# Associative Hierarchical CRFs for Object Class Image Segmentation

L'ubor Ladický Chris Russell Pushmeet Kohli
Oxford Brookes University Microsoft Research Cambridge

Philip H.S. Torr
Oxford Brookes University
http://cms.brookes.ac.uk/research/visiongroup/,
http://research.microsoft.com/en-us/um/people/pkohli/

## Abstract

*Most methods for object class segmentation are formulated as a labelling problem over a single choice of quantisation of an image space - pixels, segments or group of segments. It is well known that each quantisation has its fair share of pros and cons; and the existence of a common optimal quantisation level suitable for all object categories is highly unlikely. Motivated by this observation, we propose a hierarchical random field model, that allows integration of features computed at different levels of the quantisation hierarchy. MAP inference in this model can be performed efficiently using powerful graph cut based move making algorithms. Our framework generalises much of the previous work based on pixels or segments. We evaluate its efficiency on some of the most challenging data-sets for object class segmentation, and show it obtains state-of-the-art results.*

## 1. Introduction

Object class based image segmentation is one of the most challenging and important problems in computer vision. It aims to assign an object label to each pixel of a given image; and can be seen as a generalisation of the object recognition and localisation tasks. Over the last few years many different methods have been proposed for this problem, which can be broadly categorised on the basis of their choice of the quantisation (partitioning) of the image space[1]. Some methods are formulated in terms of pixels [26] (representing the finest quantisation), others used segments [1, 10, 32], groups of segments [19], or intersections of multiple segmentations [18], while some have gone to the extreme of looking at the whole image in order to reason about object segmentation [17]. We present a model together with an efficient optimisation technique that contains the above mentioned previous methods as special cases, thus allowing for the use of holistic models that integrate the strengths of these different approaches.

**Pixel vs Segments** Each choice of image quantisation comes with its share of advantages and disadvantages. Pixels might be considered the most obvious choice of quantisation. However, pixels by themselves contain a limited amount of information. The colour and intensity of a lone pixel is often not enough to determine its correct object label. Ren and Malik's [20] remark that *'pixels are not natural entities; they are merely a consequence of the discrete representation of images'* captures some of problems of pixel-based representation.

The last few years have seen a proliferation of unsupervised segmentation methods [5, 8, 24], that perform an initial *a priori* segmentation of the image, applied to object segmentation [1, 10, 32, 11, 22, 32], and elsewhere [13, 27]. These rely upon an initial quantisation over the image space, typically based upon a segmentation of pixels based upon spatial location and colour/texture distribution.

Based upon the assumption that the quantisation is correct a segment based *conditional random field* (CRF) is defined over the image, and inference is performed to estimate the dominant label of each segment. This quantisation of the image allows the computation of powerful region-based features which are partially invariant to scale [31].

**Use of Multiple Quantisations** Segment based methods work under the assumption that some segments share boundaries with objects in the image. This is not always the case, and this assumption may result in dramatic errors in the labelling (see figure 1). A number of techniques have been proposed to overcome errors in the image quantisation. Rabinovich *et al.* [19] suggested finding the most stable segmentation from a large collection of multiple segmentations in the hope that these would be more consistent with object boundaries. Larlus and Juri [17] proposed an approach to the problem driven by object detection. In their algorithm, rectangular regions are detected using a bag-of-words model based upon affine invariant features. These rectangles are refined using graph cuts to extract boundaries in a grab-cut [21] like approach. Such approaches face difficulty in dealing with cluttered images, in which multiple object classes intersect. Pantofaru *et al.* [18] observed that although segments may not be consistent with object boundaries, the segmentation map formed by taking the intersections of multiple segmentations often is. They proposed finding the most probable labelling of intersections of segments based upon the features of their parent segments. This scheme effectively reduces the image quantisation level. It results in more consistent segments but with

[1] We use the phrase "quantise the image" as opposed to "segment the image" in order to emphasise that a 'quantum' of the image space need not just be a collection of pixels. It could represent a sub-pixel division of the image space.

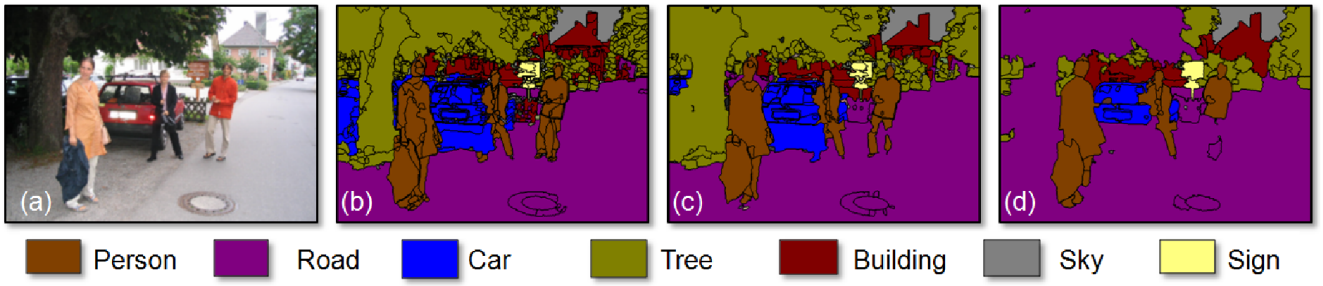



Figure 1. *Effect of image quantisation on object segmentation. (a) Original image. (b)-(d) Object class segmentations with different image quantisations. (b), (c) and (d) use three different unsupervised segmentations of the image, in this case mean-shift with different choices of kernel, to divide the image into segments. Each segment is assigned the label of the dominant object present in it. It can be seen that quantisation (b) is the best for tree, road, and car. However, quantisation (d) is better for the left person and the sign board.*

a loss in the information content/discriminative power associated with each segment.

