Relating Things and Stuff by High-Order Potential Modeling

Byung-soo Kim[1,⋆], Min Sun[1,⋆], Pushmeet Kohli[2], and Silvio Savarese[1]

[1] University of Michigan, Ann Arbor, U.S.A
[2] Microsoft Research Cambridge, UK

Abstract. In the last few years, substantially different approaches have been adopted for segmenting and detecting "things" (object categories that have a well defined shape such as people and cars) and "stuff" (object categories which have an amorphous spatial extent such as grass and sky). This paper proposes a framework for scene understanding that relates both things and stuff by using a novel way of modeling high order potentials. This representation allows us to enforce labelling consistency between hypotheses of detected objects (things) and image segments (stuff) in a single graphical model. We show that an efficient graph-cut algorithm can be used to perform maximum a posteriori (MAP) inference in this model. We evaluate our method on the Stanford dataset [1] by comparing it against state-of-the-art methods for object segmentation and detection.

1 Introduction

The last decade has seen the development of a number of methods for object detection, segmentation and scene understanding. These methods can be divided into two broad categories: methods that attempt to model and detect object categories that have distinct shape properties such as cars or humans (*things*), and methods that seek to model and identify object categories whose internal structure and spatial support are more heterogeneous such as grass or sky (*stuff*). In the first category, we find that methods based on pictorial structures [2] or generalized Hough transform [3,4] work best. These representations are appropriate for capturing shape or structural properties of *things*, and typically parameterize the object hypothesis by a bounding box. The second category of methods aim at segmenting the image into semantically consistent regions [5,6,7] and work well for *stuff*, like sky or road.

Recently, researchers have proposed methods to jointly detect *things* and segment *stuff*. Gould et al. [8] proposed a random field model incorporating both stuff-stuff, thing-stuff, and thing-horizon relationships. However, MAP inference on their model is computationally expensive and typically takes around five minutes per image. To overcome this limitation, some authors have proposed inference procedures which iteratively solve different visual tasks (e.g., detection,

⋆ Equal contributions.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 293–304, 2012.

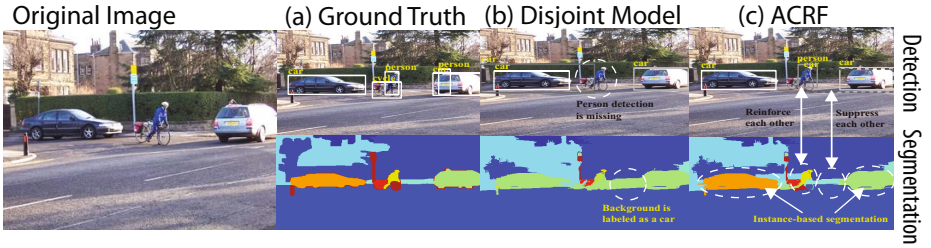

Fig. 1. Our goal is to segment the image into things (e.g., cars, humans, etc) and stuff (e.g., road, sky, etc) by combining segmentation (bottom) with object detection (top). Notice that our final ACRF recovers missing detections, and corrects mistaken segment labels since an object hypothesis and segments are reinforced or suppressed from the novel higher-order potential. At the top of each column, we show the top 3 probable bounding boxes, where light and dark boxes denote the confidence ranking from high to low, and dashed lines are used to indicate false detections. Segmentation results are shown in each bottom column, where we highlight our instance-based segmentation in (c)-bottom, where different colors represent different object instances. Notice that our final ACRF captures the key relationships and recovers many missing detections and segmentation labels by jointly reinforcing or suppressing each other. Thing-Stuff relationships are indicated by arrows connecting a bounding box and segments.

segmentation, occlusion reasoning, etc) using the outputs of state-of-the-art detection or segmentation methods as the input feature [9,10,11]. The drawback of these inference procedures is that different objective functions are optimized independently without guaranteeing that a joint solution is reached and that performances are improved at each iteration.

Ladicky et al. [12] introduce a higher-order potential to incorporate thing-stuff relationships and demonstrate that the information from object detection can be used to improve the segmentation performance. Their higher-order potential is designed so that an efficient graph-cut algorithm can be used to solve the MAP inference problem. However, the model of [12] encourages the labels of the segments to be the same as the label of the detection only when a detection is encountered. When a detection is not found, the labels of the segments are encourage to take labels other than the particular label of the detection. Therefore, the consistency of the object detections and segment labels is only weakly enforced. Finally, both [12,13] cannot be used to assign segments to object instances (i.e., object instances of same class cannot be distinguished in a labeling space.). On the contrary, our proposed model can address both issues.

