Multiclass Pixel Labeling with Non-Local Matching Constraints

Stephen Gould
Research School of Computer Science
Australian National University
stephen.gould@anu.edu.au

Abstract

*A popular approach to pixel labeling problems, such as multiclass image segmentation, is to construct a pairwise conditional Markov random field (CRF) over image pixels where the pairwise term encodes a preference for smoothness within local 4-connected or 8-connected pixel neighborhoods. Recently, researchers have considered higher-order models that encode soft non-local constraints (*e.g.*, label consistency, connectedness, or co-occurrence statistics). These new models and the associated energy minimization algorithms have significantly pushed the state-of-the-art for pixel labeling problems.*

In this paper, we consider a new non-local constraint that penalizes inconsistent pixel labels between disjoint image regions having similar appearance. We encode this constraint as a truncated higher-order matching potential function between pairs of image regions in a conditional Markov random field model and show how to perform efficient approximate MAP inference in the model. We experimentally demonstrate quantitative and qualitative improvements over a strong baseline pairwise conditional Markov random field model on two challenging multiclass pixel labeling datasets.

1. Introduction

The task of labeling each pixel in an image for the purpose of semantic understanding is a key challenge in computer vision that has received increasing attention in recent years [11, 26, 18, 10]. Here the aim is to provide a semantic (or geometric) segmentation of the image where each pixel is assigned a label from a pre-defined set of classes, *e.g.*, *sky*, *road*, *tree*, *etc*. The most successful approaches use conditional Markov random fields (CRFs), which allow local appearance information (such as color and texture) to be combined with a smoothness prior that favors labelings in which neighboring pixels are assigned the same class label.

There has been a recent trend to improve results for pixel labeling problems by incorporating higher-order terms into the CRF models. These terms bias the energy-minimizing solution of the model towards one that has a more desirable label configuration. For example, in figure-ground segmentation a preference for global connectivity [27, 22] or segmentation "tightness" [20] may be encoded. These preferences need not be hard constraints and thus can be overridden given enough contrary evidence.

In the context of multiclass image segmentation, Ladicky *et al*. [18] proposed using higher-order terms, known as consistency potentials, to promote smoothness over large superpixel regions rather than relying on the simple pairwise smoothness terms encoded by traditional CRF models. While their approach encourages a uniform label assignment over large regions, it does not allow long-range similarity constraints to be encoded, *e.g.*, that a patch in one part of the image resembles a patch in another part of the image and therefore should be labeled consistently.

Consider the side view of a car. In that view, not only do the two visible wheels of the car have similar appearance, but they also suggest similar labeling pattern for their surroundings—*i.e*. car body above and road below. In this paper, we investigate a novel higher-order potential function for encoding this type of non-local symmetry information. Specifically, we incorporate, into a unified CRF model, terms that encourage consistent pixelwise labelings between pairs of image patches with similar appearance.

Our model is motivated by the idea that appearance-based symmetry within an image plays an important part in scene understanding and can be exploited to improve segmentation of the image (*e.g*. see [1, 29]). In other words, our model is based on the observation that *similar appearance of disjoint image regions suggests similar semantic meaning for pairs of corresponding pixels in the regions*.

Our contributions are two-fold: First, we introduce a new kind of higher-order term—the *truncated higher-order matching potential*—that captures long-range similarity between image regions as soft constraints in a CRF model for pixel labeling. Second, we show how to approximately minimize the resulting energy function efficiently using a graph-cut construction. Experimental results on two challenging datasets validate our approach.

2. Background and Related Work

Many recent works on multiclass pixel labeling build on the conditional Markov random field (CRF) models introduced by He *et al*. [11] and Shotton *et al*. [26] (although the basic idea of constructing a Markov random field over image pixels dates back even further, *e.g*., [2, 9]). In these works, each pixel in the image is associated with a random variable and the distribution over the joint assignment to all random variables is defined by both local features (encoded as unary potentials) and pairwise correlations between neighboring variables.

These pairwise CRF models perform remarkably well and the introduction of efficient inference algorithms, such as α-expansion and $\alpha\beta$-swap [6], for finding good approximate maximum a posteriori (MAP) solutions, has solidified CRFs as the dominant method for multiclass image segmentation and as the foundation for more sophisticated scene understanding tasks, *e.g*., [10, 28].

Despite their success, pairwise CRF models still leave much room for improvement, *e.g*., correctly labeling pixels near object boundaries, and a number of recent works address these problems through the introduction of higher-order terms and the associated energy minimization (or MAP inference) algorithms.

Some works, for example, enforce constraints on the number of distinct labels appearing in a solution [21, 8], label co-occurrence [19], or region connectivity [27, 22]. In contrast to our approach, these works aim to encode a preference for the structure of the solution without regard to pixel appearance. As such, the models are complementary to the higher-order terms developed in our work.

Kohli *et al*. [14] introduce the idea of higher-order potentials that enforce label consistency over superpixel regions. The higher-order potentials are instantiations of the so-called robust P^n model [13] that imposes a linear penalty on the number of pixels that disagree with the most frequently occurring label within the superpixel. The term is truncated by some maximum possible penalty. Ladicky *et al*. [18] extend this image segmentation model by allowing class label predictions at the superpixel level to influence pixel-level labels.

