What, Where and How Many? Combining Object Detectors and CRFs

L'ubor Ladický, Paul Sturgess, Karteek Alahari, Chris Russell, and Philip H.S. Torr*

Oxford Brookes University
http://cms.brookes.ac.uk/research/visiongroup

Abstract. Computer vision algorithms for individual tasks such as object recognition, detection and segmentation have shown impressive results in the recent past. The next challenge is to integrate all these algorithms and address the problem of scene understanding. This paper is a step towards this goal. We present a probabilistic framework for reasoning about regions, objects, and their attributes such as object class, location, and spatial extent. Our model is a Conditional Random Field defined on pixels, segments and objects. We define a global energy function for the model, which combines results from sliding window detectors, and low-level pixel-based unary and pairwise relations. One of our primary contributions is to show that this energy function can be solved efficiently. Experimental results show that our model achieves significant improvement over the baseline methods on CamVid and PASCAL VOC datasets.

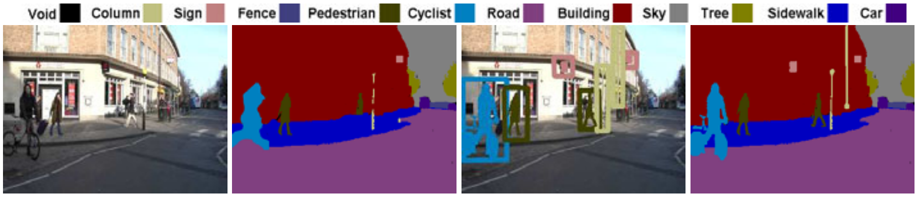
1 Introduction

Scene understanding has been one of the central goals in computer vision for many decades [1]. It involves various individual tasks, such as object recognition, image segmentation, object detection, and 3D scene recovery. Substantial progress has been made in each of these tasks in the past few years [2,3,4,5,6]. In light of these successes, the challenging problem now is to put these individual elements together to achieve the grand goal — *scene understanding*, a problem which has received increasing attention recently [6,7]. The problem of scene understanding involves explaining the whole image by recognizing all the objects of interest within an image and their spatial extent or shape. This paper is a step towards this goal. We address the problems of *what*, *where*, and *how many*: we recognize objects, find their location and spatial extent, segment them, and also provide the number of instances of objects. This work can be viewed as an integration of object class segmentation methods [3], which fail to distinguish between adjacent instances of objects of the same class, and object detection approaches [4], which do not provide information about background classes, such as grass, sky and road.

The problem of scene understanding is particularly challenging in scenes composed of a large variety of classes, such as road scenes [8] and images in the PASCAL VOC

* This work is supported by EPSRC research grants, HMGCC, the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. P. H. S. Torr is in receipt of Royal Society Wolfson Research Merit Award.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part IV, LNCS 6314, pp. 424–437, 2010.



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

Fig. 1. A conceptual view of our method. (a) An example input image. (b) Object class segmentation result of a typical CRF approach. (c) Object detection result with foreground/background estimate within each bounding box. (d) Result of our proposed method, which jointly infers about objects and pixels. Standard CRF methods applied to complex scenes as in (a) underperform on the "things" classes, e.g. inaccurate segmentation of the bicyclist and persons, and misses a pole and a sign, as seen in (b). However, object detectors tend to perform well on such classes. By incorporating these detection hypotheses (§2.2), shown in (c), into our framework, we aim to achieve an accurate overall segmentation result as in (d) (§3.3). (**Best viewed in colour**)

dataset [9]. For instance, road scene datasets contain classes with specific shapes such as person, car, bicycle, as well as background classes such as road, sky, grass, which lack a distinctive shape (Figure 1). The distinction between these two sets of classes — referred to as *things* and *stuff* respectively — is well known [10,11,12]. Adelson [10] emphasized the importance of studying the properties of *stuff* in early vision tasks. Recently, these ideas are being revisited in the context of the new vision challenges, and have been implemented in many forms [12,13,14,15]. In our work, we follow the definition by Forsyth *et al.* [11], where *stuff* is a homogeneous or reoccurring pattern of fine-scale properties, but has no specific spatial extent or shape, and a *thing* has a distinct size and shape. The distinction between these classes can also be interpreted in terms of localization. *Things*, such as cars, pedestrians, bicycles, can be easily localized by bounding boxes unlike *stuff*, such as road, sky[1].

Complete scene understanding requires not only the pixel-wise segmentation of an image, but also an identification of object instances of a particular class. Consider an image of a road scene taken from one side of the street. It typically contains many cars parked in a row. Object class segmentation methods such as [3,8,16] would label all the cars adjacent to each other as belonging to a large car segment or blob, as illustrated in Figure 2. Thus, we would not have information about the number of instances of a particular object—car in this case. On the other hand, object detection methods can identify the number of objects [4,17], but cannot be used for background (*stuff*) classes.

In this paper, we propose a method to jointly estimate the class category, location, and segmentation of objects/regions in a visual scene. We define a global energy function for the Conditional Random Field (CRF) model, which combines results from detectors (Figure 1(c)), pairwise relationships between mid-level cues such as superpixels, and low-level pixel-based unary and pairwise relations (Figure 1(b)). We also show that, unlike [6,18], our formulation can be solved efficiently using graph cut based move

[1] Naturally what is classified as things or stuff might depend on either the application or viewing scale, *e.g.* flowers or trees might be things or stuff.

