Robust Higher Order Potentials for Enforcing Label Consistency

Pushmeet Kohli
Microsoft Research Cambridge
pkohli@microsoft.com

L'ubor Ladický Philip H. S. Torr
Oxford Brookes University
{lladicky,philiptorr}@brookes.ac.uk

Abstract

This paper proposes a novel framework for labelling problems which is able to combine multiple segmentations in a principled manner. Our method is based on higher order conditional random fields and uses potentials defined on sets of pixels (image segments) generated using unsupervised segmentation algorithms. These potentials enforce label consistency in image regions and can be seen as a strict generalization of the commonly used pairwise contrast sensitive smoothness potentials. The higher order potential functions used in our framework take the form of the Robust P^n model. This enables the use of powerful graph cut based move making algorithms for performing inference in the framework [14]. We test our method on the problem of multi-class object segmentation by augmenting the conventional CRF *used for object segmentation with higher order potentials defined on image regions. Experiments on challenging data sets show that integration of higher order potentials quantitatively and qualitatively improves results leading to much better definition of object boundaries. We believe that this method can be used to yield similar improvements for many other labelling problems.*

1. Introduction

In recent years an increasingly popular way to solve various image labelling problems like object segmentation, stereo and single view reconstruction is to formulate them using image segments (so called superpixels) obtained from unsupervised segmentation algorithms [9, 10, 22]. These methods are inspired from the observation that pixels constituting a particular segment often have the same label; for instance, they may belong to the same object or may have the same surface orientation. This approach has the benefit that higher order features based on all the pixels constituting the segment can be computed and used for classification[1]. Further, it is also much faster as inference now only needs to be performed over a small number of superpixels rather than all the pixels in the image.

Methods based on grouping segments make the assumption that segments are consistent with object boundaries in the image [9], i.e. segments do not contain multiple objects. As observed by [11] and [26] this is not always the case and segments obtained using unsupervised segmentation methods are often wrong. To overcome these problems [11] and [26] use multiple segmentations of the image (instead of only one) in the hope that although most segmentations are bad, some are correct and thus would prove useful for their task. They merge the multiple superpixels using heuristic algorithms which lack any optimality guarantees and thus may produce bad results. In this paper we propose an algorithm that can compute the solution of the labelling problem (using features based on image segments) in a principled manner. Our approach couples potential functions defined on sets of pixels with conventional unary and pairwise cues using higher order CRFs. We test the performance of this method on the problem of object segmentation and recognition. Our experiments show that the results of our approach are significantly better than the ones obtained using pairwise CRF models (see figure 1). †

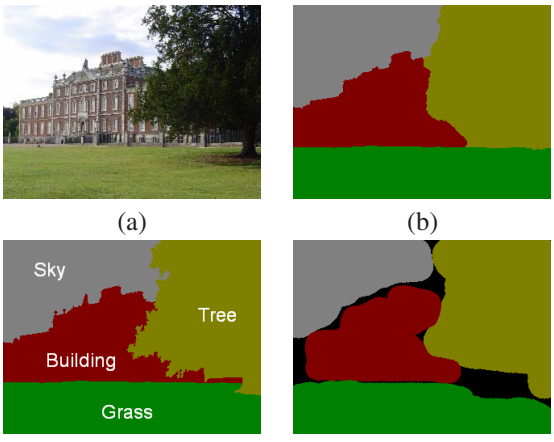


Figure 1. *Using higher order potentials for object segmentation. (a) An image from the MSRC-23 dataset. (b) The object segmentation obtained by performing inference in the pairwise CRF defined in section 2.1 which uses unary likelihood potentials from Textonboost [29]. (c) Our segmentation result which was obtained by augmenting the pairwise CRF with higher order potentials defined on image segments. The segments were generated by changing the parameters values in the mean-shift segmentation algorithm [6] (as explained in section 3.4). (d) The rough hand labelled segmentations provided in the MSRC data set. It can be clearly seen that the use of higher order potentials results in a significant improvement in the segmentation result. For instance, the branches of the tree are much better segmented.*

Object Segmentation and Recognition Combined object segmentation and recognition is one of the most challenging and fundamental problems in computer vision. The

[1] In some sense this causes the problem of scene understanding to be decoupled from the image resolution given by the hardware; it is conducted using more natural primitives that are independent of resolution.

† This work was supported by the EPSRC research grant GR/T21790/01(P), HMGCC and the IST Programme of European Community, under the PASCAL Network of Excellence.

last few years have seen the emergence of object segmentation algorithms which integrate *object specific* top-down information with *image based* low-level features [2, 8, 12, 16, 19]. These methods have produced excellent results on challenging data sets. However, they typically only deal with one object at a time in the image independently and do not provide a framework for understanding the whole image. Further, their models become prohibitively large as the number of classes increases. This prevents their application to scenarios where segmentation and recognition of many object classes is desired.