We propose a novel framework for jointly detecting things and segmenting stuff by using a novel way of modeling high order potentials. Our contributions are three-fold. Firstly, the model enables to segment objects in expanded labeling space where classes as well as instances can be distinguished (see color coded segments in Fig. 1(c)) by associating segments of thing categories to instance-specific labels. Secondly, the higher-order potential enforces two types of consistency (i.e., reinforcement and suppression) between an object hypothesis and

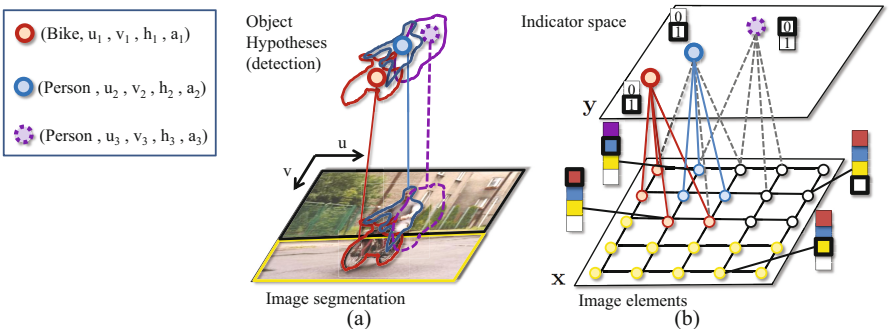

Fig. 2. Our Augmented CRF model (ACRF). In panel (a), we show an image and the indicator variables corresponding to the different object hypotheses present in it. Notice there are two person instance hypotheses in this example. The blue indicator is correct (solid-circle) and the purple indicator is mistaken (dash-circle). In panel (b), the figure shows the label space of the segmentation variables **X** and the indicator variables **Y** using the color-coded column. The interaction between variables are represented using edges. Notice that the dash-edges denote the indicator variable is turned off and suppresses the corresponding label in **X**, and the solid-edges denote the indicator variable is turned on and reinforces the corresponding label in **X**.

segments. As seen in Fig. 1(b), detections typically do not agree with the segmentation results if the detection and segmentation are applied separately. However, in our model the person segments are reinforced by a strong person detection, and the mistaken car segments are suppressed by a weak car detection from the background (Fig. 1(c)). Finally, the special design of the higher-order potential allows efficient inference which takes a few seconds per image in average using graph-cut.

Augmented CRF. Our framework extends the basic conditional random field (CRF) formulations for scene segmentation (i.e., stuff recognition) [14,6] by introducing the concept of an object instance hypothesis (Fig. 2-Top). Each hypothesis is described by object categorical label l, and 2D bounding box (u, v, h, a), where (u, v) denotes the 2D location and (h, a) denote the height and aspect ratio. We refer to our model as the augmented CRF, or ACRF, to highlight the newly added object hypothesis indicator variables. The indicator variables can take only two states, 0 or 1, which represents the absence or presence of an object instance hypothesis, respectively. The edges between two layers of ACRF highlight that labelling consistency between object detection and segment labels is enforced (Fig. 2).

Learning. We formulate the problem of learning these costs as a Structured SVM (SSVM) [15] learning problem with two types of loss functions related to the segmentation loss and detection loss, respectively (see Sec. 4 for details).

MAP Inference. Jointly estimating the segmentation variables X and object indicator variables Y (Fig. 2(c)) is challenging due to the intrinsic difference of the variable space and the presence of high-order potentials between things

and stuff. We design an efficient graph-cut-based move making algorithm by combining state-of-the-art discrete optimization techniques. Our method is based on the α-expansion move making approach [16], which works by projecting the energy minimization problem of segmentation variables X into a binary energy minimization problem to have the same space as the indicator variables Y. Our MAP inference algorithm takes only a few seconds per image in average as opposed to five minutes in [8].

Outline of the Paper. The rest of the paper is organized as follows. We describe the model representation, inference, learning, and implementation details in Sec. 2, 3, and 4, respectively. Experimental results are given in Sec. 5.