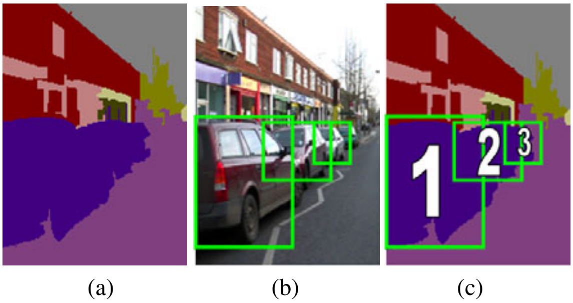


Fig. 2. (a) Object class segmentation results (without detection), (b) The detection result, (c) Combined segmentation and detection. Object class segmentation algorithms, such as [3], label all the cars adjacent to each other as belonging to one large blob. Detection methods localize objects and provide information about the number of objects, but do not give a segmentation. Our method jointly infers the number of object instances and the object class segmentation. See §2.3 for details. (**Best viewed in colour**)

making algorithms. We evaluate our approach extensively on two widely used datasets, namely Cambridge-driving Labeled Video Database (CamVid) [8] and PASCAL VOC 2009 [9], and show a significant improvement over the baseline methods.

Outline of the paper. Section 1.1 discusses the most related work. Standard CRF approaches for the object segmentation task are reviewed in Section 2.1. Section 2.2 describes the details of the detector-based potential, and its incorporation into the CRF framework. We also show that this novel CRF model can be efficiently solved using graph cut based algorithms in Section 2.3. Implementation details and the experimental evaluation are presented in Section 3. Section 4 discusses concluding remarks.

1.1 Related Work

Our method is inspired by the works on object class segmentation [3,6,8,16], foreground (*thing*) object detection [4,17], and relating *things* and *stuff* [12]. Whilst the segmentation methods provide impressive results on certain classes, they typically underperform on *things*, due to not explicitly capturing the global shape information of object class instances. On the other hand, detection methods are geared towards capturing this information, but tend to fail on *stuff*, which is amorphous.

A few object detection methods have attempted to combine object detection and segmentation sub-tasks, however they suffer from certain drawbacks. Larlus and Jurie [19] obtained an initial object detection result in the form of a bounding box, and then refined this rectangular region using a CRF. A similar approach has been followed by entries based on object detection algorithms [4] in the PASCAL VOC 2009 [9] segmentation challenge. This approach is not formulated as one energy cost function and cannot be applied to either cluttered scenes or *stuff* classes. Furthermore, there is no principled way of handling multiple overlapping bounding boxes. Tu *et al.* [15] also presented an effective approach for identifying text and faces, but leave much of the image unlabelled. Gu *et al.* [20] used regions for object detection instead of bounding boxes, but were restricted to using a single over-segmentation of the image. Thus, their approach cannot recover from any errors in this initial segmentation step. In comparison, our method does not make such *a priori* decisions, and jointly reasons about segments and objects.

The work of layout CRF [21] also provides a principled way to integrate things and stuff. However, their approach requires that things must conform to a predefined structured layout of parts, and does not allow for the integration of arbitrary detector responses. To our knowledge, the only other existing approaches that attempt to jointly estimate segmentation and detection in one optimization framework are the works of [6,18]. However, the minimization of their cost functions is intractable and their inference methods can get easily stuck in local optima. Thus, their incorporation of detector potentials does not result in a significant improvement of performance. Also, [6] focussed only on two classes (cars and pedestrians), while we handle many types of objects (*e.g.* 20 classes in the PASCAL VOC dataset). A direct comparison with this method was not possible as neither their code nor their dataset because ground truth annotations are not publicly available at the time of publication.

2 CRFs and Detectors

We define the problem of jointly estimating segmentation and detection in terms of minimizing a global energy function on a CRF model. Our approach combines the results from detectors, pairwise relationships between superpixels, and other low-level cues. Note that our framework allows us to incorporate any object detection approach into any pixel or segment based CRF.

2.1 CRFs for Labelling Problems

In the standard CRF formulation for image labelling problems [3] we represent each pixel as random variable. Each of these random variables takes a label from the set $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$, which may represent objects such car, airplane, bicycle. Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ denote the set of random variables corresponding to the image pixels $i \in \mathcal{V} = \{1, 2, \ldots, N\}$. A clique c is a set of random variables $\mathbf{X}_c$ which are conditionally dependent on each other. A labelling $\mathbf{x}$ refers to any possible assignment of labels to the random variables and takes values from the set $\mathbf{L} = \mathcal{L}^N$.

The posterior distribution $\Pr(\mathbf{x}|\mathbf{D})$ over the labellings of the CRF 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 normalizing constant called the *partition function*, $\mathcal{C}$ is the set of all cliques, and $\mathbf{D}$ the given data. The term $\psi_c(\mathbf{x}_c)$ is known as the potential function of the clique $c \subseteq \mathcal{V}$, where $\mathbf{x}_c = \{x_i : i \in c\}$. The corresponding Gibbs energy is given by: $E(\mathbf{x}) = -\log \Pr(\mathbf{x}|\mathbf{D}) - \log Z = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$. The most probable or Maximum a Posteriori (MAP) labelling $\mathbf{x}^*$ of the random field is defined as: $\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathbf{L}} \Pr(\mathbf{x}|\mathbf{D}) = \arg\min_{\mathbf{x} \in \mathbf{L}} E(\mathbf{x})$.