Our model for multiclass pixel labeling is strongly motivated by the these works, but unlike the robust P^n model, our model does not attempt to enforce label consistency over a region. Rather our model enforces a consistent label *pattern* between two separate regions.

At a conceptual level, our ideas relate to the work of Bagon *et al*. [1] on interactive figure-ground segmentation. In that work, the authors suggest that good segments are self-similar, *i.e*. can be composed from other chunks of the same segment. Our model can be thought of as an extension of this broad idea to the multiclass labeling case.

Perhaps most similar to our research are works that consider pattern templates. The work of Zhu *et al*. [30] introduce segmentation templates into a multiclass pixel labeling CRF model. A similar notion are the pattern potentials introduced by Rother *et al*. [24] and Komodakis and Paragios [16]. However, these templates and patterns are chosen *a priori* and attempt to constrain the label pattern within an image region rather than match the configuration of labels between image regions.

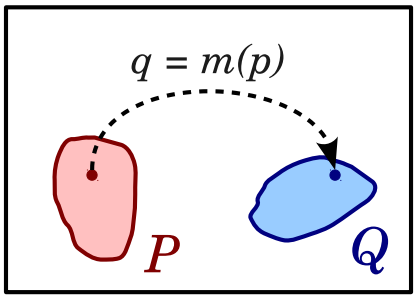

Figure 1. Schematic showing the mapping between two regions $\mathcal{P}$ and $\mathcal{Q}$. We wish to penalize labelings in which the two regions disagree on corresponding pixel assignments, y_p and y_q.

3. Higher-order Matching Potential

In this section we present our higher-order term for enforcing consistency between labels of corresponding pixels from two disjoint regions in the image. We also show how this term can be optimized in the context of move-making MAP inference for the CRF model. In Section 4 we will show how we incorporate this higher-order term into a standard CRF model for multiclass pixel labeling.

3.1. Problem Setup

Our framework for developing the higher-order matching potential is the problem of multiclass pixel labeling. In this problem every pixel in a $W \times H$ image is assigned a label from a discrete label set $\mathcal{L}$. The joint labeling over all pixels is denoted by $\boldsymbol{y} \in \mathcal{L}^{W \times H}$.

Typically, the label y_p for any given pixel p will depend on both local appearance features derived from the image and *a priori* knowledge encoded in the model. For example, most pixel labeling models incorporate a smoothness prior that encodes the fact that, in real images, neighboring pixels usually take the same label.

For our problem, we wish to encode the constraint that corresponding pixels between two matching regions in the image agree on their label. We will defer discussion of how these regions are discovered in a specific image until Section 4. For now, consider two equal-size sets of pixels $\mathcal{P}$ and $\mathcal{Q}$. Let $m : \mathcal{P} \rightarrow \mathcal{Q}$ be a one-to-one mapping (bijection) from the pixels $p \in \mathcal{P}$ to the pixels $q \in Q$. This is depicted in Figure 1. We define our *truncated higher-order*

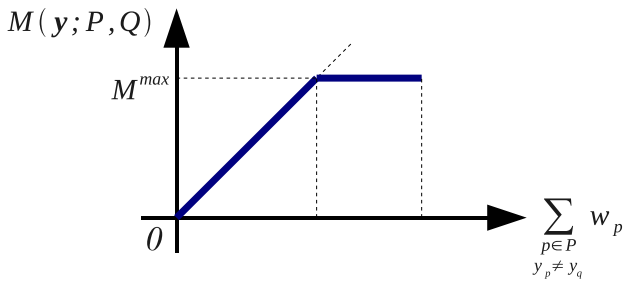

Figure 2. Truncated higher-order matching potential. The potential penalizes disagreement between labels, y_p and y_q, of corresponding pixels within matched regions, $\mathcal{P}$ and $\mathcal{Q}$, up to some maximum penalty, $M^{\max}$.

matching potential between $\mathcal{P}$ and $\mathcal{Q}$ as

$$M(\boldsymbol{y};\mathcal{P},\mathcal{Q}) \triangleq \min\left\{\sum_{p\in\mathcal{P}} [\![y_p \neq y_{m(p)}]\!] w_p,\ M^{\max}\right\} \quad (1)$$

where w_p is a non-negative per-pixel weight, and $[\![T]\!]$ is the indicator function, taking the value one if T is true and zero otherwise. The potential function is illustrated in Figure 2.

Intuitively, $M(\boldsymbol{y};\mathcal{P},\mathcal{Q})$ penalizes the (weighted) number of disagreements between labels of pixels in $\mathcal{P}$ and the corresponding pixels in $\mathcal{Q}$ up to some maximum penalty $M^{\max}$. The rate at which pixel disagreements are penalized is controlled by the parameters w_p, which can be set differently for each matching pair of pixels. The potential allows us to encode a preference for two patches in an image to take the same *label configuration*. It is not necessary, however, that all variables in the scope of the potential take the same label as enforced by other higher-order approaches, *e.g.*, [14]. For example, some pixels in the patch could be labeled as car and others as road. The potential is truncated so that we only pay a maximum penalty when this preference is wrong (*i.e.*, outweighed by other evidence).