Another interesting method, and one closely related to ours was proposed by Kohli *et al.* [15]. By formulating the labelling problem as a CRF defined over pixels, they were able to recover from misleading segments which spanned multiple object classes. Further, they were able to encouraged individual pixels within a single segment to share the same label, by defining higher order potentials (functions defined over cliques of size greater than 2) that penalised inconsistent labellings of segment.Their method can be understood as a relaxation of the hard constraint of previous methods, that the image labelling must follow the quantisation of the image space, to a softer constraint in which a penalty is paid for non-conformance.

Given the dependence of previous methods on the image partitioning (quantisation), the key question to be asked is: *What is the correct quantisation of an image and how can we find it?* This is a difficult question to answer. As we explore the quantisation hierarchy from coarse to fine, we observe that while larger segments are perceptually more meaningful and easier to label correctly, they are less likely to lie inside a single object. Indeed pragmatically, it appears that the finding of an ideal quantisation may not be possible, and that segmentation of different objects in the image may require different quantisations (see figure 1).

In this paper we propose a novel hierarchical CRF formulation of object class segmentation that allows us to unify multiple disparate quantisations of the image space, avoiding the need to make a decision of which is most appropriate. It allows for the integration of features derived from different quantisation levels (pixel, segment, and segment union/intersection). We will demonstrate how many of the state-of-the-art methods based on different fixed image quantisations can be seen as special cases of our model.

Inferring the Maximum a Posteriori solution in this framework involves the minimisation of a higher order function defined over several thousand random variables, as explained in section 2. We show that the solutions of such difficult function minimisation problems can be efficiently computed using graph-cut [3] based move-making algorithms. However, the contribution of this paper is not limited to the application of the novel hierarchical CRF framework to object class segmentation. We also propose new sophisticated potentials defined over the different levels of the quantisation hierarchy, and evaluate the efficacy of our framework on some of the most challenging data-sets for object class segmentation, and show that it outperforms state of the art methods based on individual image quantisation levels. We believe this is because: *(i)* Our methods generalise these previous methods allowing them to be represented as particular parameter choices of our hierarchical model. *(ii)* We go beyond these models by being able to use multiple hierarchies of segmentation simultaneously. *(iii)* In contrast to many previous methods that do not define any sort of cost function, or likelihood, we cleanly formulate the CRF energy of our model and show it can be minimised.

**Hierarchical Models and Context** The use of context has been well documented for object recognition and segmentation. It is particularly useful in overcoming ambiguities caused by limited evidence: this often occurs in object recognition where we frequently encounter objects at small scales or low resolution images [14]. Classical Markov and Conditional Random Field models exploit context in a local manner by encouraging adjacent pixels or segments to take the same label. To encode context at different scales Zhu *et al.* [33] introduced the hierarchical image model (HIM) built of rectangular regions with parent-child dependencies. This model captures large-distance dependencies and is solved efficiently using dynamic programming. However, it supports neither multiple hierarchies, nor dependencies between variables at the same level. To encode semantic context and to combine top-down and bottom-up approaches Tu *et al.* [30] proposed a framework with which they showed that the use of object specific knowledge helps to disambiguate low-level segmentation cues.

Our hierarchical CRF model uses a novel formulation that allows context to be incorporated at multiple levels of multiple quantisation, something not previously possible. As we will explain in section 4 it leads to improved segmentation results, while keeping the inference tractable.

## 2. Random Fields for Labelling Problems

This section introduces the pixel-based CRF used for formulating the object class segmentation problem. This formulation contains one discrete random variable per image pixel, each of which may take a value from the set of labels

$\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$. We use $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ to denote the set of random variables corresponding to the image pixels $i \in \mathcal{V} = \{1, 2, \ldots, N\}$. The neighbourhood 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 $\mathbf{L} = \mathcal{L}^N$.

The posterior distribution $\Pr(\mathbf{x}|\mathbf{D})$ over the labellings of the CRF is a *Gibbs* distribution and can be written as: $\Pr(\mathbf{x}|\mathbf{D}) = \frac{1}{Z}\exp(-\sum_{c\in\mathcal{C}}\psi_c(\mathbf{x}_c))$, where $Z$ is a normalising constant called the *partition function*, and $\mathcal{C}$ is the set of all cliques [16]. The term $\psi_c(\mathbf{x}_c)$ is known as the potential function of the clique $c \subset \mathcal{V}$ where $\mathbf{x}_c = \{x_i : i \in c\}$. The corresponding Gibbs energy is given by

$$E(\mathbf{x}) = -\log \Pr(\mathbf{x}|\mathbf{D}) - \log Z = \sum_{c\in\mathcal{C}} \psi_c(\mathbf{x}_c). \quad (1)$$

The most probable or Maximum a Posteriori (MAP) labelling $\mathbf{x}^*$ of the random field is defined as

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathbf{L}} \Pr(\mathbf{x}|\mathbf{D}) = \arg\min_{\mathbf{x}\in\mathbf{L}} E(\mathbf{x}). \quad (2)$$

**Pairwise CRFs** Most pixel labelling problem in vision are formulated as a pairwise CRF whose energy can be written as the sum of unary and pairwise potentials as

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

The unary potentials $\psi_i(x_i)$ of the CRF are defined as the negative log likelihood of variable $X_i$ taking label $x_i$, while the pairwise potential encode a smoothness prior which encourages neighbouring pixels in the image to take the same label, resulting in a *shrinkage bias* [15].

The pairwise CRF formulation suffers from a number of problems stemming from its inability to express high-level dependencies between pixels. Despite these limitations, it is widely used and very effective. Shotton *et al.* [26] applied the pairwise CRF to the object class segmentation problem. They defined the unary likelihoods potentials using the result of a boosted classifier over a region about each pixel, that they called *TextonBoost* and were able to obtain good results.