In computer vision labelling problems such as segmentation or object recognition, the energy $E(\mathbf{x})$ is typically modelled as a sum of unary, pairwise [3,22], and higher order [23] potentials. The unary potentials are based on local feature responses and capture the likelihood of a pixel taking a certain label. Pairwise potentials encourage neighbouring pixels in the image to take the same label. Similarly, a CRF can be defined

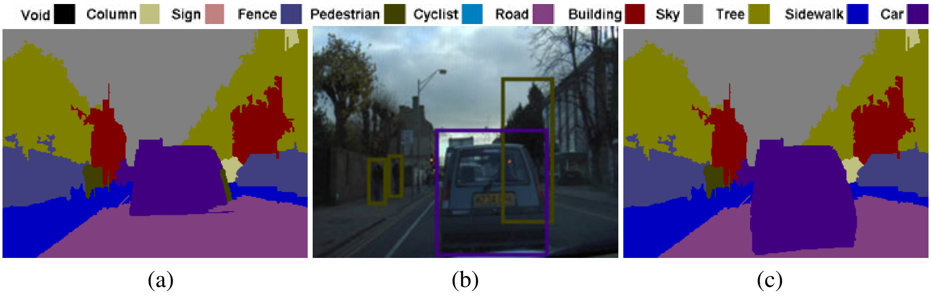


Fig. 3. (a) Segmentation without object detectors, (b) Object detections for car and pedestrian shown as bounding boxes, (c) Segmentation using our method. These detector potentials act as a soft constraint. Some false positive detections (such as the large green box representing person) do not affect the final segmentation result in (c), as it does not agree with other strong hypotheses based on pixels and segments. On the other hand, a strong detector response (such as the purple bounding box around the car) correctly relabels the road and pedestrian region as car in (c) resulting in a more accurate object class segmentation. (**Best viewed in colour**)

over segments [24,25] obtained by unsupervised segmentation [26,27] of the image. Recently, these models have been generalized to include pixels and segments in a single CRF framework by introducing higher order potentials [16]. All these models successfully reason about pixels and/or segments. However, they fail to incorporate the notion of object instances, their location, and spatial extent (which are important cues used by humans to understand a scene) into the recognition framework. Thus, these models are insufficient to address the problem of scene understanding. We aim to overcome these issues by introducing novel object detector based potentials into the CRF framework.

2.2 Detectors in CRF Framework

MAP estimation can be understood as a soft competition among different hypotheses (defined over pixel or segment random variables), in which the final solution maximizes the weighted agreement between them. These weighted hypotheses can be interpreted as potentials in the CRF model. In object class recognition, these hypotheses encourage: (i) variables to take particular labels (unary potentials), and (ii) agreement between variables (pairwise). Existing methods [16,24,25] are limited to such hypotheses provided by pixels and/or segments only. We introduce an additional set of hypotheses representing object detections for the recognition framework[2].

Some object detection approaches [4,19] have used their results to perform a segmentation within the detected areas[3]. These approaches include both the true and false positive detections, and segment them assuming they all contain the objects of interest. There is no way of recovering from these erroneous segmentations. Our approach overcomes this issue by using the detection results as hypotheses that can be rejected

[2] Note that our model chooses from a set of given detection hypotheses, and does not propose any new detections.

[3] As evident in some of the PASCAL VOC 2009 segmentation challenge entries.

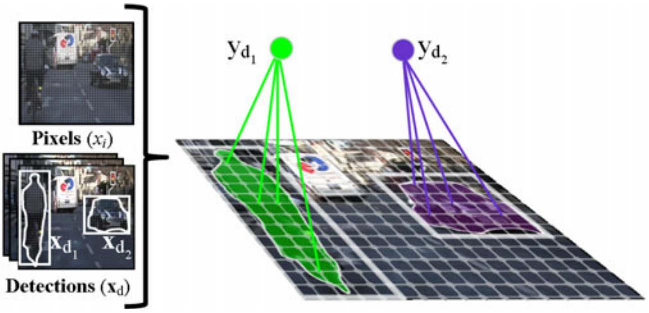


Fig. 4. Inclusion of object detector potentials into a CRF model. We show a pixel-based CRF as an example here. The set of pixels in a detection d_1 (corresponding to the bicyclist in the scene) is denoted by $\mathbf{x}_{d_1}$. A higher order clique is defined over this detection window by connecting the object pixels $\mathbf{x}_{d_1}$ to an auxiliary variable $y_{d_1} \in \{0, 1\}$. This variable allows the inclusion of detector responses as soft constraints. (**Best viewed in colour**)

in the global CRF energy. In other words, all detections act as soft constraints in our framework, and must agree with other cues from pixels and segments before affecting the object class segmentation result. We illustrate this with one of our results shown in Figure 3. Here, the false positive detection for "person" class (shown as the large green box on the right) does not affect the segmentation result in (c). Although, the true positive detection for "car" class (shown as the purple box) refines the segmentation because it agrees with other hypotheses. This is achieved by using the object detector responses[4] to define a clique potential over the pixels, as described below.