By introducing an auxiliary binary variable z, the truncated higher-order matching potential in Equation 1 can be re-written as

$$\begin{aligned} &M(\boldsymbol{y};\mathcal{P},\mathcal{Q}) \\ &= \min_{z\in\{0,1\}} \sum_{p\in\mathcal{P}} z[\![y_p \neq y_{m(p)}]\!] w_p + (1-z)M^{\max} \quad (2) \\ &= \min_{z\in\{0,1\}} \sum_{p\in\mathcal{P}} f_p\big(y_p, y_{m(p)}, z; w_p\big) + (1-z)M^{\max} \quad (3) \end{aligned}$$

where for brevity in the subsequent discussion we have replaced the summand in Equation 2 by the function f_p over variables y_p, $y_{m(p)}$, and z, and parameterized by w_p.

As discussed in Section 2, our truncated higher-order matching potential and its transformation to a minimization over the sum of terms in Equation 3 shares similarities with the robust-P^n model introduced by Kohli *et al*. [13]. However, unlike their model our terms encode a preference for matching labels between two distinct sets of pixels rather than a set of pixels and single label. Like, the robust-P^n model, our formulation can be extended to an arbitrary non-decreasing concave function over the number of mismatching labels, *e.g.*, by replacing Equation 1 with a minimization over a set of linear functions of the form $a_k \sum_{p\in\mathcal{P}} [\![y_p \neq y_{m(p)}]\!] w_p + b_k$. However, we do not consider this extension further in this paper.

Note that if $M^{\max} \geq \sum_{p\in\mathcal{P}} w_p$ then Equation 1 reduces to $M(\boldsymbol{y};\mathcal{P},\mathcal{Q}) = \sum_{p\in\mathcal{P}} [\![y_p \neq y_{m(p)}]\!] w_p$, which is simply a sum of individual Potts potentials. In this case the potential is not truncated, and the energy function can be solved efficiently by well-known techniques (*e.g.*, α-expansion or $\alpha\beta$-swap [6]). Furthermore, if $w_p \geq M^{\max}$ for all $p \in \mathcal{P}$ then $M(\boldsymbol{y};\mathcal{P},\mathcal{Q}) = M^{\max}$ for all assignments to $\boldsymbol{y}$. The case of $M^{\max} < \sum_{p\in\mathcal{P}} w_p < M^{\max}|\mathcal{P}|$ is more interesting and requires a specialized optimization approach.

3.2. Move-making Optimization

A popular approach to minimizing energy functions arising in computer vision is to use a class of algorithms known as move-making algorithms. Here optimization is performed by a series of moves, each of which projects the problem onto a restricted state-space and finds an optimal or near-optimal solution in the reduced space. When the subproblems can be solved exactly strong convergence and optimality conditions can sometimes be guaranteed.

One of the most successful move-making algorithm for energy functions appearing in multiclass labeling problems is the α-expansion algorithm of Boykov *et al*. [6]. Here, each move considers keeping a random variable's current assignment or switching its assignment to a given $\alpha \in \mathcal{L}$. It is well-known that when the energy function defined over the restricted state-space is submodular (and pairwise) the optimal α-expansion move can be found by finding the min-st-cut on a suitably constructed graph [3, 15].[1]

Concretely, let $\boldsymbol{y}^{\text{prev}}$ be the current best assignment to the variables (*i.e.*, before the α-expansion move) and let $\boldsymbol{y}^{\text{next}}$ be the assignment to the variables after the move. We define $\widetilde{y}_p^{\alpha} \in \{0,1\}$ to be the binary variable associated with multiclass variable $y_p \in \mathcal{L}$ for a given α-expansion move where the value 0 indicates no change in the corresponding variable's assignment and the value 1 indicates changing the corresponding variable's assignment to α. Therefore, after the move we have

$$y_p^{\text{next}} = \begin{cases} y_p^{\text{prev}} & \text{if } \widetilde{y}_p^{\alpha} = 0 \\ \alpha & \text{if } \widetilde{y}_p^{\alpha} = 1. \end{cases} \quad (4)$$

[1] Furthermore, for computer vision applications, very efficient implementations of graph-cut algorithms exist and are publicly available [5].

Now let $\widetilde{f}_p^{\alpha}(\widetilde{y}_p^{\alpha}, \widetilde{y}_q^{\alpha}, z; w_p)$ be the projection of f_p onto the restricted (binary) move-space. We identify five cases for this projection and partition $\mathcal{P} = \bigcup_{i=1}^{5} \mathcal{P}_i$ into five disjoint sets corresponding to each case defined as follows:

$$\mathcal{P}_1 = \left\{p \in \mathcal{P} : y_p^{\text{prev}} = \alpha \text{ and } y_q^{\text{prev}} = \alpha\right\} \quad (5)$$

$$\mathcal{P}_2 = \left\{p \in \mathcal{P} : y_p^{\text{prev}} = \alpha \text{ and } y_q^{\text{prev}} \neq \alpha\right\} \quad (6)$$

$$\mathcal{P}_3 = \left\{p \in \mathcal{P} : y_p^{\text{prev}} \neq \alpha \text{ and } y_q^{\text{prev}} = \alpha\right\} \quad (7)$$