2 Augmented CRF

Object segmentation, like other image labelling problems, is commonly formulated using Conditional Random Fields (CRF). The conventional CRF model is defined over a set of random variables $X = \{x_i\},\ i \in \mathcal{V}$ where $\mathcal{V}$ represents the set of image elements, which could be pixels, patches, super-pixels, etc (Fig. 2 (b)-Bottom). Each random variable x_i is assigned to a label from a discrete label space $\mathbf{L}$, which for the task of object-category segmentation, is considered the set $\mathcal{L}$ of object categories such as grass, road, car and people.

The energy (or cost) function $E(X)$ of the CRF is the negative logarithm of the joint posterior distribution of the model and has the following common form: $E(X) = -\log P(X|\mathcal{E}) = -\log \phi_{eRF}(X|\mathcal{E}) + \mathrm{K} = \sum_{c \in \mathcal{C}^X} \psi_c(X_c) + \mathrm{K}$, where $\mathcal{E}$ is the given evidence from the image and any additional information (e.g., object property lists), $\phi_{eRF}(X|\mathcal{E})$ takes the form of a higher order CRF model defined over image elements. $\phi_{eRF}(X|\mathcal{E})$ can be decomposed into potential ψ_c which is a cost function defined over a set of element variables (called a clique) X_c indexed by $c \in \mathcal{C}^X$, $\mathcal{C}^X$ is the set of cliques for image elements, and K is a constant related to the partition function. The problem of finding the most probable or maximum a posteriori (MAP) assignment of the CRF model is equivalent to solving the following discrete optimization problem: $X^* = \arg\min_{X \in \mathcal{L}^{|\mathcal{V}|}} E(X)$.

The standard CRF model mostly relies on bottom-up information. It is constructed using unary potentials based on local classifiers and smoothness potentials defined over pairs of neighboring pixels. Higher-order potentials such as the ones used in [6] reinforce labels of groups of image elements to be the same. This classic representation for object segmentation has led to excellent results for the stuff object categories, but has failed to replicate the same level of performance on the thing object categories.

In addition to the variables representing image elements, our model contains a set of indicator variables (later referred as indicators) $Y = \{y_j \in \{0, 1\}\}$ for every possible configuration $j \in \hat{\mathcal{Q}}$ of an object (Fig. 2 (c)-Top). The configuration set $\hat{\mathcal{Q}}$ is a Cartesian product of the space of all possible object category labels $\mathcal{L}$, all possible 2D bounding boxes in the image. For example, a configuration $j \in \hat{\mathcal{Q}}$ specifies that an instance of the object category $l_j \in \mathcal{L}$ exists at location (u_j, v_j) with height h_j and aspect ratio a_j in the image. We also associate each

object instance with a segmentation mask $\mathcal{V}_j$ which is the set of image elements associated with the object (see technical report [17]).

As mentioned earlier, variables X representing the image elements in the classical CRF formulation for object segmentation take values for the set of object categories $\mathcal{L}$ only. In contrast, in our framework, these variables take values from a set of all possible object configuration $x_i \in \mathbf{L} = \hat{\mathcal{Q}}$ (refer as *augmented labeling space*). On the one hand, this allows us to obtain segmentations of individual instances of particular object categories which the classical CRF formulations are unable to handle. On the other hand, the space of all possible detections $\hat{\mathcal{Q}}$ is clearly huge, which makes learning and inference much more challenging. We will come back to this issue later.

The joint posterior distribution of the segmentation X and indicator variables Y can be written as: $P(X, Y|\mathcal{E}) \propto \phi_{eRF}(X|\mathcal{E})\ \phi_{con}(X, Y|\mathcal{E})$. The potential function ϕ_{con} enforces that the segmentation and indicator variables take values which are consistent with each other (Fig. 2 (b)). The term is formally defined as: $\phi_{con}(X, Y|\mathcal{E}) = \prod_{j \in \hat{\mathcal{Q}}} e^{\Phi(y_j, X)}$, Hence, the model energy can be written as:

$$E(X, Y) = \sum_{c \in \mathcal{C}^X} \psi_c(X_c) + \sum_{j \in \hat{\mathcal{Q}}} \Phi(y_j, X)\ . \tag{1}$$

The first term of the energy function is defined in a manner similar to [6]. We now describe other terms of the energy function in detail in the following subsection.

