_1^\alpha + \sum_{k=1}^{K-1} (a_{k+1}^\alpha - a_k^\alpha)$
$b_{k+1}^\alpha - b_k^\alpha$
t

(a) α-expansion

(b) $\alpha\beta$-swap

Fig. 1. *st*-graphs for α-expansion and $\alpha\beta$-swap moves. The rectangles indicate replication of nodes.

4 Learning Parameters with Structured SVM

In this section, we describe how to learn the parameters including our lower linear envelop potentials. Our multi-class model needs to learn $2 + (K + 1)|\mathcal{L}|$ parameters for unary, pairwise and higher-order terms. The cross-validation approach usually employed for pairwise learning is not a feasible method for this large number of parameters. To learn parameters efficiently, we adopt a variant of the max-margin framework by [18,19].

Let an energy function $E(\boldsymbol{y}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{y})$ be parameterized as linear combination of features $\phi(\boldsymbol{y}) \in \mathbb{R}^m$ and weights $\boldsymbol{\theta} \in \mathbb{R}^m$ where m is the number of the parameters. The framework learns weights for the energy function given a training set $\mathcal{Y} = \{\boldsymbol{y}_t\}_{t=1}^T$ and the objective function with constraints is

$$\underset{\boldsymbol{\theta},\, \boldsymbol{\xi} \succeq 0}{\text{minimize}} \quad \frac{1}{2} \|\boldsymbol{\theta}\|^2 + \frac{C}{T} \sum_{t=1}^{T} \xi_t \tag{15}$$

$$\text{subject to } E(\boldsymbol{y}; \boldsymbol{\theta}) - E(\boldsymbol{y}_t; \boldsymbol{\theta}) \geq \Delta(\boldsymbol{y}, \boldsymbol{y}_t) - \xi_t \qquad \boldsymbol{y} \in \mathcal{Y}_t,\ \forall t$$

$$\mathbf{G}\boldsymbol{\theta} \geq \mathbf{0} \tag{16}$$

where $\mathcal{Y}_t \subseteq \mathcal{L}^n$ is the set of all possible assignments for t-th training example and C is a regularization constant. The loss Δ compensates the large margin for each high loss example. To solve the above quadratic program, the constrains should be satisfied by all the possible assignments $\boldsymbol{y}$. Due to the large number of the possible assignments, this optimization problem can be solved by cutting-plane method by finding the most violated ones first; other potential constraints are then guaranteed to have larger margin than the subset of constraints. The Hamming loss Δ is suitable to be decomposed to the unary potential and we can get the most violated constraints by graph-cuts algorithm ($\boldsymbol{y}^* = \text{argmin}_{\boldsymbol{y}} E(\boldsymbol{y}; \boldsymbol{\theta}) - \Delta(\boldsymbol{y}, \boldsymbol{y}_t)$).

In order to learn the parameters by the max-margin framework, our higher-order term requires re-parameterization. Let $\psi_c^H(\boldsymbol{y}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \phi(\boldsymbol{y})$. The feature vector $\phi(\boldsymbol{y}) = \{\phi^l \mid l \in \mathcal{L}, \phi^l \in \mathbb{R}^{K+1}\}$ represents the consistency of the clique by the number of pixels for each label l (i.e. $\phi_k^l(\boldsymbol{y}) = \sum_c \sum_{i \in c} [\![y_i = l]\!]$). And the coefficients are converted to the parameters to learn such as $a_k^l = \theta_k^l - \theta_{k-1}^l$ and $b_k^l = \theta_k^l - k a_k^l$ for $k = 1, \ldots, K+1$. The additional constraint (16) enforces the parameters $\boldsymbol{\theta}$ of our higher-order potentials to be the concave function (refer to [10] for further details).[4]

Max-margin learning is a powerful framework but computationally expensive on large data sets. Specifically, it demands comprehensive search for violated constraints from all training examples at every iteration. Intuitively, we can borrow some idea from stochastic gradient descent to speed convergence. Instead of testing the whole training set, taking a subset of the examples at each iteration results in decrease of the computational complexity. For example, if an example had no violated constraints found before, the example will be discarded for inference of most violated constraint. After skipping all examples, we seek violated constraints for all examples again. This method still guarantees to converge to optimal parameters because it finally satisfies all constraints for the training examples. Another heuristic speed-up is to reduce the number of iterations in move-making inference. During training, we know the ground-truth labels of the training examples, therefore we expect that the minimum energy assignment tends to belong to the ground-truth labels. By reducing the label set to only ground-truth labels in the move-making algorithm, the loss-augmented inference time as a main part of the training time, can be saved at the risk of missing the most violated constraints. Choosing the optimal move space is a topic of active research area (see [20]).