$$\mathcal{P}_4 = \left\{p \in \mathcal{P} : y_p^{\text{prev}} \neq y_q^{\text{prev}} \text{ and } y_p^{\text{prev}}, y_q^{\text{prev}} \neq \alpha\right\} \quad (8)$$

$$\mathcal{P}_5 = \left\{p \in \mathcal{P} : y_p^{\text{prev}} = y_q^{\text{prev}} \neq \alpha\right\} \quad (9)$$

where $q = m(p)$ is the pixel in $\mathcal{Q}$ matched to pixel $p \in \mathcal{P}$. It is easy to show that $\widetilde{f}_p^{\alpha}$ can be represented by the following pseudo-Boolean function [3]:

$$\widetilde{f}_p^{\alpha} = \begin{cases} 0 & \text{if } p \in \mathcal{P}_1 \\ w_p z\left(1 - \widetilde{y}_q^{\alpha}\right) & \text{if } p \in \mathcal{P}_2 \\ w_p z\left(1 - \widetilde{y}_p^{\alpha}\right) & \text{if } p \in \mathcal{P}_3 \\ w_p z\left(1 - \widetilde{y}_p^{\alpha}\widetilde{y}_q^{\alpha}\right) & \text{if } p \in \mathcal{P}_4 \\ w_p z\left(\widetilde{y}_p^{\alpha} + \widetilde{y}_q^{\alpha} - 2\widetilde{y}_p^{\alpha}\widetilde{y}_q^{\alpha}\right) & \text{if } p \in \mathcal{P}_5 \end{cases} \quad (10)$$

Unfortunately, for $p \in \mathcal{P}_5$ the pseudo-Boolean function $\widetilde{f}_p^{\alpha}$ is non-submodular. Typical approaches for dealing with non-submodular terms are:

- approximate the potential by one that is submodular;
- relax the problem (*e.g.*, using QPBO [25] or dual decomposition [17]) and round the (fractional) result onto the variable state-space;
- coordinate descent (*e.g.*, hold the z's for each higher-order term fixed and optimize over $\boldsymbol{y}$, then hold $\boldsymbol{y}$ fixed and optimize over the z's) or exhaustive search over the z's (if the number of higher-order potentials is small).

We take the first approach and approximate the f_p term as

$$f_p(y_p, y_{m(p)}, z; w_p) \approx w_p [\![y_p \neq y_{m(p)}]\!] \quad (11)$$

whenever this situation occurs during an α-expansion move, *i.e.*, when the labels for p and q are equal before the move (and not equal to α). We stress that this approximation is only applied to the subset of terms meeting the above condition (*i.e.*, $p \in \mathcal{P}_5$), and note that it is an over-approximation of the original potential function f_p (and is tight when $z = 1$). The effect is to favor configurations where the corresponding pixels continue to agree on their labels.

With this approximation our energy function is always graph-representable and an optimal α-expansion move (with respect to the approximate energy function) can be found. For $p \in \mathcal{P}_4$ we use the construction of Ishikawa [12] to convert $\widetilde{f}_p^{\alpha}$ from a cubic pseudo-Boolean function to a quadratic one. Specifically we introduce an auxiliary binary variable a_p for all $p \in \mathcal{P}_4$ and write $\widetilde{f}_p^{\alpha}$ as

$$w_p \left(z + \min_{a_p \in \{0,1\}} 2a_p - \tilde{y}_p^{\alpha} a_p - \tilde{y}_q^{\alpha} a_p - z a_p \right). \quad (12)$$

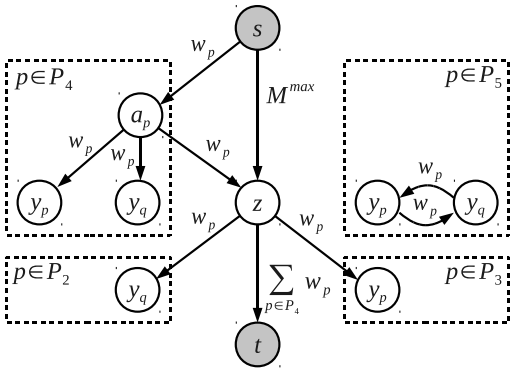

Figure 3. Graph construction for performing an α-expansion move on the truncated higher-order matching potential, $M(\boldsymbol{y}; \mathcal{P}, \mathcal{Q})$. The a_p are auxiliary nodes introduced by the Ishikawa construction for $p \in \mathcal{P}_4$. The other nodes correspond to variables in the potential function. The minimum st-cut in the graph corresponds to the optimal assignment to $\boldsymbol{y}$. Variables corresponding to nodes in the $\mathcal{S}$ partition are assigned label 1. The remaining variables are assigned label 0.

Figure 3 shows the complete st-graph construction for a single truncated higher-order matching potential. As is standard for approximate moves, at the end of each α-expansion step we compare the solution found with the current best assignment, and keep the one with lower energy. This process is continued, repeatedly iterating through all $\alpha \in \mathcal{L}$, until no improvement in the energy can be found. This approach appears to work well in practice even though we cannot provide a formal guarantee on the quality of the solution at convergence due to our over-approximation of terms from $\mathcal{P}_5$.