Shotton *et al.* [29] recently proposed a method (*Textonboost*) to overcome this problem. In contrast to using explicit models to encode object shape they used a boosted combination of *texton* features which jointly modeled shape and texture. They combine the result of textons with colour and location based likelihood terms in a condition random field (CRF). Although their method produced good segmentation and recognition results, the rough shape and texture model caused it to fail at object boundaries. The problem of extracting accurate boundaries of objects is considerably more challenging. In what follows we show that incorporation of higher order potentials defined on superpixels dramatically improves the object segmentation result. In particular, it leads to segmentations with much better definition of object boundaries as shown in figure 1.

Higher Order CRFs Higher order random fields are not new to computer vision. They have been frequently used to model image textures [18, 20, 24]. The initial work in this regard has been quite promising and higher order CRFs have been shown to improve results for problems such as image denoising and restoration [24], and texture segmentation [13]. However, the lack of efficient algorithms for performing inference in these models has limited their applicability. Traditional inference algorithms such as BP become computationally expensive for higher order CRFs. Recent work has been partly successful in improving their performance for certain classes of potential functions. Lan *et al.* [18] proposed approximation methods for BP to make efficient inference possible in higher order MRFs. This was followed by the recent work of Potetz [21] in which he showed how belief propagation can be efficiently performed in graphical models containing moderately large cliques. However, as these methods were based on BP, they were quite slow and took minutes or hours to converge.

Kohli *et al.* [13] recently introduced a class of higher order potentials called the P^n Potts model and showed that they can be minimized using the graph cuts based move making algorithms, namely, α-expansion and $\alpha\beta$-swap [4]. The higher order potential functions used in our framework take the form of the Robust P^n model. This model is more general than the P^n Potts model and cannot be minimized using the method of [13]. We have shown that energy functions composed of these *robust* potentials can be minimized using α-expansion and $\alpha\beta$-swap algorithms [14]. The complexity of our algorithm increases linearly with the size of the clique which makes it able to handle cliques composed of thousands of latent variables.

Overview of the Paper This paper proposes a general framework for solving labelling problems which has the ability of utilizing higher order potentials defined on segments. We test this framework on the problem of object segmentation and recognition by integrating potentials encouraging label consistency in segments with conventionally used unary and pairwise potentials. Inference in this framework is performed using graph cut based move making algorithms [14]. To summarize, the novelties of our approach include:

1. A novel higher order region consistency potential which is a strict generalization of the commonly used pairwise contrast sensitive smoothness potential.
2. The application of higher order CRFs for object segmentation and recognition which integrate the above mentioned higher order potentials with conventional unary and pairwise potentials based on colour, location, texture, and smoothness.

An outline of the paper follows. In section 2 we discuss the basic theory of conditional random fields. We then show how pairwise CRFs can be used to model labelling problems like object segmentation. In section 3 we augment the pairwise CRF model by incorporating novel higher order potentials based on super-pixel segmentations. The experimental results of our method are given in section 4. These include qualitative and quantitative results on well known and challenging data sets for object segmentation and recognition. The conclusions and directions for future work are listed in section 5.

2. Preliminaries

We start by providing the basic notation used in the paper. Consider a discrete random field $\mathbf{X}$ defined over a lattice $\mathcal{V} = \{1, 2, \ldots, N\}$ with a neighbourhood system $\mathcal{N}$. Each random variable $X_i \in \mathbf{X}$ is associated with a lattice point $i \in \mathcal{V}$ and takes a value from the label set $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$. The neighborhood system $\mathcal{N}$ of the random field is defined by the sets $\mathcal{N}_i, \forall i \in \mathcal{V}$, where $\mathcal{N}_i$ denotes the set of all neighbours of the variable X_i. A clique c is a set of random variables $\mathbf{X}_c$ which are conditionally dependent on each other. Any possible assignment of labels to the random variables will be called a *labelling* (denoted by $\mathbf{x}$) which takes values from the set $\mathbf{L} = \mathcal{L}^N$.

A random field is said to be *Markov* with respect to a neighborhood system $\mathcal{N} = \{\mathcal{N}_v | v \in \mathcal{V}\}$ if and only if it satisfies the positivity property: $\Pr(\mathbf{x}) > 0 \;\; \forall \mathbf{x} \in \mathcal{X}^n$, and the Markovian property:

$$\Pr(x_v | \{x_u : u \in \mathcal{V} - \{v\}\}) = \Pr(x_v | \{x_u : u \in \mathcal{N}_v\}), \quad (1)$$

for all $v \in \mathcal{V}$. Here we refer to $\Pr(X = \mathbf{x})$ by $\Pr(\mathbf{x})$ and $\Pr(X_i = x_i)$ by $\Pr(x_i)$. A conditional random field (CRF) may be viewed as an MRF globally conditioned on the data.