Implicit Representation of Inactive Object Configurations. It is easy to see that the space of all possible configuration space $\hat{\mathcal{Q}}$ is huge, which would make learning and performing inference in the above model completely infeasible. However, in real world images, only a few possible configurations are actually present. Thus, most indicator variables $y_j,\ j \in \hat{\mathcal{Q}}$ are inactive (take value 0), and similarly the label set for the segmentation variables is typically quite small. We use an object detector that has been trained on achieving high recall rate to generate the set of *plausible* object configuration space $\mathcal{Q}$ instances that are likely to be present in any given image. In this way, we reduce the problem into a manageable size so that the inference algorithm can handle it in practice.

2.1 Relating Object Hypotheses Y and Segments X

The function $\Phi(y_j, X)$ is a likelihood term that enforces consistency in the assignments of the jth indicator variable y_j and a set of segmentation variables X. It is formally defined as:

$$\Phi(y_j, X) = \begin{cases} \inf & \text{if } y_j \neq \delta_j(X) \\ \gamma_{l_j} \cdot |\mathcal{V}_j| & \text{if } y_j = \delta_j(X) = 1 \\ 0 & \text{if } y_j = \delta_j(X) = 0 \end{cases}, \tag{2}$$

where j is any possible object configuration in $\mathcal{Q}$, the function $\delta_j(X)$ indicates whether the indicator j shares a consistent object category label with image elements in $\mathcal{V}_j$, and is defined as:

$$\delta_j(X) = \begin{cases} 1 & \text{if } R_j(X) = \frac{|\mathcal{V}_j(X)|}{|\mathcal{V}_j|} \geq R(l_j) \\ 0 & \text{otherwise} \end{cases} , \tag{3}$$

where $|\mathcal{V}_j(X)| = |\{i; x_i = l_j \texttt{ for } i \in \mathcal{V}_j\}|$ is the number of elements in $\mathcal{V}_j$ assigned with label l_j, $|\mathcal{V}_j|$ is the total number of elements in $\mathcal{V}_j$, $R_j(X)$ is the consistency percentage, and $R(l_j) \in [0\ 1]$ is an object category-specific consistency threshold. Hence, the first condition in the above function ensures that $y_j = 1$ if and only if the detection j shares an object label with at least $R(l_j)$ percent of the pixels (or image element) in $\mathcal{V}_j$ (i.e. $R_j(X) \geq R(l_j)$). The remaining conditions in Eq. 2 shows that if the detection is considered correct by our model, the energy is penalized by $\gamma_j \cdot |\mathcal{V}_j|$, where γ_j is inversely proportional to the detection confidence.

3 Inference

We now show that the MAP inference problem in our ACRF model can be solved by minimizing the energy function using an efficient graph cut based expansion move making algorithm [16].

Standard move making algorithms repeatedly project the energy minimization problem into a smaller subspace in which a sub-problem is efficiently solvable. Solving this sub-problem produces a change to the solution (referred to as a move) which results in a solution having lower or equal energy. The *optimal* move leads to the largest possible decrease in the energy.

The *expansion* move algorithm projects the problem into a Boolean label sub-problem. In an α-expansion move, every segmentation variable X can either retain its current label or transit to the label α. One iteration of the algorithm involves making moves for all α in $\mathcal{L}$ successively. Under the assumption that the projection of the energy is pairwise and submodular, it can be exactly solved using graph cuts [18,19]. We derive graph construction only for energy terms related to indicator variables Y, for all other terms, the constructions are introduced in [6,16].

The energy terms related to the instance indicator variables are $\Phi(y_j, X)$ in Eq. 2. We observe that, when $y_j = 1$

$$\Phi(y_j, X) = \begin{cases} \inf & \text{if } \delta_j(X) = 0 \\ \gamma_j & \text{if } \delta_j(X) = 1 \end{cases} \approx \gamma_j \frac{1 - R_j(X)}{1 - R(l_j)} . \tag{4}$$

When $y_j = 0$

$$\Phi(y_j, X) = \begin{cases} \inf & \text{if } \delta_j(X) = 1 \\ 0 & \text{if } \delta_j(X) = 0 \end{cases} \approx \gamma_j \frac{R_j(X)}{R(l_j)} . \tag{5}$$