**The Robust $P^N$ model** The pairwise CRF formulation of [26] was extended by [15] with the incorporation of robust higher order potentials defined over segments. Their formulation was based upon the observation that pixels lying within the same segment are more likely to take the same label. The energy of the higher order CRF proposed by [15] was of the form

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}} \psi_i(x_i) + \sum_{i\in\mathcal{V}, j\in\mathcal{N}_l} \psi_{ij}(x_i, x_j) + \sum_{c\in\mathcal{S}} \psi_c^h(\mathbf{x}_c), \quad (4)$$

where $\mathcal{S}$ is a set of cliques (or segments), and $\psi_c$ are higher order potentials defined over them. Their higher order potentials took the form of Robust $P^N$ model defined as

$$\psi_c^h(\mathbf{x}_c) = \min_{l\in\mathcal{L}}(\gamma_c^{\max}, \gamma_c^l + k_c^l N_c^l), \quad (5)$$

satisfying $\gamma_c^l \leq \gamma_c^{\max}, \forall l \in \mathcal{L}$, where $N_c^l = \sum_{i\in c}\delta(x_i \neq l)$ is the number of inconsistent pixels with the label $l$.

The potential takes cost $\gamma_c^l$ if all pixels in the segment take the label $l$. Each inconsistent pixel is penalised with a cost $k_c^l$. The maximum cost of the potential is truncated to $\gamma_c^{max}$. By setting $\gamma_c^l = 0\ \forall l \in \mathcal{L}$ this potential penalises inconsistent segments and thus encourages label consistency in segments. The weighted version of this potential is

$$\psi_c^h(\mathbf{x}_c) = \min_{l\in\mathcal{L}}(\gamma_c^{\max}, \gamma_c^l + \sum_{i\in c} w_i k_c^l \delta(x_i \neq l)), \quad (6)$$

where $w_i$ is the weight of the variable $x_i$.

This framework enabled the integration of multiple quantisations of the image space in a principled manner. However unlike our work, their choice of potential was independent of the choice of label and only encouraged pixels within the same segment to take the same label. Similarly, their model is unable to encode the conditional dependencies between segments. These potentials greatly increase the expressiveness of our model, as detailed in section 3.

**$P^N$-Based Hierarchical CRFs** As shown in [23], the higher-order $P^N$ potentials of (6) are equivalent to the minimisation of a pairwise graph defined over the same clique $\mathbf{X}_c$ and a single auxiliary variable $y_c$, that takes values from an extended label set $\mathcal{L}^E = \mathcal{L} \cup \{L_F\}$. The cost function over $\mathbf{X}_c \cup \{y_c\}$takes the form

$$\psi_c^p(\mathbf{x}_c, y_c) = \phi_c(y_c) + \sum_{i\in c} \phi_c(y_c, x_i). \quad (7)$$

where the unary potential over $Y$, $\phi_c(y_c)$ associates the cost $\gamma_c^l$ with $y_c$ taking a label in $\mathcal{L}$, and $\gamma_c^{\max}$ with $y_c$ taking the *free* label $L_F$. The pairwise potentials $\phi_c(y_c, x_i)$ are defined

$$\phi_c(y_c, x_i) = \begin{cases} 0 & \text{if } y_c = L_F \text{ or } y_c = x_i \\ w_i k_c^l & \text{otherwise, where } l = x_i. \end{cases} \quad (8)$$

Then

$$\psi_c^h(\mathbf{x}_c) = \min_y \psi_c^p(\mathbf{x}_c, y_c). \quad (9)$$

By ensuring that the pairwise edges between the auxiliary variable and its children satisfy the constraint $\sum_i w_i k_c^l \geq 2\phi_c(l), \forall l \in \mathcal{L}$, we can guarantee that the labels of these auxiliary variable carry a clear semantic meaning. If this constraint is satisfied an auxiliary variable may takes state $l \in \mathcal{L}$ in a minimal cost labelling, if and only if, the weighted majority of its child nodes take state $l$. State $L_F$

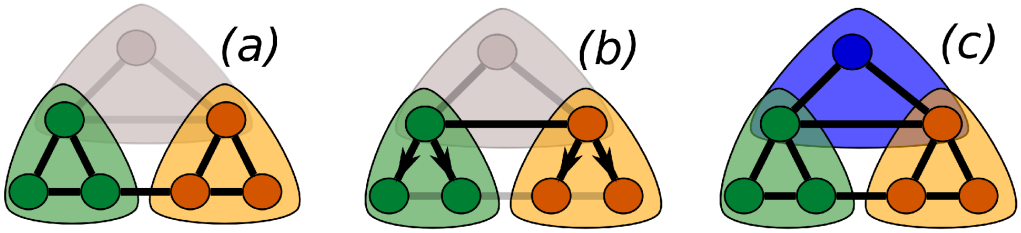



Figure 2. Existing models as special cases of our hierarchical model. *The lowest layer of the image represents the pixel layer, the middle layer potentials defined over super-pixels or segments, and the third layer represents our hierarchical terms.* (a) *shows the relationships permitted in a pixel-based* CRF *with Robust* $P^N$ *potentials.* (b) *shows relationships contained within a super-pixel-based* CRF *(the directed edges indicate the one way dependence between the labellings of pixels and super-pixels).* (c) *Our hierarchical* CRF. *See section 3.*

indicates a heterogeneous labelling of a segment in which no label holds a significant majority. We now extend the model to include pairwise dependencies between auxiliary variables

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}} \psi_i(x_i) + \sum_{i\in\mathcal{V}, j\in\mathcal{N}_l} \psi_{ij}(x_i, x_j) + \min_{\mathbf{y}} \Big( \sum_{c\in\mathcal{S}} \psi_c^p(x, y_c) + \sum_{c,d\in\mathcal{S}} \psi_{cd}(y_c, y_d) \Big). \quad (10)$$

These pairwise terms can be understood as encouraging consistency between neighbouring cliques. This framework can be further generalised to a hierarchical model [23] where the connection between layers take the form of (7) and the weights $\phi_c(y_c, \underline{x})$ are proportional to the number of the pixels in the *"base layer"* belonging to the clique $c$.