The posterior distribution $\Pr(\mathbf{x}|\mathbf{D})$ over the labellings of the conditional random field is a *Gibbs* distribution and can be written in the form: $\Pr(\mathbf{x}|\mathbf{D}) = \frac{1}{Z}\exp(-\sum_{c\in\mathcal{C}}\psi_c(\mathbf{x}_c))$, where Z is a normalizing constant known as the partition function, and $\mathcal{C}$ is the set of all cliques [17]. The term $\psi_c(\mathbf{x}_c)$ is known as the potential function of the clique c where $\mathbf{x}_c = \{x_i, i \in c\}$. The corresponding Gibbs energy is given by

$$E(\mathbf{x}) = -\log\Pr(\mathbf{x}|\mathbf{D}) - \log Z = \sum_{c\in\mathcal{C}}\psi_c(\mathbf{x}_c). \quad (2)$$

The most probable or maximum a posterior (MAP) labelling $\mathbf{x}^*$ of the random field is defined as: $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathbf{L}}\Pr(\mathbf{x}|\mathbf{D}) = \arg\min_{\mathbf{x}\in\mathbf{L}}E(\mathbf{x})$.

2.1. Pairwise CRFs for Object Segmentation

The CRF models commonly used for object segmentation are characterized by energy functions defined on unary and pairwise cliques as:

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}}\psi_i(x_i) + \sum_{i\in\mathcal{V},j\in\mathcal{N}_i}\psi_{ij}(x_i,x_j). \quad (3)$$

Here $\mathcal{V}$ corresponds to the set of all image pixels, $\mathcal{N}$ is a neighbourhood defined on this set which is commonly chosen to be either a 4 or 8 neighbourhood. The labels constituting the label set $\mathcal{L}$ represent the different objects. The random variable x_i denotes the labelling of pixel i of the image. Every possible assignment of the random variables $\mathbf{x}$ (or configuration of the CRF) defines a segmentation.

The unary potential ψ_i of the CRF is defined as the negative log of the likelihood of a label being assigned to pixel i. It can be computed from the colour of the pixel and the appearance model for each object. However, colour alone is not a very discriminative feature and fails to produce accurate segmentations. This problem can be overcome by using sophisticated potential functions based on colour, texture, location, and shape priors as shown by [1, 5, 16, 25, 29]. The unary potential used by us can be written as:

$$\psi_i(x_i) = \theta_T\psi_T(x_i) + \theta_{col}\psi_{col}(x_i) + \theta_l\psi_l(x_i) \quad (4)$$

where θ_T, θ_{col}, and θ_l are parameters weighting the potentials obtained from TextonBoost(ψ_T) [29], colour(ψ_{col}) and location(ψ_l) respectively.

The pairwise terms ψ_{ij} of the CRF take the form of a contrast sensitive Potts model:

$$\psi(x_i,x_j) = \begin{cases} 0 & \text{if } x_i = x_j, \\ g(i,j) & \text{otherwise,} \end{cases} \quad (5)$$

where the function $g(i,j)$ is an edge feature based on the difference in colors of neighboring pixels [3]. It is typically defined as:

$$g(i,j) = \theta_p + \theta_v\exp(-\theta_\beta||I_i - I_j||^2), \quad (6)$$

where I_i and I_j are the colour vectors of pixel i and j respectively. θ_p, θ_v, and θ_β are model parameters whose values are learned using training data. We refer the reader to [3, 25, 29] for more details.

Inferring the most probable segmentation The object segmentation problem can be solved by finding the least energy configuration of the CRF defined above. As the pairwise potentials of the energy function (3) are of the form of a Potts model it can be minimized approximately using the well known α-expansion algorithm [4]. The resulting segmentation can be seen in figure 1. We also tried other energy minimization algorithms such as sequential tree-reweighted message passing (TRW-S) [15, 31]. The α-expansion algorithm was preferred because it was faster and gave a solution with lower energy compared to TRW-S.

Need for higher order CRFs The use of Potts model[4] potentials in the CRF model makes it favour smooth object boundaries. Although this improves results in most cases it also introduces an undesirable side effect. Smoothness potentials make the model incapable of extracting the fine contours of certain object classes such as trees and bushes. As seen in the results, segmentations obtained using pairwise CRFs tend to be oversmooth and quite often do not match the actual object contour. In the next section we show how these results can be significantly improved by using higher order potentials derived from multiple segmentations obtained from an unsupervised image segmentation method.

3. Incorporating Higher Order Potentials

Methods based on grouping regions for segmentation generally make the assumption that all pixels constituting a particular segment (or region) belong to the same object [9]. This is not always the case, and image segments quite often contain pixels belonging to multiple object classes. For instance, in the segmentations shown in figure 2 the bottom image segment contains some 'building' pixels in addition to all the grass pixels.

Unlike other object segmentation algorithms which use the label consistency in segments as a hard constraint, our method uses it as a *soft constraint*. This is done by using higher order potentials defined on the image segments generated using unsupervised segmentation algorithms. Specifically, we augment the pairwise CRF model explained in the previous section by incorporating higher order potentials defined on sets or regions of pixels. The Gibbs energy of this higher order CRF can now be written as:

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}}\psi_i(x_i) + \sum_{i\in\mathcal{V},j\in\mathcal{N}_i}\psi_{ij}(x_i,x_j) + \sum_{c\in\mathcal{S}}\psi_c(\mathbf{x}_c), \quad (7)$$