5 Experiments

Denoising with Synthetic Data. This synthetic experiment is a toy example that verifies the overall performance of the algorithm on a denoising task. The input data is an artificially generated checkerboard image. Here we generate 8×8 checkerboards and each square contains 16×16 variables assigned with one of five labels consistently. The unary potential is $\psi_i^U(y_i) = \theta^{unary} x_i(y_i)$ and generated with noise as $x_i(y_i) = \mathcal{U}[1,2] - 0.3\mathcal{U}[0,1][\![y_i = y_i^*]\!]$ where y_i^* is a ground truth label for the pixel i and $\mathcal{U}[0,1]$ is the uniform distribution. The pairwise potential is $\psi_{ij}^P(y_i, y_j) = \theta^{pair} [\![y_i \neq y_j]\!]$ and defined between every pair of adjacent variables. Each checkerboard coincides with a consistency clique for the input of the higher-order term. We set the number of the linear envelop functions per label to $K = 5$.

Figure 2 compares the denoised images inferred with all the parameters learned by the max-margin learning. As expected, the pairwise model does not recover the noisy image perfectly (Fig. 2b and 2f). However, we can see the images 2c and 2d, where parameters have been learned by α-expansion, are coincident with the ground truth. When $\alpha\beta$-swap is used, the performance is slightly worse (see 2g and 2h). With the higher-order terms involved, the value of the pairwise

[4] To ensure the energy function remains submodular, θ^{pair} must be non-negative.

parameter became negligible,[5] which indicates that the higher-order term is more dominant than the pairwise term because the higher-order consistency term is the generalized form of the pairwise term in the experiment. Another interesting point is that move-making algorithms are approximate solutions to the global minimum energy [15] and the approximate inference for generating constraints in learning can lead the max-margin framework to perform poorly [21], which explains that the learning with $\alpha\beta$-swap move left some noise in the image.

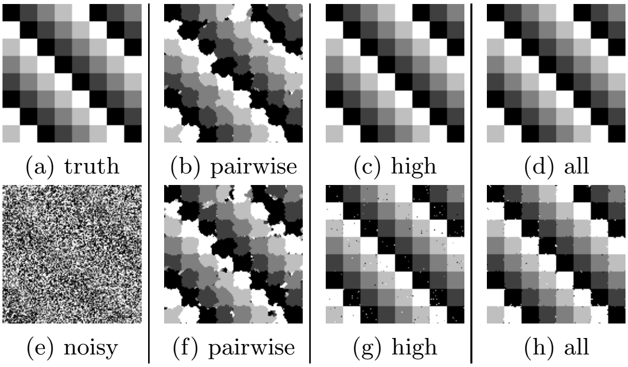

(a) truth (b) pairwise (c) high (d) all
(e) noisy (f) pairwise (g) high (h) all

Fig. 2. Results from our synthetic data. 2a is the ground truth with 5 different labels and 2e is the noisy image. 2b - 2d are results by α-expansion. 2f - 2h are results by $\alpha\beta$-swap move.

Multi-class Segmentation. We evaluate the performance of our model conducting semantic segmentation on the MSRC data set (23 classes)[6]. While there exist very powerful models for achieving state-of-the-art results on this data set (e.g., Ladicky et al. [8] use sophisticated features and a hierarchical CRF model), our interest is in evaluating the learning algorithm for higher-order consistency potentials. We, therefore, choose a simpler baseline model consisting of a pairwise smoothness prior and unary potentials learned from local color and texture features. As is standard with this data set, we removed the two scarcest classes ('horse' and 'mountain') by filling with the 'void' label. Based on the previously published works on this data set, we divided the image set into 315 training and 276 test images. This separation repeated with five random shuffles.