4. Multiclass Pixel Labeling Model

In this section we describe our multiclass pixel labeling model. We begin by describing the components of the standard pairwise energy function. We then introduce the truncated higher-order matching potentials into this model.

4.1. Pairwise CRF Model

Our model extends the standard pairwise CRF model for multiclass pixel labeling [11, 26]. Here the energy for a pixel labeling $\boldsymbol{y} \in \mathcal{L}^{W \times H}$ given image features $\boldsymbol{x}$ is defined over unary and pairwise terms as

$$E(\boldsymbol{y}; \boldsymbol{x}) = \sum_p \psi_p(y_p) + \lambda \sum_{pq} \psi_{pq}(y_p, y_q) \quad (13)$$

where ψ_p is the unary potential for assigning label y_p to pixel p and ψ_{pq} is a contrast-dependent smoothing prior that penalizes adjacent pixels p and q for taking different labels.[2]

[2]For brevity we omit the features $\boldsymbol{x}$ from the arguments of the potential functions ψ_p and ψ_{pq}. It should be understood that all potential functions are conditioned on $\boldsymbol{x}$.

The non-negative constant λ trades-off the strength of the smoothness prior against the unary potential and is chosen by cross-validation on the training set.

An implementation of the model we use is provided by the Darwin software library (version 0.9)[3]. Briefly, the unary term ψ_p is constructed by learning one-versus-all boosted decision tree classifiers for each label in $\mathcal{L}$. The input to the boosted classifiers are 669-element per-pixel feature vectors comprised of 17-dimensional filter bank responses, dense HOG descriptors, RGB color, and the normalized x and y coordinates of the pixel. For the filter bank and HOG features we also compute the mean and standard deviation over pixels within the same row and column as p and over 5×5 pixel regions in a 3×3 grid centered on p. Once the one-versus-all boosted decision tree classifiers are learned, their outputs are calibrated via a multiclass logistic classifier [23] and $\psi_p(y_p)$ taken as the negative log-likelihood predicted by the multiclass logistic for class y_p.

As is typical for pixel labeling CRFs [4], our contrast-dependent smoothness prior ψ_{pq} takes the form

$$\psi_{pq}(y_p, y_q) = \begin{cases} \exp\left(-\frac{\|x_p - x_q\|^2}{2\beta}\right), & \text{if } y_p \neq y_q \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

where x_p and x_q are the RGB color vectors for pixels p and q, respectively, and β is the mean square-difference between color vectors over all adjacent pixels in the image.

4.2. Higher-Order CRF Model

We append to the pairwise CRF (Equation 13) one truncated higher-order matching potential (Equation 1) for each pair of matching regions $(\mathcal{P}^{(t)}, \mathcal{Q}^{(t)})$ to give

$$E(\boldsymbol{y}; \boldsymbol{x}) = \underbrace{\sum_p \psi_p(y_p)}_{\text{unary term}} + \lambda \underbrace{\sum_{pq} \psi_{pq}(y_p, y_q)}_{\text{smoothness term}} + \mu \underbrace{\sum_t M_t(\boldsymbol{y}; \mathcal{P}^{(t)}, \mathcal{Q}^{(t)})}_{\text{higher-order term}} \quad (15)$$

where t indexes the region pairs. The constant μ trades-off the strength of the truncated higher-order matching potentials against other terms in the model and is set by finding the optimal value on the training set of images.

In our experiments, we find matching regions by densely sampling rectangular patches of size 32×32 to 96×96 in 16 pixel increments. Patches with nearly uniform appearance were discarded. For each remaining patch $\mathcal{P}$, we compute the normalized cross-correlation (NCC) between the patch and image in a sliding-window fashion (excluding the original location of the patch), *i.e.*, for each candidate match $\mathcal{Q}$, we compute

$$NCC(\mathcal{P}, \mathcal{Q}) = \frac{\sum_{p \in \mathcal{P}} x_p^T x_{m(p)}}{\sqrt{\sum_{p \in \mathcal{P}} \|x_p\|^2 \cdot \sum_{q \in \mathcal{Q}} \|x_q\|^2}} \quad (16)$$

where x_p is the 3-element RGB feature vector for pixel p.[4] The mapping $m : \mathcal{P} \to \mathcal{Q}$ is defined in the obvious way, *i.e.*, $q = m(p)$ is pixel q that has the same relative offset from the top-left of rectangular region $\mathcal{Q}$ as pixel p has from the top-left of rectangular region $\mathcal{P}$. To capture reflective symmetry, we also compute the normalized cross-correlation with a horizontally flipped version of the patch.

We discard any candidate pair $(\mathcal{P}, \mathcal{Q})$ whose NCC is below 0.9 and then perform non-maximal neighborhood suppression (with 0.5 area-of-overlap criterion) on the remaining pairs to remove densely overlapping matches. The resulting pairs of matched regions are used to construct the truncated higher-order matching potential terms.