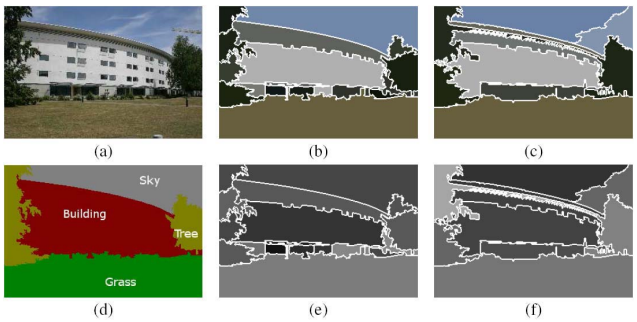


Figure 2. *Quality sensitive region consistency prior. (a) An image from the MSRC data set. (b) and (c) Two different segmentations of the image obtained using different parameter values for the mean-shift algorithm. (d) A hand labelled object segmentation of the image. (e) and (f) The value of the variance based quality function $G(c)$ (see equation 10) computed over the segments of the two segmentations. Segments with high quality values are darker. It can be clearly seen that segments which contain multiple object classes have been assigned low quality. For instance, the top segment of the left tree in segmentation (c) includes a part of the building and thus is brighter in the image (f) indicating low quality. Potentials defined on such segments will have a lower labelling inconsistency cost and will have less influence in the CRF.*

where $\mathcal{S}$ refers to the set of all regions or segments, and ψ_c are higher order potentials defined on them. We will now describe in detail how these potentials are defined.

3.1. Region based consistency potential

The *region consistency potential* is similar to the smoothness prior present in pairwise CRFs [3]. It favours all pixels belonging to a segment taking the same label, and as will be shown later is particularly useful in obtaining object segmentations with fine boundaries. It takes the form of a $\mathcal{P}^n$ Potts model [13]:

$$\psi_c^p(\mathbf{x}_c) = \begin{cases} 0 & \text{if } x_i = l_k, \forall i \in c, \\ \theta_p^h |c|^{\theta_\alpha} & \text{otherwise.} \end{cases} \tag{8}$$

where $|c|$ is the cardinality of the pixel set c which in our case is the number of pixels constituting superpixel c. The expression $\theta_p^h |c|^{\theta_\alpha}$ gives the label inconsistency cost, i.e. the cost added to the energy of a labelling in which different labels have been assigned to the pixels constituting the segment. The parameters θ_p^h and θ_α are learned from the training data by cross validation as described in section 4. The reader should note that this potential cannot be expressed in a pairwise CRF model.

3.2. Quality sensitive consistency potential

Not all segments obtained using unsupervised segmentation are equally good, for instance, some segments may contain multiple object classes. A region consistency potential defined over such a segment will encourage an incorrect labelling of the image. This is because the potential (8) does not take the quality or *goodness* of the segment. It assigns the same penalty for breaking 'good' segment as it assigns to 'bad' ones. This problem of the consistency potential can be overcome by defining a quality sensitive higher order potential (see figure 2). This new potential works by modulating the label inconsistency cost with a function of the quality of the segment (which is denoted by $G(c)$). Any method for estimating the segment quality can be used in our framework. A good example would be the method of [23] which uses inter and intra region similarity to measure the quality or goodness of a segment. Formally, the potential function is written as:

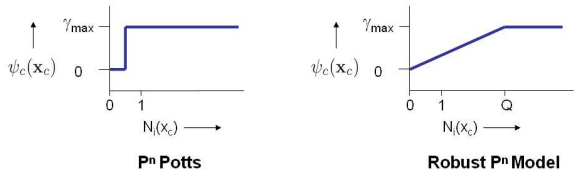


Figure 3. *Behaviour of the rigid P^n Potts potential and the Robust P^n model potential. The figure shows how the cost enforced by the two higher order potentials changes with the number of variables in the clique not taking the dominant label i.e. $N_i(\mathbf{x}_c) = min_k(|c| - n_k(\mathbf{x}_c))$.*

$$\psi_c^v(\mathbf{x}_c) = \begin{cases} 0 & \text{if } x_i = l_k, \forall i \in c, \\ |c|^{\theta_\alpha}(\theta_p^h + \theta_v^h G(c)) & \text{otherwise.} \end{cases} \tag{9}$$

For our experiments, we use the variance of the response of a unary feature evaluated on all constituent pixels of a segment to measure the quality of a segment, i.e.

$$G(c) = \exp\left(-\theta_\beta^h \frac{\| \sum_{i \in c} (f(i) - \mu)^2 \|}{|c|}\right), \tag{10}$$

where $\mu = \frac{\sum_{i \in c} f(i)}{|c|}$ and $f()$ is a function evaluated on all constituent pixels of the superpixel c. If we restrict our attention to only pairwise cliques i.e. $|c| = 2$, the variance sensitive potential becomes $\psi_c^v(x_i, x_j) =$

$$\begin{cases} 0 & \text{if } x_i = x_j, \\ |c|^{\theta_\alpha}(\theta_p^h + \theta_v^h \exp(-\theta_\beta^h \frac{\|f(i) - f(j)\|^2}{4})) & \text{otherwise.} \end{cases} \tag{11}$$

This is the same as the pairwise potential (5) commonly used in pairwise CRFs for different image labelling problems [3, 25]. Thus, the variance sensitive potential can be seen as a higher order generalization of the contrast preserving potential. The variance function response over two segmentations of an image is shown in figure 2.