Hence, $\Phi(y_j, X)$ can be approximated by

$$\Phi(y_j, X) = \gamma_j (y_j \frac{1 - R_j(X)}{1 - R(l_j)} + (1 - y_j) \frac{R_j(X)}{R(l_j)}) . \tag{6}$$

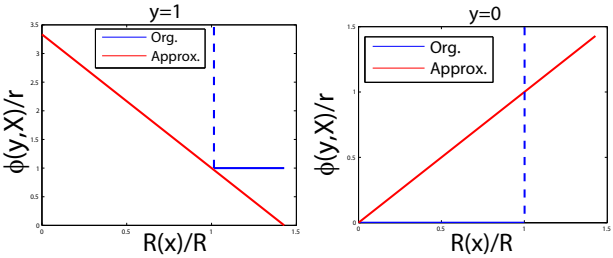

Fig. 3. Comparison between the original function $\Phi(y, X)$ (blue line) and the approximated function (red lines) in Eq. 5 and 4. The left panel shows the case when $y = 1$. The right panel shows the case when $y = 0$. Notice the dash blue lines indicate the sharp transition from finite values to infinite values.

The effect of the approximation in Eq. 5 and 4 are shown in Fig. 3. Instead of imposing an infinite cost when $\delta(X) \neq y$, our approximation imposes an cost which is linearly proportional to the consistency percentage $R(X)$. When $y = 1$, the ratio between the consistency percentage and the consistency threshold $R(X)/R(l)$ are *reinforced* to be large, which means the more elements in X labeled as l the better (Fig. 3-Left). On the contrary, When $y = 0$, the ratio between the consistency percentage and the consistency threshold $R(X)/R(l)$ are *suppressed* to be small, which means the less elements in X labeled as l the better (Fig. 3-Right). In the next section, we show that the approximated higher-order potential becomes pair-wise and submodular when applying the standard transformation function for the α-expansion move.

3.1 α-Expansion Move Energy

We first define the transformation function $T_\alpha(x_i; t_i)$ for the α-expansion move which transforms the label of a random variable x_i as:

$$T_\alpha(x_i; t_i) = \begin{cases} \alpha, & \text{if } t_i = 0 \\ x_i, & \text{if } t_i = 1 \end{cases} \tag{7}$$

The corresponding α-expansion move energy for the term in Eq. 6 can be written as: $\Phi(y_j, T_\alpha(X; T)) =$

$$\begin{cases} \gamma_j(\frac{y_j}{1-R(l_j)}(1 - R_j(X) + \sum_{i\in\mathcal{V}_j(X)} \frac{(1-t_i)}{|\mathcal{V}_j|}) \\ +\frac{1-y_j}{R(l_j)}(\sum_{i\in\mathcal{V}_j(X)} \frac{(t_i)}{|\mathcal{V}_j|})), \text{if } \alpha \neq l_j \\ \gamma_j(\frac{1-y_j}{R(l_j)}(R_j(X) + \sum_{i\in\mathcal{V}_j\setminus\mathcal{V}_j(X)} \frac{(1-t_i)}{|\mathcal{V}_j|}) \\ +\frac{y_j}{1-R(l_j)}(\sum_{i\in\mathcal{V}_j\setminus\mathcal{V}_j(X)} \frac{(t_i)}{|\mathcal{V}_j|})), \text{if } \alpha = l_j \end{cases} \tag{8}$$

where $T = \{t_i\}$ and $\mathcal{V}_j \setminus \mathcal{V}_j(X)$ is the remaining set of elements in $\mathcal{V}_j$ with labels (i.e.,$\{x_i \neq l_j; i \in \mathcal{V}_j\}$). Notice that when $\alpha \neq l_j$ the function is submodular in

(y_j, t_i), but when $\alpha = l_j$ it is submodular in $(\overline{y}_j, t_i)$, where $\overline{y}_j = 1 - y_j$ is the negation of y_j. After the transformation, the original model energy becomes a pairwise and submodular function of T, Y, and $\overline{Y}$ as follows,

$$E(T, Y, \overline{Y}) = \sum_{c \in \mathcal{C}^X} \psi_c(T_c) + \sum_{j \in \hat{\mathcal{Q}}_1} \Phi(y_j, T) + \sum_{j \in \hat{\mathcal{Q}}_2} \Phi(\overline{y}_j, T) \ . \tag{9}$$

where $\hat{\mathcal{Q}}_1 = \{y_j; l_j \neq \alpha\}$ and $\hat{\mathcal{Q}}_1 = \{y_j; l_j = \alpha\}$. Therefore, we will construct the graph using T, partially using indicator y_j, and partially using the negation of indicator $\overline{y}_j$ depending on whether $l_j = \alpha$.