The baseline (unary, pairwise) models follow the standard pixelwise MRF models.[7] The unary potentials are encoded by boosted decision trees via multiclass logistic regression classifier. The features are derived from 17 filters over images. The pairwise potentials are contrast-dependent smoothness terms for 8 neighbors defined as $\psi^P_{ij}(y_i, y_j) = [\![y_i \neq y_j]\!] \exp\left\{-\frac{1}{2\beta}\|x_i - x_j\|^2\right\}$ where β is the average squared distance between adjacent color vectors and x_i and x_j are RGB

[5] The ratio is about 42 times ($\theta^{unary} = 7.33 \times 10^{-4}$ and $\theta^{pair} = 1.76 \times 10^{-5}$).

[6] `http://www.cs.cmu.edu/~tmalisie/projects/bmvc07/`

[7] `http://drwn.anu.edu.au`

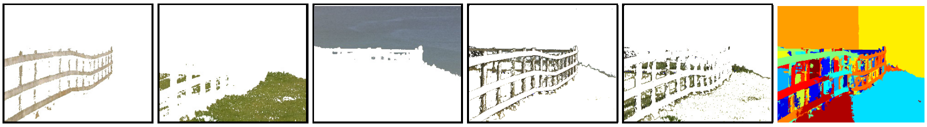


Fig. 3. Separate pixel sets for non-local regions and contiguous regions. The five pixel sets by GMM are illustrated followed by mean-shift clustering.

color vectors for pixel i and j. Learning for the baseline model was conducted by cross-validation given the limited number of pairwise parameters.

Regarding the higher-order consistency model, we used Gaussian Mixture Model (GMM) clustering for higher-order cliques: five groups of clusters over pixel colors.[8] As shown in Figure 3, the superpixels vary in size and shape unlike the checkerboards. Different to contiguous clustering, the GMM clustering defines non-local regions globally sharing common features (i.e. colors or textures) over similar objects in images, which enables overlapped or disjointed object parts to belong to each original object (see Fig. 3). We set the number of linear functions per class as $K = 3$ and the regularization constant as $C = 1$.

Table 1. Averaged results with standard deviation for five experiments on 21-class MSRC data set ((B)='Baseline', (V)='Validation', (R)='Reduced Set')

Model	Evaluation Accuracy		Time (d:h:m)
	Overall	Class	
Unary(B)	58.44±1.04	42.40±1.08	26m
Unary+Pair(B)	73.23±2.39	58.69±4.21	2h53m
Unary+Pair	71.65±1.20	59.18±1.42	1h19m
Unary+Pair(R)	71.19±1.25	58.64±1.50	9m
Unary+Pair(V)	76.09±1.21	64.66±1.55	1h5m
Unary+Pair(RV)	74.11±1.39	62.38±1.75	43m
Unary+High	73.22±1.22	61.23±1.27	8h8m
Unary+High(V)	73.06±1.17	60.86±0.75	7h51m
Full	**79.47**±1.38	69.55±0.94	2d15h21m
Full(R)	75.51±1.59	66.48±1.92	11h46m
Full(V)	**79.68**±1.42	69.48±1.17	2d6h28m
Full(RV)	76.02±1.74	67.48±1.95	1d9h46m

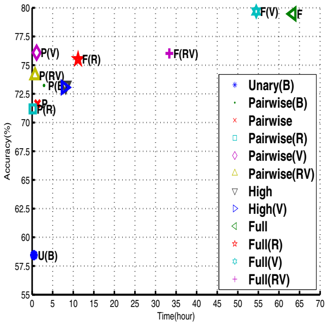


Table 1 shows overall results from various models.[9] The full models performed best in accuracy (79.7%). Unlike the previous experiment with the synthetic data, the experiments with consistency terms alone do not always outperform the result from the pairwise experiment. We recall that simple clustering over

[8] Other clustering methods such as mean-shift clustering or mixture of clustered regions can be considered.