3.3. Making the potentials robust

The P^n Potts model enforces label consistency very rigidly and thus might not be able to deal with inaccurate superpixels or resolve conflicts between overlapping regions of pixels. This phenomenon is illustrated in figure 4 wherein a part of the bird is merged with the 'sky' superpixel and

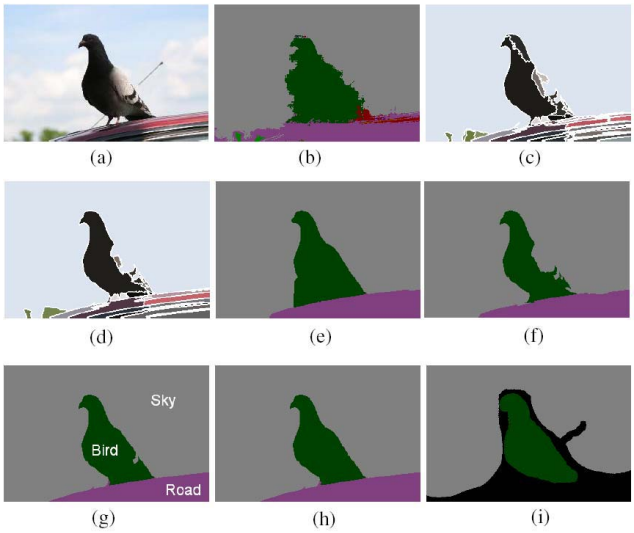


Figure 4. *Object segmentation and recognition using the Robust P^n higher order potentials (12). (a) Original Image. (b) Labelling from unary likelihood potentials from Textonboost [29]. (c) and (d) Segmentations obtained by varying the parameters of the Mean shift algorithm for unsupervised image segmentation [6]. (e) Result obtained using pairwise potential functions as described in [29]. (f) Result obtained using P^n Potts model potentials defined on the segments (or superpixels) shown in (c) and (d). These higher order potentials encourage all pixels in a superpixel to take the same label. The P^n Potts model enforces label consistency in regions very rigidly thus causing certain pixels belonging to the 'bird' to erroneously take the label 'sky' as they were included in the 'sky' superpixel. This problem can be overcome by using the Robust P^n model potentials defined in (12) which are robust and allow some variables in the clique to take different labels. (g) and (h) show results obtained by using the robust potentials with truncation parameter Q equal to $0.1|c|$ and $0.2|c|$ respectively. Here $|c|$ is equal to the size of the superpixel over which the Robust P^n model potential is defined. (i) Hand labelled segmentation from the* MSRC *dataset.*

results in an inaccurate segmentation. Intuitively, this problem can be resolved using the *Robust* higher order potentials defined as:

$$\psi_c^v(\mathbf{x}_c) = \begin{cases} N_i(\mathbf{x}_c)\frac{1}{Q}\gamma_{\max} & \text{if } N_i(\mathbf{x}_c) \leq Q, \\ \gamma_{\max} & \text{otherwise,} \end{cases} \quad (12)$$

where $N_i(\mathbf{x}_c)$ denotes the number of variables in the clique c not taking the dominant label, i.e. $N_i(\mathbf{x}_c) = \min_k(|c| - n_k(\mathbf{x}_c))$, $\gamma_{\max} = |c|^{\theta_\alpha}(\theta_p^h + \theta_v^h G(c))$, and Q is the truncation parameter which controls the rigidity of the higher order clique potential. This potential takes the form of the Robust P^n model introduced by us in [14], where we showed how energy functions composed of such potentials can be minimized using move making algorithms such as α-expansion and $\alpha\beta$-swap.

Unlike the P^n Potts model, this potential function gives rise to a cost that is a linear truncated function of the number of inconsistent variables (see figure 3). This enables the robust potential to allow some variables in the clique to take different labels. In the image shown in figure 4, the Robust P^n model potential allows some pixels of the 'sky' segment to take the label 'bird' thus producing a much better segmentation. Experiment results are shown for multiple values of the truncation parameter Q. More qualitative results can be seen in figure 7.

Figure 5. *Generating multiple segmentations. The figure shows the segmentations obtained by using different parameters in the mean-shift algorithm. The parameters used for generating the segmentation are written below it in the format (h_s, h_r), where h_s and h_r are the bandwidth parameters for the spatial and range (colour) domains.*

3.4. Generating multiple segmentations

We now explain how the set $\mathcal{S}$ of segments used for defining the higher order energy function (7) was generated. Our framework is quite flexible and can handle multiple overlapping or non-overlapping segments. The computer vision literature contains algorithms for sampling the likely segmentations of an image [30] or for generating multi-scale segmentations [27]. However, following in the footsteps of [26] we choose to generate multiple segmentations by varying the parameters of the mean shift segmentation algorithm [6]. This method belongs to the class of unsupervised segmentation algorithms which work by clustering pixels on the basis of low level image features [28, 6, 7]. They have been shown to give decent results which have proved to be useful for many applications [10, 11, 32].