Let $\mathcal{D}$ denote the set of object detections, which are represented by bounding boxes enclosing objects, and corresponding scores that indicate the strength of the detections. We define a novel clique potential ψ_d over the set of pixels $\mathbf{x}_d$ belonging to the d-th detection (*e.g.* pixels within the bounding box), with a score H_d and detected label l_d. Figure 4 shows the inclusion of this potential graphically on a pixel-based CRF. The new energy function is given by:

$$E(\mathbf{x}) = E_{pix}(\mathbf{x}) + \sum_{d \in \mathcal{D}} \psi_d(\mathbf{x}_d, H_d, l_d), \tag{1}$$

where $E_{pix}(\mathbf{x})$ is any standard pixel-based energy. The minimization procedure should be able to reject false detection hypotheses on the basis of other potentials (pixels and/or segments). We introduce an auxiliary variable $y_d \in \{0, 1\}$, which takes value 1 to indicate the acceptance of d-th detection hypothesis. Let ϕ_d be a function of this variable and the detector response. Thus the detector potential $\psi_d(.)$ is the minimum of the energy values provided by including ($y_d = 1$) and excluding ($y_d = 0$) the detector hypothesis, as given below:

$$\psi_d(\mathbf{x}_d, H_d, l_d) = \min_{y_d \in \{0,1\}} \phi_d(y_d, \mathbf{x}_d, H_d, l_d). \tag{2}$$

[4] This includes sliding window detectors as a special case.

We now discuss the form of this function $\phi_d(\cdot)$. If the detector hypothesis is included ($y_d = 1$), it should: (a) Encourage consistency by ensuring that labellings where all the pixels in $\mathbf{x_d}$ take the label l_d should be more probable, *i.e.* the associated energy of such labellings should be lower; (b) Be robust to partial inconsistencies, *i.e.* pixels taking a label other than l_d in the detection window. Such inconsistencies should be assigned a cost rather than completely disregarding the detection hypothesis. The absence of the partial inconsistency cost will lead to a hard constraint where either all or none of the pixels in the window take the label l_d. This allows objects partially occluded to be correctly detected and labelled.

To enable a compact representation, we choose the potential ψ_d such that the associated cost for partial inconsistency depends only on the number of pixels $N_d = \sum_{i \in \mathbf{x_d}} \delta(x_i \neq l_d)$ disagreeing with the detection hypothesis. Let $f(\mathbf{x}_d, H_d)$ define the strength of the hypothesis and $g(N_d, H_d)$ the cost taken for partial inconsistency. The detector potential then takes the form:

$$\psi_d(\mathbf{x}_d, H_d, l_d) = \min_{y_d \in \{0,1\}} (-f(\mathbf{x}_d, H_d) y_d + g(N_d, H_d) y_d). \tag{3}$$

A stronger classifier response H_d indicates an increased likelihood of the presence of an object at a location. This is reflected in the function $f(\cdot)$, which should be monotonically increasing with respect to the classifier response H_d. As we also wish to penalize inconsistency, the function $g(\cdot)$ should be monotonically increasing with respect to N_d. The number of detections used in the CRF framework is determined by a threshold H_t. The hypothesis function $f(\cdot)$ is chosen to be a linear truncated function using H_t as:

$$f(\mathbf{x}_d, H_d) = w_d |\mathbf{x}_d| \max(0, H_d - H_t), \tag{4}$$

where w_d is the detector potential weight. This ensures that $f(\cdot) = 0$ for all detections with a response $H_d \leq H_t$. We choose the inconsistency penalizing function $g(\cdot)$ to be a linear function of the number of inconsistent pixels N_d of the form:

$$g(N_d, H_d) = k_d N_d, \quad k_d = \frac{f(\mathbf{x}_d, H_d)}{p_d |\mathbf{x}_d|}, \tag{5}$$

where the slope k_d was chosen such that the inconsistency cost equals $f(\cdot)$ when the percentage of inconsistent pixels is p_d.

Detectors may be applied directly, especially if they estimate foreground pixels themselves. However, in this work, we use sliding window detectors, which provide a bounding box around objects. To obtain a more accurate set of pixels $\mathbf{x}_d$ that belong to the object, we use a local colour model [28] to estimate foreground and background within the box. This is similar to the approach used by submissions in the PASCAL VOC 2009 segmentation challenge. Any other foreground estimation techniques may be used. See §3 for more details on the detectors used. Note that equation (1) could be defined in a similar fashion over superpixels.

2.3 Inference for Detector Potentials

One of the main advantages of our framework is that the associated energy function can be solved efficiently using graph cut [29] based move making algorithms (which outperform message passing algorithms [30,31] for many vision problems). We now show that our detector potential in equation (3) can be converted into a form solvable using $\alpha\beta$-swap and α-expansion algorithms [2]. In contrast, the related work in [6] suffers from a difficult to optimize energy. Using equations (3), (4), (5), and $N_d = \sum_{i \in \mathbf{x_d}} \delta(x_i \neq l_d)$, the detector potential $\psi_d(\cdot)$ can be rewritten as follows:

$$\begin{aligned}\psi_d(\mathbf{x}_d, H_d, l_d) &= \min(0, -f(\mathbf{x}_d, H_d) + k_d \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)) \\ &= -f(\mathbf{x}_d, H_d) + \min(f(\mathbf{x}_d, H_d), k_d \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)). \qquad (6)\end{aligned}$$