To make our approach robust to small deformations in object boundaries and misalignments within a matched region we weight each pixelwise match by how well the individual pixel colors agree (see Figure 4). Specifically, for a given pair of regions $(\mathcal{P}^{(t)}, \mathcal{Q}^{(t)})$, we set

$$w_p^{(t)} = \frac{1}{|\mathcal{P}^{(t)}|} \cdot \frac{2 x_p^T x_{m(p)}}{\|x_p\|^2 + \|x_{m(p)}\|^2} \quad (17)$$

We then set the maximum penalty for the potential to

$$M^{\max} = \kappa \sum_{p \in \mathcal{P}^{(t)}} w_p^{(t)} \quad (18)$$

where $\kappa \in [0, 1]$ is a parameter shared between all higher-order potentials. In our work we set κ to 0.85. This completes the specification for each truncated higher-order matching term.

5. Experimental Results

We performed experiments on the multiclass pixel labeling task and compared results on CRF models with and without our truncated higher-order matching potentials. Our experiments were conducted on two standard datasets: (i) the 21-class MSRC dataset [7] consisting of 591 images, and (ii) the 8-class Stanford Background dataset [10] consisting of 715 images. We use the same train/test split as [26] for the MSRC dataset (335 training and 256 evaluation images). For the Stanford Background dataset we follow previous works and randomly partition the images into sets of 572 and 143 images for training and testing, respectively.

[3] http://drwn.anu.edu.au/

[4] The use of RGB features is not critical to our approach and could have been replaced with other pixelwise features, *e.g.*, CIELab, HOG, *etc*. Moreover, our algorithm does not appear to be sensitive to the choice of NCC matching score.

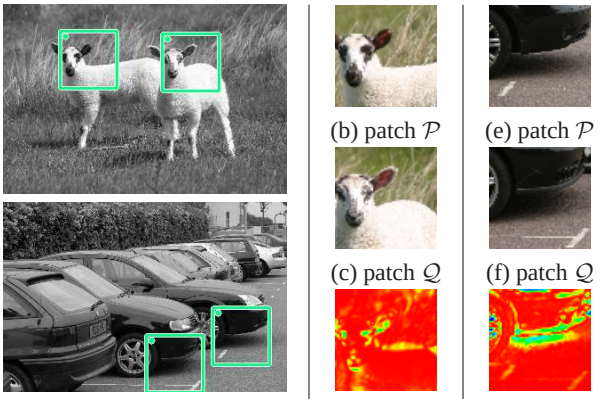

Figure 4. Illustration of weights ((d) and (g)) assigned to pixels within matched regions ((b) and (c), and (e) and (f), respectively). Panels (d) and (g) are colored with red indicating a higher weight.

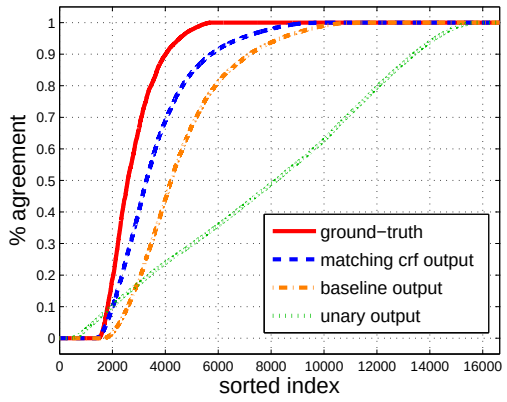

Figure 5. Plot showing percentage agreement between corresponding pixels in matched regions with respect to (i) ground-truth labels, (ii) output from our model with higher-order matching potentials, (iii) baseline CRF output, and (iv) unary model output.

We learned all parameters of our models from the subset of training images with the objective of maximizing (global-averaged) pixelwise accuracy. Results are reported on the hold-out set of evaluation images. Inference in our model is very fast, typically taking less than two seconds per image. Note that this does not include the time taken to compute image features nor perform the region matching, both of which take considerably longer.

Figure 5 shows a cumulative plot of label agreement between pixels within matching regions (on the MSRC dataset). The vast majority of matches demonstrate near perfect agreement supporting our claim that similarity in appearance implies similarity in semantic label. However, approximately 10% of the matches show very poor agreement. Our method is robust to these inconsistent matches.

Quantitative results from are shown in Table 1. For both datasets the inclusion of the truncated higher-order matching potential provides a small increase in accuracy: **1.2%** for the MSRC dataset and **0.98%** for the Stanford Background dataset. The table also shows results from a variant of our model without truncation (*i.e.*, setting $\kappa = 1$ in Eqn. (18)). We note that our MSRC result is below the state-of-the-art result of 86% by Ladicky *et al*. [18]. However, as noted in their work, significant improvement comes from strong unary terms as a result of better pixel-level features. The purpose of our work is to explore the higher-order matching potential rather than engineer new features.

Qualitative results are shown in Figure 6 and Figure 7 for the MSRC and Stanford Background datasets, respectively. Note that for visualization, column (b) only shows the top ten non-overlapping matched regions—the model used to generate the results in column (d) contains many more matches. As can be seen, many of the matches are over regions where the baseline CRF model already labels the scene correctly and therefore have no effect on the final labeling. The remaining matches, however, can result in significant improvements to the segmentation quality.

While our model allows pixels in the image mislabeled by the baseline CRF model to be corrected, it can also result in a degradation in performance, for example, when incorrectly classified pixels on one side of the match have a stronger influence on *a priori* correctly labeled pixels on the other (see, for example, the last two rows of Figure 6).