The kernel used in the mean shift algorithm is defined as the product of spatial and range kernels. The spatial domain contains the (x, y) coordinates, while the range domain contains pixel colour information in LUV space. An assumption of Euclidian metric in both of them allows the use of a single bandwidth parameter for each domain, h_s for spatial and h_r for range. The segmentation results obtained using 2 different spatial $\{7, 18\}$ and 3 different range parameter values $\{6.5, 9.5, 15\}$ are shown in figure 5. It can be seen that the results do not change dramatically on small images by modifying h_s. The only difference occurs on very noisy parts of the image like trees and bushes. By increasing the range parameter h_r we can get a range of segmentations which vary from over-segmented to under-segmented. We decided to use three segmentations with parameters $(h_s, h_r) = \{(7, 6.5), (7, 9.5), (7, 15)\}$.

Figure 6. *Qualitative object segmentation and recognition results. The first column shows the original image from the Sowerby-7 dataset. Column 2 shows the result of performing inference in the pairwise* CRF *model described in section 2.1. The result obtained using the P^n Potts potential (9) is shown in column 3. The results of using the Robust P^n potential (12) is shown in column 4. The hand labelled segmentation used as ground truth is shown in column 5.*

Figure 7. *Some qualitative results. Please view in colour. First Row: Original Image. Second Row: Unary likelihood labelling from Textonboost [29]. Third Row: Result obtained using a pairwise contrast preserving smoothness potential as described in [29]. Fourth Row: Result obtained using the P^n Potts model potential [13]. Fifth Row: Results using the Robust P^n model potential (12) with truncation parameter $Q = 0.1|c|$, where $|c|$ is equal to the size of the superpixel over which the Robust P^n higher order potential is defined. Sixth Row: Hand labelled segmentations. Observe that the results obtained using the Robust P^n model are significantly better than those obtained using other methods. For instance, the leg of the sheep and bird have been accurately labelled which was missing in other results. Same can be said about the tail and leg of the dog, and the wings of the aeroplane.*

4. Experiments

This section describes our experiments. For comparative evaluation of our method we implemented the state of the art TextonBoost [29] algorithm which uses a pairwise CRF. We then augmented the CRF by adding higher order potentials defined on segments obtained from mean-shift [6].

Datasets We tested both the pairwise CRF and higher order CRF models on the MSRC-23 [29] and Sowerby-7 [9] datasets. The MSRC dataset contains 23 object classes and comprises of 591 colour images of 320×213 resolution. The Sowerby dataset contains 7 object classes and comprises of 104 colour images of 96×64 resolution. In our experiments, 50% of the images in the dataset were used for training and the remaining were used for testing.

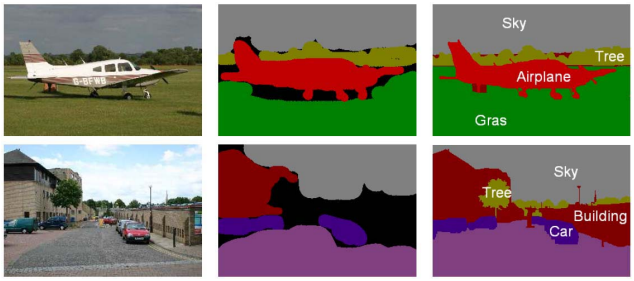


Figure 8. *Accurate hand labelled segmentations which were used as ground truth. The figure shows some images from the* MSRC *data set (column 1), the hand labelled segmentations that came with the data set (column 2), and the new segmentations hand labelled by us which were used as ground truth (column 3).*

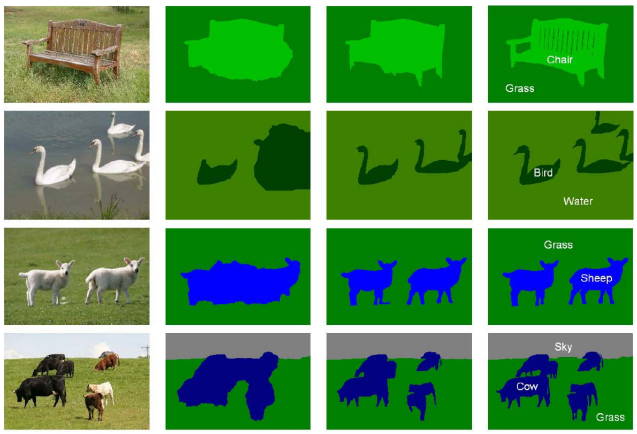
4.1. Setting CRF parameters

The optimal values for different parameters of the higher order CRF were found in a manner similar to the one used for the pairwise CRF in [29]. The model parameters were learned by minimizing the overall pixelwise classification error rate on a set of validation images - a subset of training images which were not used for training unary potentials.