The energy of our new hierarchical model is of the form

$$E^{(0)}(\mathbf{x}) = \sum_{i\in\mathcal{S}^{(0)}} \psi_i(x_i^{(0)}) + \sum_{ij\in\mathcal{N}^{(0)}} \psi_{ij}(x_i^{(0)}, x_j^{(0)}) + \min_{\mathbf{x}^{(1)}} E^{(1)}(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}), \quad (11)$$

where $E^{(1)}(\mathbf{x}^{(0)}, \mathbf{x}^{(1)})$ is recursively defined as: $E^{(n)}(\mathbf{x}^{(n-1)}, \mathbf{x}^{(n)})$

$$= \sum_{c\in\mathcal{S}^{(n)}} \psi_c^p(\mathbf{x}_c^{(n-1)}, x_c^{(n)}) + \sum_{cd\in\mathcal{N}^{(n)}} \psi_{cd}(x_c^{(n)}, x_d^{(n)}) + \min_{\mathbf{x}^{(n+1)}} E^{(n+1)}(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}). \quad (12)$$

Where $\mathbf{x}^{(0)}$ refers to the state of the base level, and $\mathbf{x}^{(n)}|n \geq 1$ the state of auxiliary variables. Under certain reasonable conditions [23], the auxiliary variables retain their semantic interpretation and this energy can be solved with graph-cut based move making algorithms [4].

## 3. Relation to Previous Models

In this section, we draw comparisons with the current state of the art models for object segmentation [10, 18, 19, 32] and show that at certain choices of the parameters of our model, these methods fall out as special cases (illustrated in figure 2). Thus, our method not only generalise the standard pairwise CRF formulations over pixels, but also the previous work based on super-pixels and (as we shall see) provides a global optimisation framework allowing us to combine features at different quantisation levels.

We will now show that our model is a generalisation of two classes of pre-existing model: *(i)* CRFs based upon disjoint segments [1, 10, 32] (see section 1 and figure 2(b)), and *(ii)* CRFs based upon the intersection of segments [18]. energy of this model is given in (10), and further assume that there are no unary or pairwise potentials defined over individual pixels.

**Equivalence to CRFs based on Segments** In this case, $c \in \mathcal{S}$ are disjoint (non-overlapping)[2]. To insure that $y_c \neq L_F, \forall c \in \mathcal{C}$, we assign a high value to $\gamma_c^{\max}, \forall c \in \mathcal{C}$. As only the potential $\psi^p(\mathbf{x}_c, y_c)$ acts upon $x_i : i \in c$, all pixels in $c$ will take the same label. In this case, the optimal labelling will always be *segment consistent* (i.e. the labelling of pixels within any segment is homogeneous) and the potential $\psi_c^p(\mathbf{x}_c, y_c)$ can now be considered as a unary potential over the auxiliary (segment) variable $y_c$. This allows us to rewrite (10) as:

$$E(\mathbf{y}) = \sum_{c\in\mathcal{S}} \psi_c(y_c) + \sum_{cd\in\mathcal{N}^{(1)}} \psi_{cd}(y_c, y_d) \quad (13)$$

which is exactly the same as the cost associated with the pairwise CRF defined over segments with $\psi_c(y_c = l) = \gamma_c^l$ as the unary cost and $\psi_{cd}$ as the pairwise cost for each segment. In this case, our model becomes equivalent to the pairwise CRF models defined over segments [1, 10, 19, 32].

**Equivalence to Models of Segment Intersections** The model is defined as above, but allowed to contain multiple overlapping segmentations. If we set $w_i k_c^l = \gamma_c^{\max}, \forall i \in \mathcal{V}, l \in \mathcal{L}, c \in \mathcal{S}$, then $y_c \neq L_F$ only if $x_i = y_c, \forall i \in c$. In this case, only the potentials $\sum_{c\ni i} \psi_c^p(\mathbf{x}_c, y_c)$ act on $x_i$.

Consider a pair of pixels $i, j$ that lie in the same intersection of segments *i.e.* $\{c \in S : c \ni i\} = \{c \in S : c \ni j\}$. Then, in a minimal labelling, either $\exists y_c = x_i$, and hence $x_j = y_c = x_i$, or $\forall c \ni i : y_c = L_F$. In the second case there are no constraints acting on $x_i$ or $x_j$, and a minimal cost labelling can be chosen such that $x_i = x_j$.

Consequentially, there is always a minimal cost labelling consistent with respect to the intersection of segments, in this sense our model is equivalent to that proposed in [18].

## 4. Hierarchical CRF for Object Segmentation

Having described the definition and intuition behind the $P^N$-based hierarchical CRF framework, in this section we

[2] This is equivalent to the case where only one particular quantisation of the image space is considered.

describe the set of potentials we use in the object-class segmentation problem. This set includes unary potentials for both pixels and segments, pairwise potentials between pixels and between segments and connective potentials between pixels and their containing segments.

**Robustness to Misleading Segmentations** As discussed before, the quantisation of image space obtained using unsupervised segmentation algorithms may be misleading - segments may contain multiple object classes. Assigning the same label to all pixels of such segments will result in an incorrect labelling. This problem can be overcome by using segment quality measures proposed by [19, 20] which can be used to distinguish the *good* segments from *misleading* ones. These measures can be seamlessly integrated in our hierarchical framework by modulating the strength of the potentials defined over segments. Formally, this is achieved by modifying the potentials $\psi_c^h(\mathbf{x}_c, y_c)$ according to a quality sensitive measure $Q(c)$ for any segment $c$.