This potential takes the form of a Robust P^N potential [23], which is defined as:

$$\psi_h(\mathbf{x}) = \min(\gamma_{max}, \min_l(\gamma_l + k_l \sum_{i \in \mathbf{x}} \delta(x_i \neq l))), \qquad (7)$$

where $\gamma_{max} = f(\cdot), \gamma_l = f(\cdot), \forall l \neq d$, and $\gamma_d = 0$. Thus it can be solved efficiently using $\alpha\beta$-swap and α-expansion algorithms as shown in [23]. The detection instance variables y_d can be recovered from the final labelling by computing y_d as:

$$y_d = \arg \min_{y'_d \in \{0,1\}} (-f(\mathbf{x}_d, H_d) y'_d + g(N_d, H_d) y'_d). \qquad (8)$$

3 Experimental Evaluation

We evaluated our framework on the CamVid [8] and PASCAL VOC 2009 [9] datasets.

CamVid. The Cambridge-driving Labeled Video Database (CamVid) consists of over 10 minutes of high quality 30 Hz footage. The videos are captured at 960 × 720 resolution with a camera mounted inside a car. Three of the four sequences were shot in daylight, and the fourth sequence was captured at dusk. Sample frames from the day and dusk sequences are shown in Figures 1 and 3. Only a selection of frames from the video sequences are manually annotated. Each pixel in these frames was labelled as one of the 32 candidate classes. We used the same subset of 11 class categories as [8,32] for experimental analysis. We have detector responses for the 5 *thing* classes, namely Car, Sign-Symbol, Pedestrian, Column-Pole, and Bicyclist. A small number of pixels were labelled as *void*, which do not belong to one of these classes and are ignored. The dataset is split into 367 training and 233 test images. To make our experimental setup the same as [8,32], we scaled all the images by a factor of 3.

PASCAL VOC 2009. This dataset was used for the PASCAL Visual Object Category segmentation contest 2009. It contains 14,743 images in all, with 20 foreground (*things*) classes and 1 background (*stuff*) class. We have detector responses for all foreground classes. Each image has an associated annotation file with the bounding boxes and the object class label for each object in the image. A subset of these images are also annotated with pixel-wise segmentation of each object present. We used only these images for training our framework. It contains 749 training, 750 validation, and 750 test images.

3.1 CRF Framework

We now describe the baseline CRF formulation used in our experiments. Note that any CRF formulation based on pixels or segments could have been used. We use the Associative Hierarchical CRF model [16], which combines features at different quantization levels of the image, such as pixels, segments, and is a generalization of commonly used pixel and segment-based CRFs. We have a base layer of variables corresponding to pixels, and a hierarchy of auxiliary variables, which encode mid-level cues from and between segments. Furthermore, it assumes that pixels in the same segment obtained using unsupervised segmentation methods, are highly correlated, but are not required to take the same label. This allows us to incorporate multiple segmentations in a principled approach.

In our experiments we used a two level hierarchy based on pixels and segments. Three segmentations are used for the CamVid dataset and six for the PASCAL VOC 2009 dataset; these were obtained by varying parameters of the MeanShift algorithm [26], similar to [16,32].

Pixel-based potentials. The pixel-based unary potential is identical to that used in [16,32], and is derived from *TextonBoost* [3]. It estimates the probability of a pixel taking a certain label by boosting weak classifiers based on a set of shape filter responses. Shape filters are defined by triplets of feature type, feature cluster, and rectangular region and their response for a given pixel is the number of features belonging to the given cluster in the region placed relative to the given pixel. The most discriminative filters are found using the Joint Boosting algorithm [14]. Details of the learning procedure are given in [3,16]. To enforce local consistency between neighbouring pixels we use the standard contrast sensitive Potts model [22] as the pairwise potential on the pixel level.

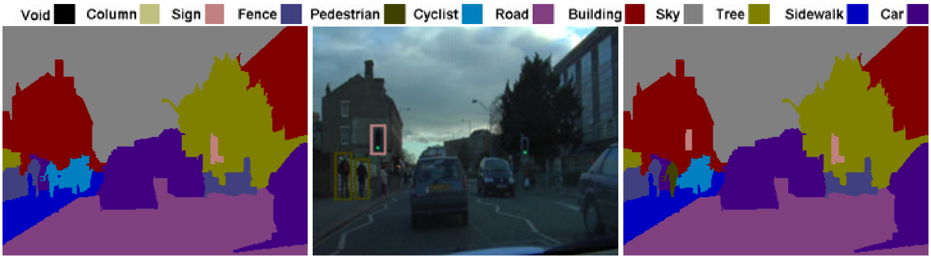
Segment-based potentials. We also learn unary potentials for variables in higher layers (*i.e.* layers other than the base layer), which represent segments or super-segments (groups of segments). The segment unary potential is also learnt using the Joint Boosting algorithm [14]. The pairwise potentials in higher layers (*e.g.* pairwise potentials between segments) are defined using a contrast sensitive (based on distance between colour histogram features) Potts model. We refer the reader to [16] for more details on these potentials and the learning procedure.