A simple method for selecting parameter values is to perform cross-validation for every combination of unary, pairwise and higher order parameters within a certain discretized range. Unfortunately, the space of possible parameter values is high dimensional and doing an exhaustive search is infeasible even with very few discretization levels for each parameter. We used a heuristic to overcome this problem. First we learned the weighting between unary potentials from colour, location and Textonboost. Then we kept these weights constant and learned the optimal parameters for pairwise potentials. Pairwise and higher order potentials have similar functionality in the framework, thus learning of higher order parameters from the model with optimal unary and pairwise parameters would lead to very low weights of higher order potentials. Instead we learned optimal higher order parameters in CRF with only unary and higher order potentials and in the last step the ratio between pairwise and higher order potentials. The final trained coefficients for the MSRC dataset were $\theta_T = 0.52$, $\theta_{col} = 0.21$, $\theta_l = 0.27$, $\theta_p = 1.0$, $\theta_v = 4.5$, $\theta_\beta = 16.0$, $\theta_\alpha = 0.8$, $\theta_p^h = 0.2$, $\theta_v^h = 0.5$, $\theta_\beta^h = 12.0$. Parameter learning for higher order CRFs is an ongoing topic of research.

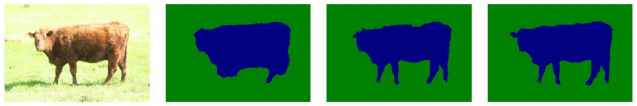
4.2. Quantitative Segmentation Results

The results of our experiments show that integration of higher order P^n Potts model potentials quantitatively and



(a) Original (b) Pairwise CRF (c) Robust Pn Model (d) Ground Truth

Figure 9. *Qualitative results of our method. (a) Original Images. (b) Segmentation result obtained using the pairwise CRF (explained in section 2.1). (c) Results obtained by incorporating the Robust P^n higher order potential (12) defined on segments. (d) Hand labelled result used as ground truth.*



(a) (b) (c) (d)

Figure 10. *The relationship between qualitative and quantitative results. (a) Original Image. (b) Segmentation result obtained using the pairwise CRF (explained in section 2.1). Overall pixelwise accuracy for the result is 95.8%. (c) Results obtained by incorporating the Robust P^n higher order potential (12) defined on segments. Overall pixelwise accuracy for this result is 98.7%. (d) Hand labelled result used as ground truth. It can be seen that even a small difference in the pixelwise accuracy can produce a massive difference in the quality of the segmentation.*

qualitatively improves segmentation results. The use of the robust potentials lead to further improvements (see figure 4,6, 7 and 9). Inference on both the pairwise and higher order CRF model was performed using the graph cut based expansion move algorithm [4, 14]. The optimal expansion moves for the energy functions containing the Robust P^n potential (12) were computed using the method of [14].

Ground Truth The hand labelled 'ground truth' images that come with the MSRC-23 data set are quite rough. In fact qualitatively they always looked worse than the results obtained from our method. The hand labelled images suffer from another drawback. A significant numbers of pixels in these images have not been assigned any label. These unlabelled pixels generally occur at object boundaries and are critical in evaluating the accuracy of a segmentation algorithm. It should be noted that obtaining an accurate and fine segmentation of the object is important for many tasks in computer vision.

In order to get a good estimate of our algorithm's accuracy, we generated accurate segmentations which preserved the fine object boundaries present in the image. Generating these segmentations is quite time consuming. It takes between 15-60 minutes to hand label one image. We hand labelled 27 images from the MSRC data set. Figure 8 shows the original hand labelled images of the MSRC data set and the new segmentations manually labelled by us which were used as ground truth.

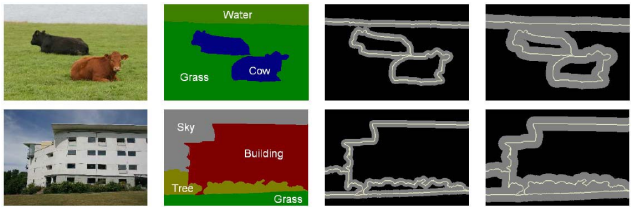


Image (MSRC) Ground Truth Trimap (8 Pixel) Trimap (16 Pixel)

Figure 11. *Boundary accuracy evaluation using trimap segmentations. The first column shows some images from the MSRC dataset [29]. The ground truth segmentations of these image are shown in column 2. Column 3 shows the trimaps used for measuring the pixel labelling accuracy. The evaluation region is coloured gray and was generated by taking an 8 pixel band surrounding the boundaries of the objects. The corresponding trimaps for an evaluation band width of 16 pixels is shown in column 4.*

Evaluating Accuracy Typically the performance of a segmentation algorithm is measured by counting the total number of mislabelled pixels in the image. We believe this measure is not appropriate for measuring the segmentation accuracy if the user is interested in obtaining accurate segmentations as alpha mattes with fine object boundaries. As only a small fraction of image pixels lie on the boundary of an object, a large qualitative improvement in the quality of the segmentation will result in only a small increase in the percentage pixel-wise accuracy. This phenomenon is illustrated in figure 10.

With this fact in mind, we evaluate the quality of a segmentation by counting the number of pixels misclassified in the region surrounding the actual object boundary and not over the entire image. The error was computed for different widths of the evaluation region. The evaluation regions for some images from the MSRC dataset are shown in figure 11. The accuracy of different segmentation methods is plotted in the graph shown in figure 12.