[9] The experiment was performed on a workstation with an Intel Xeon E5520 CPU (2.27GHz) and 32GB RAM (10 cores were used).

pixel colors decomposed objects apart into some classes (i.e. the human has different colors of limbs and clothing) and sometimes, a clique combines more than one object classes together. In case of the pairwise parameter learning, the validation process was effective on gaining the optimal parameters while the learning without validation missed the best parameters. For max-margin learning, validation may be required due to: a) loss in constraints may not represent sufficiently large margin for minimum energy, b) the move-making inference is approximate solution, and c) some amount of 'void' labeling can not be estimated properly in learning. Naturally, the training and the inference times increase as the number of parameters and variables increase. Our heuristic methods such as skipping non-violated examples and cutting down label set in inference reduced the training time significantly. However, cutting down the label set sometimes results in degraded accuracy as shown in Table 1.

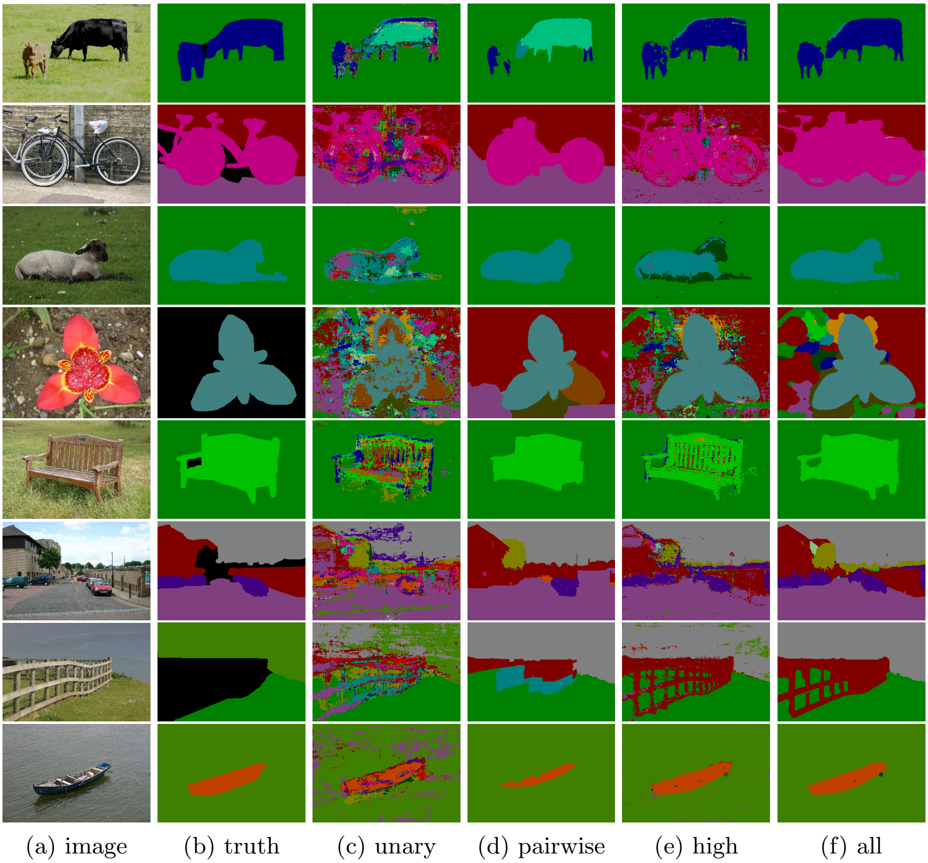


Fig. 4. Improved examples with the higher-order consistency terms. The black region represents 'void'. Best viewed in color.