In the previous section we decomposed the energy (12) into a set of potentials $\psi_c(\mathbf{x}_c)$. In this section we will decompose them further, writing $\psi_c(\mathbf{x}_c) = \lambda_c \xi_c(\mathbf{x}_c)$, where $\xi_c$ is a feature based potential over $c$ and $\lambda_c$ its weight. Initially we will discuss the learning of potentials $\xi_c(\mathbf{x}_c)$, and later discuss the learning of the weights $\lambda_c$

**Potentials for Object Class Segmentation** For our application we used potentials defined over a three levels hierarchy. We refer to elements of each layer as pixels, segments and super-segments respectively.

The unary potentials at the pixel level are computed using a boosted dense feature classifier (described below), while the pairwise terms $\psi_{ij}(\cdot)$ take the form of the classical contrast sensitive potentials. These encourage neighbouring pixels in the image (having a similar colour) to take the same label. We refer the reader to [2, 21, 26] for details.

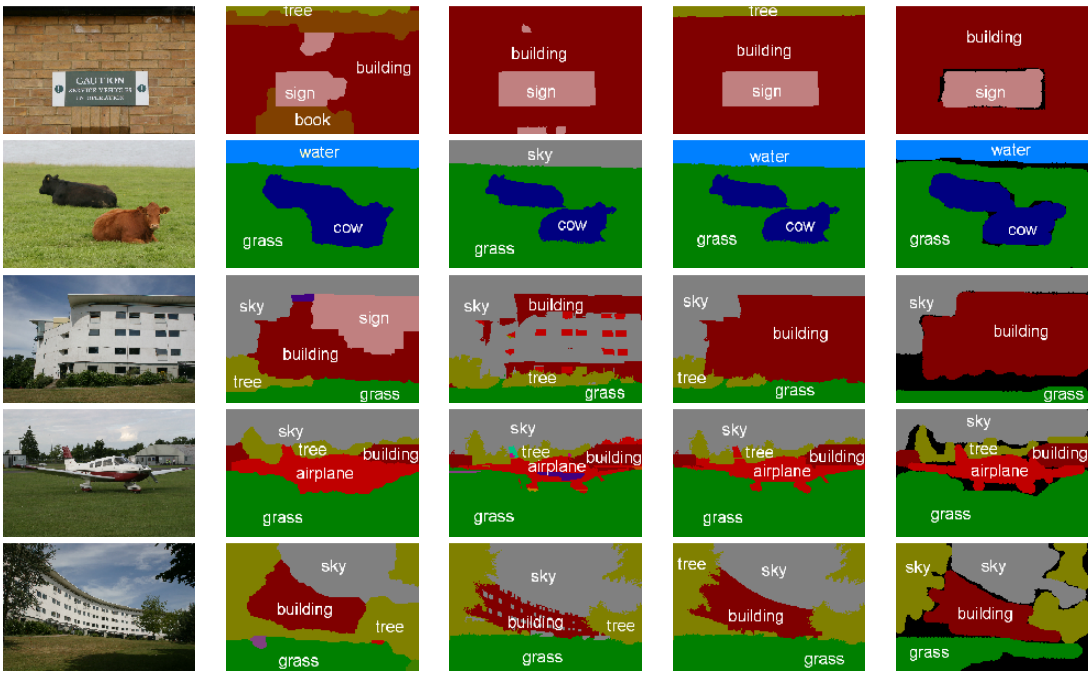
Unsupervised segments are initially found using multiple applications of a fine scale mean-shift algorithm [5]. The pixels contained within such a segment, are typically of uniform colour, and often belong to the same object class. Consequentially, they contain little novel local information, but are strong predictors of consistency. As such, the unary potentials we learn at this level are uniform, due to the lack of unique features, however as they are strongly indicative of local consistency, the penalty associated with breaking them is high. To encourage neighbouring segments with similar texture to take the same label, we used pairwise potentials based on the Euclidean distance of normalised histograms of colour ($\xi_{cd}(y_c, y_d)$) between corresponding auxiliary variables.

"Super-segments" are based upon a coarse mean-shift segmentation, performed over the result of the previous segmentations. These super-segments contain significantly more internal information than their smaller children. To take advantage of this, we propose unary segment potential based on the histograms of features (described below). This potential can be also be used in the segment-based CRF approaches as a unary potential.

**Unary Potentials from Dense Features** This unary potential is derived from *TextonBoost* [26], and allows us to perform texture based segmentation, at the pixel level, within the same framework. The features used for constructing these potentials are computed on every pixel of the image which is why we call them *dense*. TextonBoost estimates the probability of a pixel taking a certain label by boosting weak classifiers based on a set of shape filter responses. The shape filters are defined by a [texton $t$, rectangular region $r$] pair and their feature response $v_{[t,r]}(i)$ for given point $i$ is the number of textons $t$ in the region $r$ placed relative to the point $i$. Corresponding weak classifiers are comparisons of shape filter response to thresholds. The most discriminative shape filters are found using multi-class Gentle Ada-Boost [29].

We observed that textons were unable to discriminate between some classes of similar textures. This motivated us to extend the *TextonBoost* framework by boosting classifiers defined on multiple dense features (such as colour, textons, histograms of oriented gradients (HOG) [6], and pixel location) together. The dense-feature shape filters are defined by triplets [feature type $f$, feature cluster $t$, rectangular region $r$] and their feature response $v^f_{[t,r]}(i)$ for given point $i$ is the number of features of type $f$ belonging to cluster $t$ in the region $r$ placed relative to the point $i$. The pool of weak classifiers contains a comparisons of responses of dense-feature shape filters against a set of thresholds $\theta$. See [26] for further details of the procedure. Our results show that the boosting of multiple features together results in a significant improvement of the performance (note the improvement from the 72% of [26] to 81% of our similar pixel-based CRF in figure 4). Further improvements were achieved using exponentially instead of linearly growing thresholds and Gaussian instead of uniform distribution of rectangles around the point. The potential is incorporated into the framework in the standard way as a negative log likelihood.