5. Summary

In this paper we proposed a novel framework for labelling problems which is capable of utilizing features based on sets of pixels in a principled manner. We tested this approach on the problem of multi-class object segmentation and recognition. Our experiments showed that incorporation of P^n Potts and Robust P^n model type potential functions (defined on segments) in the conditional random field model for object segmentation dramatically improved results around object boundaries. We believe this method is generic and can be used to solve many other labelling problems. In the future we would like to investigate the use of more sophisticated higher order po-

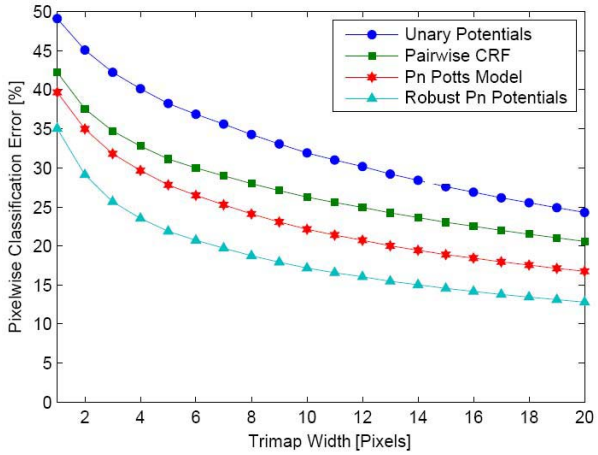


Figure 12. *Pixelwise classification error in our results. The graph shows how the overall pixelwise classification error varies as we increase the width of the evaluation region.*

tentials based on the shape and appearance of image segments. We believe that such potentials would be more discriminative and will result in even better performance.

References

[1] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*, pages I: 428–441, 2004. 3

[2] E. Borenstein and J. Malik. Shape guided object segmentation. In *CVPR (1)*, pages 969–976, 2006. 2

[3] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, pages I: 105–112, 2001. 3, 4

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001. 2, 3, 7

[5] M. Bray, P. Kohli, and P. H. S. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *ECCV (2)*, pages 642–655, 2006. 3

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002. 1, 5, 6

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 5

[8] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *CVPR (2)*, pages 695–702, 2004. 2

[9] X. He, R. S. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *ECCV (1)*, pages 338–351, 2006. 1, 3, 6

[10] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Trans. Graph.*, 24(3):577–584, 2005. 1, 5

[11] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, pages 654–661, 2005. 1, 5

[12] R. Huang, V. Pavlovic, and D. Metaxas. A graphical model framework for coupling mrfs and deformable models. In *CVPR*, volume II, pages 739–746, 2004. 2

[13] P. Kohli, M. Kumar, and P. Torr. P^3 and beyond: Solving energies with higher order cliques. In *CVPR*, 2007. 2, 4, 6

[14] P. Kohli, L. Ladicky, and P. Torr. Graph cuts for minimizing robust higher order potentials. *Technical report, Oxford Brookes University*, 2008. 1, 2, 5, 7

[15] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1568–1583, 2006. 3

[16] M. Kumar, P. Torr, and A. Zisserman. Obj cut. In *CVPR (1)*, pages 18–25, 2005. 2, 3

[17] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001. 3

[18] X. Lan, S. Roth, D. P. Huttenlocher, and M. J. Black. Efficient belief propagation with learned higher-order markov random fields. In *ECCV (2)*, pages 269–282, 2006. 2

[19] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In *ECCV*, pages 581–594, 2006. 2

[20] R. Paget and I. D. Longstaff. Texture synthesis via a noncausal nonparametric multiscale markov random field. *IEEE Transactions on Image Processing*, 7(6):925–931, 1998. 2

[21] B. Potetz. Efficient belief propagation for vision using linear constraint nodes. In *CVPR*, 2007. 2

[22] A. Rabinovich, S. Belongie, T. Lange, and J. M. Buhmann. Model order selection and cue combination for image segmentation. In *CVPR (1)*, pages 1130–1137, 2006. 1

[23] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, pages 10–17, 2003. 4

[24] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *CVPR*, pages 860–867, 2005. 2

[25] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, pages 309–314, 2004. 3, 4

[26] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR (2)*, pages 1605–1614, 2006. 1, 5

[27] E. Sharon, A. Brandt, and R. Basri. Segmentation and boundary detection using multiscale intensity measurements. In *CVPR (1)*, pages 469–476, 2001. 5

[28] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000. 5

[29] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV (1)*, pages 1–15, 2006. 1, 2, 3, 5, 6, 7

[30] Z. Tu and S. C. Zhu. Image segmentation by data-driven markov chain monte carlo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):657–673, 2002. 5

[31] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. Map estimation via agreement on trees: message-passing and linear programming. *IEEE Transactions on Information Theory*, 51(11):3697–3717, 2005. 3

[32] J. Wang, P. Bhat, A. Colburn, M. Agrawala, and M. F. Cohen. Interactive video cutout. *ACM Trans. Graph.*, 24(3):585–594, 2005. 5