3.2 Detection-Based Potentials

The object detections are included in the form of a higher order potential over pixels based on detector responses, as detailed in §2.2. The implementation details of this potential are described below. In order to jointly estimate the class category, location, and segmentation of objects, we augment the standard CRF using responses of two of the most successful detectors[5]: (i) histogram-based detector proposed in [17]; and (ii) parts-based detector proposed in [4]. Other detector methods could similarly be incorporated into our framework.

In [17], histograms of multiple features (such as bag of visual words, self-similarity descriptors, SIFT descriptors, oriented edges) were used to train a cascaded classifier

[5] We thank the authors of [4,17] for providing their detections on the PASCAL VOC 2009 dataset.



(a) (b) (c)

Fig. 5. (a) Segmentation without object detectors, (b) Object detection results on this image showing pedestrian and sign/symbol detections, (c) Segmentation using all the detection results. Note that one of the persons (on the left side of the image) is originally labelled as bicyclist (shown in cyan) in (a). This false labelling is corrected in (c) using the detection result. We also show that unary potentials on segments (traffic light on the right), and object detector potentials (traffic light on the left) provide complementary information, thus leading to both the objects being correctly labelled in (c). Some of the regions are labelled incorrectly (the person furthest on the left) perhaps due to a weak detection response. (**Best viewed in colour**)

composed of Support Vector Machines (SVM). The first stage of the cascade is a linear SVM, which proposes candidate object windows and discards all the windows that do not contain an object. The second and third stages are more powerful classifiers using quasi-linear and non-linear SVMs respectively. All the SVMs are trained with ground truth object instances [9]. The negative samples (which are prohibitively large in number) are obtained by bootstrapping each classifier, as follows. Potential object regions are detected in the training images using the classifier. These potential object regions are compared with the ground truth, and a few of the incorrect detections are added to the training data as negative samples. The SVM is then retrained using these negative and the positive ground truth samples.

In [4] each object is composed of a set of deformable parts and a global template. Both the global template and the parts are represented by HOG descriptors [33], but computed at a coarse and fine level respectively. The task of learning the parts and the global template is posed as a latent SVM problem, which is solved by an iterative method. The negative samples are obtained by bootstrapping the classifier, as described above.

Both these methods produce results as bounding boxes around the detected objects along with a score, which represents the likelihood of a box containing an object. A more accurate set of pixels belonging to the detected object is obtained using local foreground and background colour models [28]. In our experiments we observed that the model is robust to change in detector potential parameters. The parameter p_d (from equation (5)) can be set anywhere in the range $10\% - 40\%$. The parameter H_t (which defines the detector threshold, equation (4)) can be set to 0 for most of the SVM-based classifiers. To compensate the bias towards foreground classes the unary potentials of background class(es) were weighted by factor w_b. This bias weight and the detector potential weight w_d were learnt along with the other potential weights on the validation set using the greedy approach presented in [16]. The CRF was solved efficiently using the graph cut based α-expansion algorithm [2,23].

Table 1. We show quantitative results on the CamVid test set on both recall and intersection vs union measures. 'Global' refers to the overall percentage of pixels correctly classified, and 'Average' is the average of the per class measures. Numbers in bold show the best performance for the respective class under each measure. Our method includes detectors trained on the 5 "thing" classes, namely Car, Sign-Symbol, Pedestrian, Column-Pole, Bicyclist. We clearly see how the inclusion of our detector potentials ('Our method') improves over a baseline CRF method ('Without detectors'), which is based on [16]. For the recall measure, we perform better on 8 out of 11 classes, and for the intersection vs measure, we achieve better results on 9 classes. Note that our method was optimized for intersection vs union measure. Results, where available, of previous methods [8,32] are also shown for reference.

	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist	**Global**	**Average**
Recall[6]													
[8]	46.2	61.9	89.7	68.6	42.9	89.5	**53.6**	46.6	0.7	60.5	22.5	69.1	53.0
[32]	**84.5**	72.6	**97.5**	72.7	34.1	**95.3**	34.2	45.7	8.1	77.6	28.5	**83.8**	59.2
Without detectors	79.3	76.0	96.2	74.6	**43.2**	94.0	40.4	47.0	**14.6**	81.2	31.1	83.1	61.6
Our method	81.5	**76.6**	96.2	**78.7**	40.2	93.9	43.0	**47.6**	14.3	**81.5**	**33.9**	**83.8**	**62.5**
Intersection *vs* Union[7]													
[32]	**71.6**	60.4	**89.5**	58.3	19.4	86.6	**26.1**	35.0	7.2	63.8	22.6	-	49.2
Without detectors	70.0	**63.7**	**89.5**	58.9	17.1	86.3	20.0	**35.8**	9.2	**64.6**	23.1	-	48.9
Our method	71.5	**63.7**	89.4	**64.8**	**19.8**	**86.8**	23.7	35.6	**9.3**	**64.6**	**26.5**	-	**50.5**