6. Discussion

Much recent work on pixel labeling problems has focused on the addition of higher-order energy terms to encode preferences for particular label configurations. We have explored one such term that encodes a novel preference for consistent label assignments between two matching image regions. We showed how to perform efficient approximate inference in models with such terms using a graph-cut construction. Furthermore, we demonstrated the model on two standard multiclass pixel labeling datasets.

Our work suggest a number of interesting areas for further research. First, it would be interesting to consider pixel mappings that are not necessarily one-to-one. For example, matching two regions at different scales would result in a many-to-one mapping from pixels at the fine scale to those at the coarse scale. Non-rectangular matches could also be explored, *e.g.*, between superpixels.

A second exciting area for future work is in matching regions over multiple images instead of within a single image. This may be particularly relevant for videos or collections of images with significant semantic overlap, *e.g.*, taken in the same geographic vicinity.

Acknowledgments. This work was supported by the Australian Research Council and the NCI National Facility at the ANU. We thank the anonymous reviewers for their feedback in improving this paper.

		Global-averaged				Class-averaged			
	Test Set	Baseline		Higher-Order		Baseline		Higher-Order	
Dataset	Images	Unary	Pairwise	$\kappa = 1$	$\kappa = 0.85$	Unary	Pairwise	$\kappa = 1$	$\kappa = 0.85$
MSRC	256	73.83	79.73	80.75	**80.97**	64.15	69.03	70.76	**71.06**
Stanford	143	73.01	78.63	79.07	**79.61**	68.27	72.24	72.45	**72.87**

Table 1. Pixelwise semantic labeling accuracy for 21-class MSRC [7] and 8-class Stanford Background [10] datasets. Compares baseline unary and pairwise CRF model against model with non-truncated and truncated higher-order matching potentials.