4 Learning

The full CRF model in Eq. 1 contains several terms. In order to balance the importance of different terms, we introduce a set of linear weights for each term as follows,

$$W^T \Psi(X.Y) = \sum_{c \in \mathcal{C}} w_c \psi_c(X_c) + \sum_{j \in \mathcal{Q}_1} w^u(l_j)(\Phi(y_j, X)) \tag{10}$$

where w_c models weights for unary, pair-wise, and higher-order terms in X, and $w^u(l)$ is the object category specific weight for the consistency potential between Y and X.

Assume that a set of example images, ground truth segment object category labels, and ground truth object bounding boxes $\{I^n, X^n, Y^n\}_{n=1,\ldots,N}$ are given. The SSVM problem is as follows,

$$\begin{aligned} &\min_{W, \xi \geq 0} \ W^T W + C \sum_n \xi^n(X, Y) \qquad (11) \\ &\text{s.t.} \quad \xi^n(X, Y) = \max_{X,Y} (\triangle(X, Y; X^n, Y^n) + W^T \Psi(X^n, Y^n) - W^T \Psi(X, Y)), \forall n \ , \end{aligned}$$

where W concatenates all the model parameters which are linearly related to the potentials $\Psi(X, Y)$; C controls the relative weight of the sum of the violated terms $\{\xi^n(X, Y)\}$ with respect to the regularization term; $\triangle(X, Y; X^n, Y^n)$ is the loss function that generates large loss when the X or Y is very different from X^n or Y^n. The designed loss functions and the algorithm we used to solve this optimization problem are described in the technical report [17].

The remaining model parameters are set as follows. The object category-specific $R(l)$ in Eq. 2 are estimated using the median values observed in training data.

5 Experiments

We compare our full ACRF model with [1,20,21,12,13] on Stanford Background (refer as Stanford) dataset [1]. As opposed to other datasets, such as MSRC [14], Stanford dataset contains a large number of cluttered scenes and "things"

(a) Global Accuracy					
[1]	[21]	[20]	[13]	[12]	ACRF
76.4	76.9	77.5	80.0	80.2	**82.0**

(b)	Back-ground	Car	Person	Motor-bike	Bus	Boat	Cow	Sheep	Bi-cycle	Global	Avg.
CRF	**77.4**	49.1	39.9	**15.3**	**76.3**	18.9	65.0	**70.4**	17.3	79.9	47.7
ACRF	77.1	**56.7**	**61.7**	9.3	69.7	**36.9**	**88.1**	62.8	**64.2**	**82.0**	**58.5**

Fig. 4. Segmentation performance comparison on the Stanford dataset. (a) Global segmentation accuracy of our ACRF model compared with state-of-the-art methods, where "Global" is the overall percentage of pixels correctly classified. (b) System analysis of our model. The CRF row shows the results by using only the stuff-stuff relationship component (first term in Eq. 1) of our ACRF model. The last row shows results of the full ACRF model. Notice "Avg." is the average of the percentage over eight foreground classes and one background class.

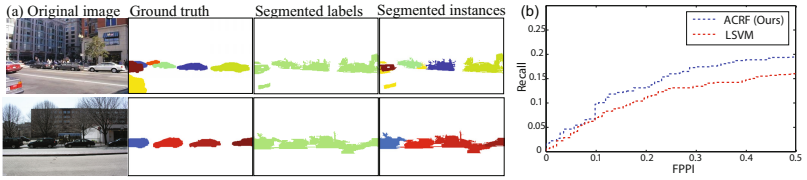

Fig. 5. (a) Typical thing segmentation results on the Stanford dataset. Notice that our model can obtain instance-based segmentations (last column) due to the ability to reason in the augmented labeling space $\hat{\mathcal{Q}}$. (b) Recall v.s. FPPI curves of our ACRF and LSVM on Stanford dataset. Our ACRF achieves better recall at different FPPI values.

object instances per image which makes segmenting and detecting "things" to particularly challenging tasks.