**Histogram-based Segment Unary Potentials** We now explain the unary potential defined over segments and super-segments. For many classification and recognition problems, the distributions of dense feature responses are more discriminative than any feature alone. For instance, the sky can be either 'black' (night) or 'blue' (day), but is never 'half-black' and 'half-blue'. This consistency in the colour of object instances can be used as a region based feature for improving object segmentation results. The unary potential of an auxiliary variable representing a segment is learnt (using the normalised histograms of multiple clustered dense features) using multi-class Gentle AdaBoost [29], where the pool of weak classifiers is a set of




Original Image | Pixel-based CRF | Segment-based CRF | Hierarchical CRF | Ground Truth

Figure 3. *Qualitative, results on the MSRC-21 data set, comparing non-hierarchical(i.e. pairwise models) approaches defined over pixels (similar to TextonBoost [26]) or segments (similar to [32, 18, 22] described in section 3) against our hierarchical model. Regions marked black in the hand-labelled ground truth image are unlabelled.*

| | Global | Average | Building | Grass | Tree | Cow | Sheep | Sky | Aeroplane | Water | Face | Car | Bicycle | Flower | Sign | Bird | Book | Chair | Road | Cat | Dog | Body | Boat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [25] | 72 | 67 | 49 | 88 | 79 | **97** | **97** | 78 | **82** | 54 | **87** | 74 | 72 | 74 | 36 | 24 | **93** | **51** | 78 | 75 | 35 | **66** | 18 |
| [26] | 72 | 58 | 62 | **98** | 86 | 58 | 50 | 83 | 60 | 53 | 74 | 63 | 75 | 63 | 35 | 19 | 92 | 15 | 86 | 54 | 19 | 62 | 07 |
| [1] | 70 | 55 | 68 | 94 | 84 | 37 | 55 | 68 | 52 | 71 | 47 | 52 | 85 | 69 | 54 | 05 | 85 | 21 | 66 | 16 | 49 | 44 | 32 |
| [32] | 75 | 62 | 63 | 98 | 89 | 66 | 54 | 86 | 63 | 71 | 83 | 71 | 79 | 71 | 38 | 23 | 88 | 23 | 88 | 33 | 34 | 43 | **32** |
| Pixel-based CRF | 81 | 72 | 73 | 92 | 85 | 75 | 78 | 92 | 75 | 76 | 86 | 79 | **87** | 96 | **95** | **31** | 81 | 34 | 84 | **53** | 61 | 60 | 15 |
| Robust $P^N$ CRF | 83 | 73 | 74 | 92 | 86 | 75 | 83 | 94 | 75 | 83 | 86 | 85 | 84 | 95 | 94 | 30 | 86 | 35 | 87 | **53** | **73** | 63 | 16 |
| Segment-based CRF | 75 | 60 | 64 | 95 | 78 | 53 | 86 | **99** | 71 | 75 | 70 | 71 | 52 | 72 | 81 | 20 | 58 | 20 | 89 | 26 | 42 | 40 | 05 |
| Hierarchical CRF | **86** | **75** | **80** | 96 | 86 | 74 | 87 | **99** | 74 | **87** | 86 | **87** | 82 | **97** | **95** | 30 | 86 | 31 | **95** | 51 | 69 | **66** | 09 |

Figure 4. *Quantitative results on the MSRC data set. The table shows % pixel accuracy $N_{ii}/\sum_j N_{ij}$ for different object classes. 'Global' refers to the overall error $\frac{\sum_{i\in\mathcal{L}} N_{ii}}{\sum_{i,j\in\mathcal{L}} N_{ij}}$, while 'average' is $\sum_{i\in\mathcal{L}} \frac{N_{ii}}{|\mathcal{L}|\sum_{j\in\mathcal{L}} N_{ij}}$. $N_{ij}$ refers to the number of pixels of label $i$ labelled $j$.*

triplets $[f, t, \theta]$. Here $f$ is the normalised histogram of the feature set, $t$ is the cluster index, and $\theta$ a threshold. Aside from a larger set of features being considered, the selection and learning procedure is identical to [26].

The segment potential is incorporated into the energy using Robust $P^N$ potentials (5) with parameters

$$\gamma_c^l = \lambda_s |c| \min(-H_l(c) + K, \alpha^h), \tag{14}$$

where $H_l(c)$ is the response given by the Ada-boost classifier to clique $c$ taking label $l$, $\alpha^h$ a truncation threshold $\gamma_c^{\max} = |c|(\lambda_p + \lambda_s \alpha^h)$, and $K = \log \sum_{l'\in\mathcal{L}} e^{H_{l'}(c)}$ a normalising constant.

For our experiments, the cost of pixel labels differing from an associated segment label was set to $k_c^l = (\gamma_c^{\max} - \gamma_c^l)/0.1|c|$. This means that up to 10% of the pixels can take a label different to the segment label without the segment variable changing its state to *free*.

**Model Details** For both dense unary and histogram-based segment potentials 4 dense features were used - colour with 128 clusters, location with 144 clusters, texton and HOG descriptor [6] with 150 clusters. 5000 weak classifiers were used in the boosting process.

**Learning Weights for Hierarchical CRFs** Having learnt potentials $\xi_c(\mathbf{x}_c)$ as described earlier, the problem remains of how to assign appropriate weights $\lambda_c$. This weighting, and the training of hierarchical models in general is not an easy problem and there is a wide body of literature dealing with it [12, 11, 28]. The approach we take to learn these weights uses a coarse to fine, layer-based, local search scheme over a validation set

We first introduce additional notation: $\mathcal{V}^{(i)}$ will refer to the variables contained in the $i^{\text{th}}$ layer of the hierarchy, while $\mathbf{x}^{(i)}$ is the labelling of $\mathcal{V}^{(i)}$ associated with a MAP estimate over the truncated hierarchical CRF consisting of the random variables $\mathbf{v}' = \{v \in \mathcal{V}^{(k)} : k \geq i\}$. Given the validation data we can determine a dominant label $L_c$ for each segment $c$, such that $L_F = l$ when $\sum_{i\in l} \Delta(x_i = l) = 0.5|c|$, if there is no such dominant label, we set $L_c = L_F$.