3.3 Results

Figures 2, 3 and 5 show qualitative results on the CamVid dataset. Object segmentation approaches do not identify the number of instances of objects, but this information is recovered using our combined segmentation and detection model (from y_d variables, as discussed in §2.3), and is shown in Figure 2. Figure 3 shows the advantage of our soft constraint approach to include detection results. The false positive detection here (shown as the large green box) does not affect the final segmentation, as the other hypotheses based on pixels and segments are stronger. However, a strong detector hypothesis (shown as the purple box) refines the segmentation accurately. Figure 5 highlights the complementary information provided by the object detectors and segment-based potentials. An object falsely missed by the detector (traffic light on the right) is recognized based on the segment potentials, while another object (traffic light on the left) overlooked by the segment potentials is captured by the detector. More details are provided in the figure captions. Quantitative results on the CamVid dataset are shown in Table 1. For the recall measure, our method performs the best on 5 of the classes, and shows near-best ($< 1\%$ difference in accuracy) results on 3 other classes. Accuracy of "things" classes improved by 7% on average. This measure does not consider false positives, and creates a bias towards smaller classes. Therefore, we also provide results with the intersection *vs* union measure in Table 1. We observe that our method shows improved results on almost all the classes in this case.

[6] Defined as $\frac{\text{True Positive}}{\text{True Positive + False Negative}}$.

[7] Defined as $\frac{\text{True Positive}}{\text{True Positive + False Negative + False Positive}}$; also used in PASCAL VOC challenges.

Table 2. Quantitative analysis of VOC 2009 test dataset results [9] using the intersection vs union performance measure. Our method is ranked **third** when compared the 6 best submissions in the 2009 challenge. The method UOCTTI_LSVM-MDPM is based on an object detection algorithm [4] and refines the bounding boxes with a GrabCut style approach. The method BROOKESMSRC_AHCRF is the CRF model used as an example in our work. We perform better than both these baseline methods by 3.1% and 7.3% respectively. Underlined numbers in bold denote the best performance for each class.

	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	**Average**
BONN_SVM-SEGM	**83.9**	64.3	21.8	21.7	**32.0**	**40.2**	**57.3**	49.4	**38.8**	5.2	**28.5**	22.0	19.6	33.6	45.5	33.6	27.3	**40.4**	18.1	33.6	**46.1**	**36.3**
CVC_HOCRF	80.2	**67.1**	**26.6**	**30.3**	31.6	30.0	44.5	41.6	25.2	5.9	27.8	11.0	23.1	**40.5**	**53.2**	32.0	22.2	37.4	**23.6**	40.3	30.2	34.5
UOCTTI_LSVM-MDPM	78.9	35.3	22.5	19.1	23.5	36.2	41.2	50.1	11.7	8.9	**28.5**	1.4	5.9	24.0	35.3	33.4	**35.1**	27.7	14.2	34.1	41.8	29.0
NECUIUC_CLS-DTCT	81.8	41.9	23.1	22.4	22.0	27.8	43.2	**51.8**	25.9	4.5	18.5	18.0	**23.5**	26.9	36.6	**34.8**	8.8	28.3	14.0	35.5	34.7	29.7
LEAR_SEGDET	79.1	44.6	15.5	20.5	13.3	28.8	29.3	35.8	25.4	4.4	20.3	1.3	16.4	28.2	30.0	24.5	12.2	31.5	18.3	28.8	31.9	25.7
BROOKESMSRC_AHCRF	79.6	48.3	6.7	19.1	10.0	16.6	32.7	38.1	25.3	5.5	9.4	25.1	13.3	12.3	35.5	20.7	13.4	17.1	18.4	37.5	36.4	24.8
Our method	81.2	46.1	15.4	24.6	20.9	36.9	50.0	43.9	28.4	**11.5**	18.2	**25.4**	14.7	25.1	37.7	34.1	27.7	29.6	18.4	**43.8**	40.8	32.1

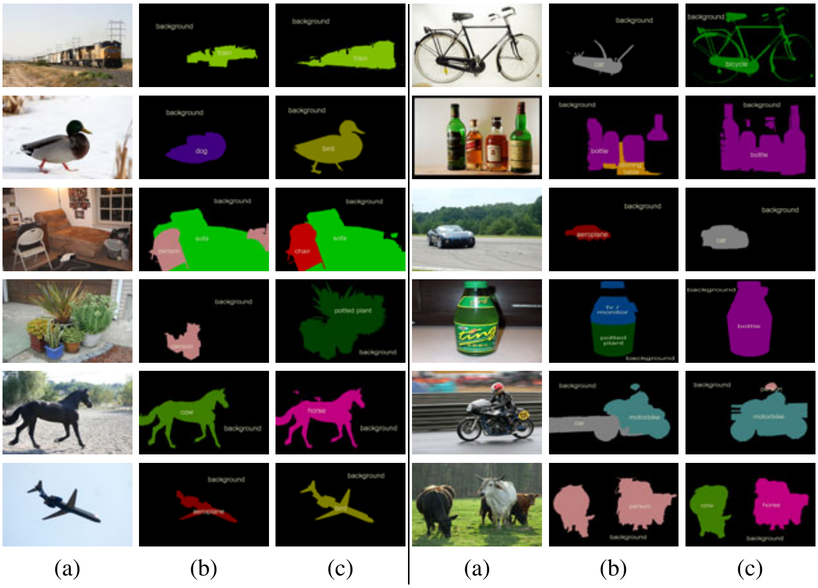


Fig. 6. (a) Original test image from PASCAL VOC 2009 dataset [9], (b) The labelling obtained by [16] without object detectors, (c) The labelling provided by our method which includes detector based potentials. Note that no groundtruth is publicly available for test images in this dataset. Examples shown in the first five rows illustrate how detector potentials not only correctly identify the object, but also provide very precise object boundaries, e.g. bird (second row), car (third row). Some failure cases are shown in the last row. This was caused by a missed detection or incorrect detections that are very strong and dominate all other potentials. (**Best viewed in colour**)