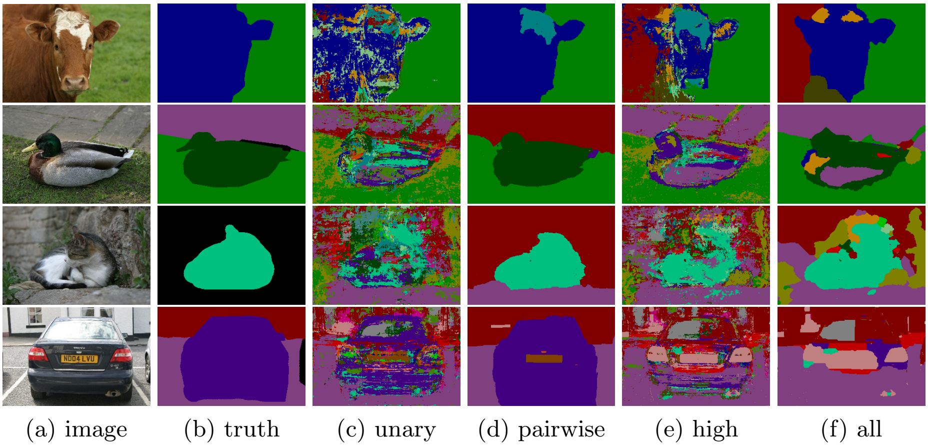


Fig. 5. Degraded examples including the higher-order consistency terms

Figure 4 and Figure 5 show some examples of where the inclusion of the consistency potential improved and degraded the results respectively. We can see that the higher-order terms provide clear contours and accurate object segmentation comparing to the pairwise term model. Note that our model with the definition of non-local regions, for example, segmented the fence correctly with holes in it, which differentiates from the use of contiguous local regions (see Fig. 3). However, the higher-order terms can also degrade performance. Due to the coarse color-based clustering, some objects have been labeled with a few different labels instead of a single one. We suspect including multiple overlapping or hierarchical clustering, this difficulty would be alleviated and performance improved.

6 Conclusion

In this paper we addressed the problem of parameter learning for multi-class lower linear envelop energy functions. We first showed how to perform approximate inference via move-making and thus how to employ max-margin parameter learning. Our results demonstrated that the higher-order terms were very successful in MRF inference tasks such as image denoising and semantic segmentation. However, the higher-order terms impose heavy computation cost on inference and learning due to the large number of parameters and variables. In our max-margin learning procedure, we proposed adaptive methods based on subset optimization for reducing the training time.

It remains an open question how to best define the regions for the higher-order terms. We utilized simple color based non-local regions for the higher-order cliques and saw the improved results. But, we believe that there is still room for improvement, for example, through more sophisticated clique discovery and learning higher-order terms conditioned on refined image features.

Acknowledgments. This research is supported by the Australian Research Council. Kyoungup Park is funded by National ICT Australia (NICTA).

References

1. He, X., Zemel, R., Carreira-Perpinán, M.: Multiscale conditional random fields for image labeling. In: CVPR (2004)
2. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: *TextonBoost*: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
3. Hoiem, D., Efros, A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision 75, 151–172 (2007)
4. Roth, S., Black, M.: Fields of experts. International Journal of Computer Vision 82, 205–229 (2009)
5. Kohli, P., Kumar, M., Torr, P.: P3 & beyond: Solving energies with higher order cliques. In: CVPR (2007)
6. Kohli, P., Ladickỳ, L., Torr, P.: Robust higher order potentials for enforcing label consistency. International Journal of Computer Vision 82, 302–324 (2009)
7. Kohli, P., Kumar, M.: Energy minimization for linear envelope MRFs. In: CVPR (2010)
8. Ladicky, L., Russell, C., Kohli, P., Torr, P.: Associative hierarchical CRFs for object class image segmentation. In: ICCV (2009)
9. Szummer, M., Kohli, P., Hoiem, D.: Learning CRFs Using Graph Cuts. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 582–595. Springer, Heidelberg (2008)
10. Gould, S.: Max-margin learning for lower linear envelope potentials in binary markov random fields. In: ICML (2011)
11. Komodakis, N.: Efficient training for pairwise or higher order CRFs via dual decomposition. In: CVPR (2011)
12. Boros, E., Hammer, P.: Pseudo-boolean optimization. Discrete Applied Mathematics 123, 155–225 (2002)
13. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? PAMI (2004)
14. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for markov random fields with smoothness-based priors. PAMI (2008)
15. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI (2001)
16. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. PAMI (2004)
17. Besag, J.: On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society. Series B (Methodological), 259–302 (1986)
18. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: ICML (2004)
19. Taskar, B., Chatalbashev, V., Koller, D., Guestrin, C.: Learning structured prediction models: A large margin approach. In: ICML (2005)
20. Batra, D., Kohli, P.: Making the right moves: Guiding alpha-expansion using local primal-dual gaps. In: CVPR (2011)
21. Finley, T., Joachims, T.: Training structural SVMs when exact inference is intractable. In: ICML (2008)