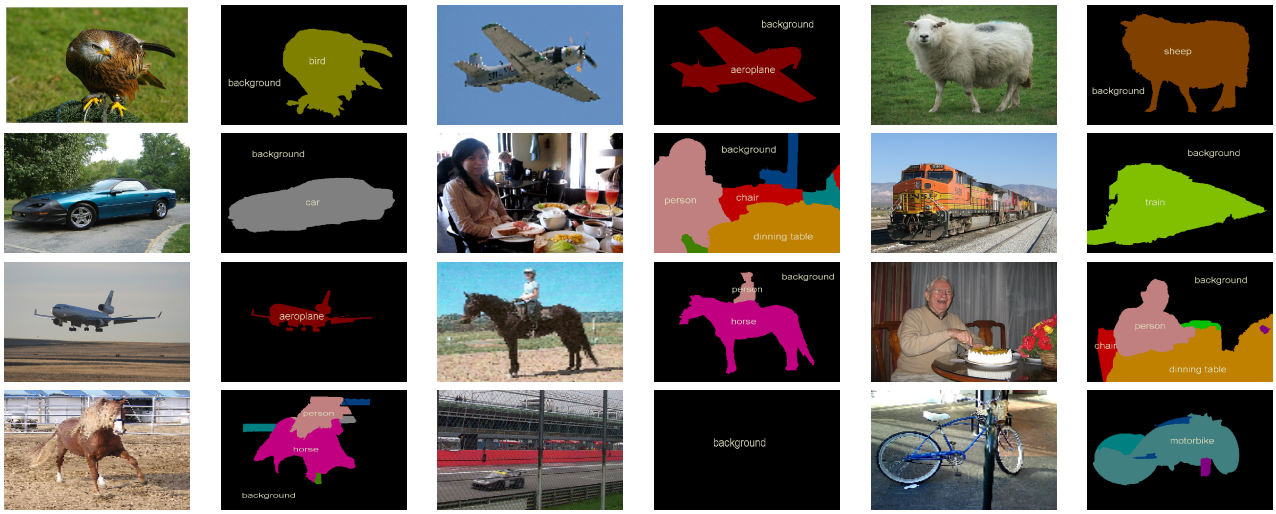



Figure 5. *Qualitative results on the VOC-2008 data set.* **Successful segmentations** *(top 3 rows) and* **standard failure cases** *(bottom) - from left to right, context error, detection failure and misclassification.*

We note that at a given level of the hierarchy, the label of a clique $x_c^{(i)}$ must correspond to the dominant label of of this clique in the ground truth (or $L_F$) for its containing pixels to be correctly labelled. Based on this observation, we propose a simple heuristic which we optimise for each layer.

At each layer, we seek to minimise the discrepancy between the dominant ground truth label of a clique $L_C$, and the value $x_c^{(i)}$ of the MAP estimate. Formally, we choose parameters $\lambda$ to minimise

$$C(\mathbf{x}^{(i)}) = \sum_{c \in \mathcal{V}^{(i)}} \Delta(x_c^{(i)} \neq L_c \wedge L_c \neq L_F). \qquad (15)$$

We optimise (15) layer by layer. The full method is given in algorithm 1, where we use $\lambda_1^{(i)}$ to refer the weighting of unary potentials in the $i^{\text{th}}$ layer, for $\lambda_2^{(i)}$ the weight of the pairwise terms and $\lambda_h^{(i+1)}$ a scalar modifier of all terms in the $(i+1)^{\text{th}}$ layer or greater. $\Theta$ is an arbitrary constant that controls the precision of the final assignment of $\lambda$.

**for** $i$ from $n$ down to 1 **do**
  Let $s_1, s_2, s_h, d_1, d_2, d_h = 1$;
  **while** $s_1, s_2$ *or* $s_h \geq \Theta$ **do**
    **for** $t \in \{1, 2, h\}$ **do**
      $\lambda_t'^{(i)} \leftarrow \lambda_t^{(i)} + d_t s_t$;
      Perform MAP estimate of $\mathbf{x}_i$ using $\lambda_t'$ instead of $\lambda_t$;
      **if** $C(\mathbf{x}_i)$ has decreased **then**
        $\lambda_t \leftarrow \lambda_t'$
      **else**
        $s_t \leftarrow s_t/2, d_t \leftarrow -d_t$

**Algorithm 1**: Weight Learning Scheme.

An alternative and elegant approach to this is that of [9] which we intend to investigate in future work.

## 5. Experiments

We evaluated our framework on two data sets: PASCAL VOC 2008 [7] and MSRC-21 [26].

**Generation of multiple nested segmentations** Normally generation of multiple segmentations is performed by varying the parameters controlling unsupervised segmentation methods [5, 8, 24]. In our experiments, we used mean-shift [5] to generate each set of nested segmentations with both fine and coarse mean-shift kernels. Firstly, a fine kernel based mean shift is applied to create the finest segmentation and then a coarse kernel based mean-shift is performed over the previous result. Multiple nested segmentations can be obtained by varying parameters of both kernels.