Qualitative results on PASCAL VOC 2009 test set are shown in Figure 6. Our approach provides very precise object boundaries and recovers from many failure cases. For example, bird (second row), car (third row), potted plant (fourth row) are not only correctly identified, but also segmented with accurate object boundaries. Quantitative results on this dataset are provided in Table 2. We compare our results with the 6 best submissions from the 2009 challenge, and achieve the third best average accuracy. Our method shows the best performance in 3 categories, and a close 2nd/3rd in 10 others. Note that using the detector based work (UOCTTI_LSVM-MDPM: 29.0%) and pixel-based method (BROOKESMSRC_AHCRF: 24.8%) as examples in our framework, we improve the accuracy to 32.1%. Both the BONN [34] and CVC [35] methods can be directly placed in our work, and should lead to an increase in performance.

4 Summary

We have presented a novel framework for a principled integration of detectors with CRFs. Unlike many existing methods, our approach supports the robust handling of occluded objects and false detections in an efficient and tractable manner. We believe the techniques described in this paper are of interest to many working in the problem of object class segmentation, as they allow the efficient integration of any detector response with any CRF. The benefits of this approach can be seen in the results; our approach consistently demonstrated improvement over the baseline methods, under the intersection *vs* union measure.

This work increases the expressibility of CRFs and shows how they can be used to identify object instances, and answer the questions: "*What object instance is this?*", "*Where is it?*", and "*How many of them?*", bringing us one step closer to complete scene understanding.

References

1. Barrow, H.G., Tenenbaum, J.M.: Computational vision. IEEE 69, 572–595 (1981)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI 23, 1222–1239 (2001)
3. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
4. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
5. Hoiem, D., Efros, A., Hebert, M.: Closing the loop on scene interpretation. In: CVPR (2008)
6. Gould, S., Gao, T., Koller, D.: Region-based segmentation and object detection. In: NIPS (2009)
7. Li, L.-J., Socher, R., Fei-Fei, L.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: CVPR (2009)
8. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 44–57. Springer, Heidelberg (2008)
9. Everingham, M., et al.: The PASCAL Visual Object Classes Challenge (VOC) Results (2009)
10. Adelson, E.H.: On seeing stuff: the perception of materials by humans and machines. In: SPIE, vol. 4299, pp. 1–12 (2001)

11. Forsyth, D.A., et al.: Finding pictures of objects in large collections of images. In: Buxton, B.F., Cipolla, R. (eds.) ECCV 1996. LNCS, Part II, vol. 1065, pp. 335–360. Springer, Heidelberg (1996)
12. Heitz, G., Koller, D.: Learning spatial context: Using stuff to find things. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 30–43. Springer, Heidelberg (2008)
13. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV (2007)
14. Torralba, A., Murphy, K., Freeman, W.T.: Sharing features: Efficient boosting procedures for multiclass object detection. In: CVPR, vol. 2, pp. 762–769 (2004)
15. Tu, Z., et al.: Image parsing: Unifying segmentation, detection, and recognition. IJCV (2005)
16. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Associative hierarchical crfs for object class image segmentation. In: ICCV (2009)
17. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV (2009)
18. Wojek, C., Schiele, B.: A dynamic conditional random field model for joint labeling of object and scene classes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 733–747. Springer, Heidelberg (2008)
19. Larlus, D., Jurie, F.: Combining appearance models and markov random fields for category level object segmentation. In: CVPR (2008)
20. Gu, C., Lim, J., Arbelaez, P., Malik, J.: Recognition using regions. In: CVPR (2009)
21. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: CVPR (2006)
22. Boykov, Y., Jolly, M.-P.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: ICCV, vol. 1, pp. 105–112 (2001)
23. Kohli, P., Ladicky, L., Torr, P.H.S.: Robust higher order potentials for enforcing label consistency. In: CVPR (2008)
24. He, X., Zemel, R.S., Carreira-Perpiñán, M.Á.: Learning and incorporating top-down cues in image segmentation. In: CVPR, vol. 2, pp. 695–702 (2004)
25. Yang, L., Meer, P., Foran, D.J.: Multiple class segmentation using a unified framework over mean-shift patches. In: CVPR (2007)
26. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space. PAMI (2002)
27. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI 22, 888–905 (2000)
28. Rother, C., Kolmogorov, V., Blake, A.: GrabCut. In: SIGGRAPH, pp. 309–314 (2004)
29. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. PAMI 26, 1124–1137 (2004)
30. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. In: CVPR (2004)
31. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI 28, 1568–1583 (2006)
32. Sturgess, P., Alahari, K., Ladicky, L., Torr, P.H.S.: Combining appearance and structure from motion features for road scene understanding. In: BMVC (2009)
33. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
34. Li, F., Carreira, J., Sminchisescu, C.: Object recognition as ranking holistic figure-ground hypotheses. In: CVPR (2010)
35. Gonfaus, J.M., Boix, X., van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzalez, J.: Harmony potentials for joint classification and segmentation. In: CVPR (2010)