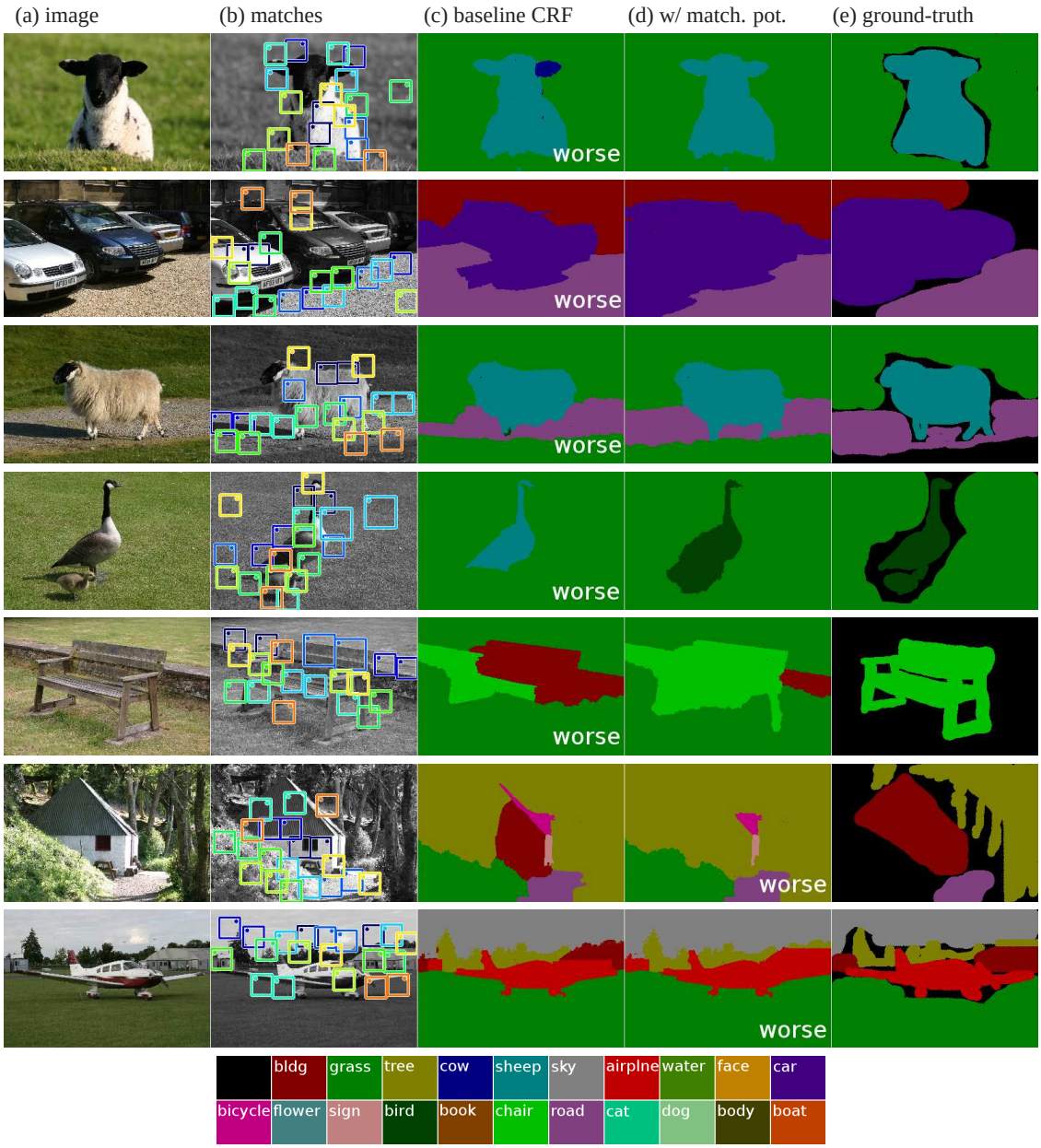

Figure 6. **Best viewed in color.** Example results from our multiclass pixel labeling experiments on the 21-class MSRC dataset [7]. Each row shows a different instance. The test image is shown in column (a) and top ten non-overlapping matches shown in (b). Matched regions are color-coded and the orientation of the match indicated by the dot in the upper-left or upper-right corner of the region. Semantic class predictions for the baseline model and one with truncated higher-order matching potentials are shown in columns (c) and (d), respectively.

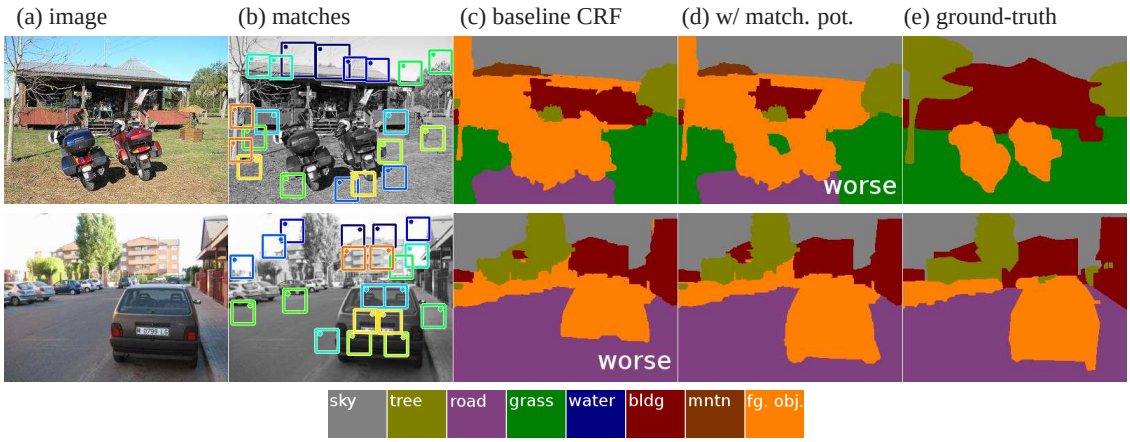

Figure 7. **Best viewed in color.** Example results from our multiclass pixel labeling experiments on the 8-class Stanford Background dataset [10]. See Figure 6 for description of panels.

References

[1] S. Bagon, O. Boiman, and M. Irani. What is a good image segment? a unified approach to segment extraction. In *ECCV*, 2008.

[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Stats. Society Series B*, 48:259–302, 1986.

[3] E. Boros and P. L. Hammer. Pseudo-boolean optimization. *Discrete Appl. Math.*, 123:155–225, 2002.

[4] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *ICCV*, 2001.

[5] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 2004.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *ICCV*, 1999.

[7] A. Criminisi. Microsoft research cambridge (MSRC) object recognition image database (version 2.0), 2004.

[8] A. Delong, A. Osokin, H. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. In *CVPR*, 2010.

[9] G. Geman and D. Geman. Stochastic relaxations, Gibbs distributions and the Bayesian restoration of images. *PAMI*, pages 721–741, 1984.

[10] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.

[11] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[12] H. Ishikawa. Higher-order clique reduction in binary graph cut. In *CVPR*, 2009.

[13] P. Kohli, M. Kumar, and P. Torr. P3 & beyond: Solving energies with higher order cliques. In *CVPR*, 2007.

[14] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008.

[15] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 2004.

[16] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *CVPR*, 2009.

[17] N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *ICCV*, 2007.

[18] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical CRFs for object class image segmentation. In *ICCV*, 2009.

[19] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut based inference with co-occurrence stats. In *ECCV*, 2010.

[20] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp. Image segmentation with a bounding box prior. In *ICCV*, 2009.

[21] Y. Lim, K. Jung, and P. Kohli. Energy minimization under constraints on label counts. In *CVPR*, 2010.

[22] S. Nowozin and C. H. Lampert. Global connectivity potentials for random field models. In *CVPR*, 2009.

[23] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 1999.

[24] C. Rother, P. Kohli, W. Feng, and J. Jia. Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, 2009.

[25] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary MRFs via extended roof duality. In *CVPR*, 2007.

[26] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recog. and segm. In *ECCV*, 2006.

[27] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *CVPR*, 2008.

[28] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segmentation. In *CVPR*, 2010.

[29] G. Zeng and L. V. Gool. Multi-label image segmentation via point-wise repetition. In *CVPR*, 2008.

[30] L. Zhu, Y. Chen, Y. Lin, C. Lin, and A. Yuille. Recursive segmentation and recognition templates for 2D parsing. In *NIPS*, 2008.