**MSRC-21** The MSRC segmentation data set contains 591 images of resolution $320 \times 213$ pixels, accompanied with a hand labelled object segmentation of 21 object classes. Pixels on the boundaries of objects are not labelled in these segmentations. The division into training, validation and test sets occupied 45% 10% and 45% the images. Methods are typically compared using global criteria or average-per-class criteria (see figure 4 for details). For these experiments, the hierarchy was composed of 3 pairs of nested segmentations. The parameters of the mean-shift kernels were arbitrarily chosen as $(6, 5), (12, 10)$; $(6, 7.5), (12, 15)$; and $(6, 9), (12, 18)$. The first value refers to the planar distance between points, and the second refers to the distance in the LUV colour space.

**PASCAL VOC 2008** This data set was used for the PASCAL Visual Object Category segmentation contest 2008. It is especially challenging given the presence of significant background clutter, illumination effects and occlusions. It contains 511 training, 512 validation and 512 segmented test images of 20 foreground and 1 background classes. The organisers also provided $10,057$ images for which only the bounding boxes of the objects present in the image are marked. We did not use these additional images for training

| | Average | Background | Aeroplane | Bicycle | Bird | Boat | Bottle | Bus | Car | Cat | Chair | Cow | Dining table | Dog | Horse | Motor bike | Person | Potted plant | Sheep | Sofa | Train | TV/monitor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XRCE | **25.4** | 75.9 | 25.8 | 15.7 | 19.2 | **21.6** | 17.2 | **27.3** | **25.5** | **24.2** | 7.9 | **25.4** | 9.9 | **17.8** | **23.3** | **34.0** | 28.8 | **23.2** | **32.1** | **14.9** | **25.9** | **37.3** |
| UIUC / CMU | 19.5 | **79.3** | 31.9 | **21.0** | 8.3 | 6.5 | **34.3** | 15.8 | 22.7 | 10.4 | 1.2 | 6.8 | 8.0 | 10.2 | 22.7 | 24.9 | 27.7 | 15.9 | 4.3 | 5.5 | 19.0 | 32.1 |
| MPI | 12.9 | 75.4 | 19.1 | 7.7 | 6.1 | 9.4 | 3.8 | 11.0 | 12.1 | 5.6 | 0.7 | 3.7 | 15.9 | 3.6 | 12.2 | 16.1 | 15.9 | 0.6 | 19.7 | 5.9 | 14.7 | 12.5 |
| Hierarchical CRF | 20.1 | 75.0 | **36.9** | 4.8 | **22.2** | 11.2 | 13.7 | 13.8 | 20.4 | 10.0 | **8.7** | 3.6 | **28.3** | 6.6 | 17.1 | 22.6 | **30.6** | 13.5 | 26.8 | 12.1 | 20.1 | 24.8 |

Figure 6. *Quantitative analysis of VOC2008 results [7] based upon performance criteria ($\frac{\sum_{i\in\mathcal{L}} N_{ii}}{|\mathcal{L}|(-N_{ii}+\sum_{j\in\mathcal{L}} N_{ij}+N_{ji})}$). Note that all other methods used classification and detection priors trained over a much larger data set that included unsegmented images.*

our framework. For this data set we used a two-level hierarchy. The methods are evaluated using average-per-class criteria [7] that penalises the performance of classes $i$ and $j$ given a mislabelling of $i$ as $j$ (see figure 6). Note that it is not equivalent to the percentage of pixels correctly labelled.

**Quantitative and Qualitative Results** Comparisons of our performances against other methods is given in figures 4 and 6. The results on the MSRC data set clearly show that our hierarchical CRF framework outperforms all existing pixel and segment-based methods. Similar results were obtained on the VOC2008 data set, where the only comparable methods used classification and detection priors trained over a much larger set of images.

## 6. Conclusions and Future Work

We have presented a generalisation of many previous super-pixel based methods within a principled CRF framework. Our approach enabled the integration of features and contextual priors defined over multiple image quantisations in one optimisation framework that supports efficient MAP estimation using graph cut based move making algorithms. In order to do this, we have examined the use of auxiliary variables in CRFs which have been relatively neglected in computer vision over the past twenty years.

The flexibility and generality of our framework allowed us to propose and use novel pixel and segment based potential functions and achieve state-of-the-art results on some of the most challenging data sets for object class segmentation. We believe that the use of the hierarchical CRF will yield similar improvements for other labelling problems.

This work was supported by EPSRC, HMGCC and the PASCAL2 Network of Excellence. Professor Torr is in receipt of a Royal Society Wolfson Research Merit Award. We would further like to thank Mark Everingham for his work in maintaining the VOC data sets, and his consistently helpful responses.

## References

[1] D. Batra, R. Sukthankar, and C. Tsuhan. Learning class-specific affinities for image labelling. In *CVPR*, 2008.

[2] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, 2001.

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

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001.

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 2002.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004.

[9] T. Finley and T. Joachims. Training structural svms when exact inference is intractable. In *ICML*, 2008.

[10] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008.

[11] X. He, R. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *ECCV*, 2006.

[12] G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006.

[13] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[14] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

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

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

[17] D. Larlus and F. Jurie. Combining appearance models and markov random fields for category level object segmentation. In *CVPR*, 2008.

[18] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations. In *ECCV*, 2008.

[19] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.

[20] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003.

[21] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[22] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[23] C. Russell, L. Ladicky, P. Kohli, and P. Torr. Exact and approximate inference over hierarchical crfs. *Technical report*, 2009.

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 2000.

[25] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[26] J. Shotton, J. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, pages 1–15, 2006.

[27] H. Tao, H. Sawhney, and R. Kumar. A global matching framework for stereo computation. In *ICCV*, 2001.

[28] B. Taskar, V. Chatalbashev, and D. Koller. Learning associative markov networks. In *Proc. ICML*, page 102. ACM Press, 2004.

[29] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection, 2004.

[30] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *ICCV*, 2003.

[31] J. Wang, P. Bhat, A. Colburn, M. Agrawala, and M. Cohen. Interactive video cutout. *ACM Trans. Graph.*, 2005.

[32] L. Yang, P. Meer, and D. J. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *CVPR*, 2007.

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