For the experiments below, we use the same pre-trained LSVM detectors [2] to obtain a set of object-instance hypotheses for "things" categories (e.g., car, person, and bike). The object depths are inferred by combining both cues from the size and the bottom positions of the object bounding boxes similar to [22,10]. The responses from off-the-shelf stuff classifiers are used as the unary stuff potentials in our model. We model different types of pair-wise stuff relationships using a codebook representation similar to [23].

Stanford Dataset. Stanford dataset [1] contains 715 images from challenging urban and rural scenes. On top of 8 background ("stuff") categories, we annotate 9 foreground ("things") object categories - car, person, motorbike, bus, boat, cow, sheep, bicycle, others. We follow the 5-fold cross-validation scheme which splits the data into different 572 training and 143 test images. We use the same STAIR Vision Library [24] used in [1] to obtain the stuff unary potentials. Pixel-wise segmentation performance are shown in Fig. 4. Our ACRF model outperforms all state-of-the-art methods [20,1,21,12,13] [1] (Fig. 4(a)). A system analysis of our model (Fig. 4(b)) shows that the performances of most foreground classes (five out of eight) are significantly improved when additional components are added on top of the baseline CRF model, while the performance

[1] We implement [12,13] by ourselves and evaluate the performance.

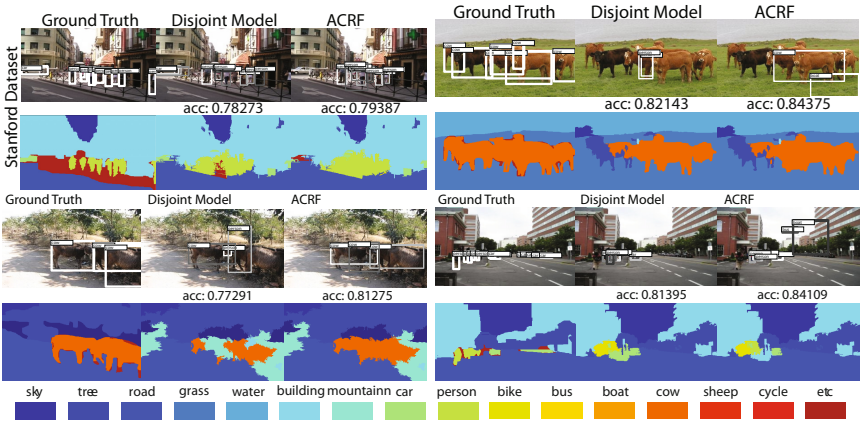

Fig. 6. Typical results on Stanford. Every set of results compare ground truth annotation, disjointed model (disjointedly applied object detection and segmentation), CRF+Det, ACRF, from left to right, respectively. The odd rows show the top K object hypotheses (color-coded bounding boxes representing the confidence ranking from light to dark), where K is the number of recalled objects in the ACRF result. The even rows show the segmentation results (color-code as shown at the bottom).

of the background classes remain almost unchanged. As a result, the full ACRF model obtains the best performance for six out of eight foreground classes and a 10.8% average improvement over the baseline model. Typical results are shown in Fig. 6-Top. We highlight that our model can generate object instance-based segmentations due to the ability to reason in the *augmented labeling space* $\hat{\mathcal{Q}}$ (Fig. 5(a)). Our method can predict the numbers of object instances per image accuractely with an average errors of 0.27.

Another advantage of using our model is the ability to improve detection accuracy. We measured detection performance in terms of Recall v.s. False Positive Per Image (FPPI) in Fig. 5(b), where detection results from 5-fold validations are accumulated and shown in one curve. The performance of the proposed model is compared with the pre-trained LSVM [2]. Our model achieves consistent higher recall than the LSVM baseline as shown in Fig. 5(b).

6 Conclusion

We have presented a unified CRF-based framework for jointly detecting and segmenting "things" and "stuff" categories in natural images. We have shown that our framework incorporates in a coherent fashion various types of (geometrical and semantic) contextual relationships by introducing a novel high order potential model. Our new formulation generalizes previous results based on CRF where the focus was only to reinforce agreement between detections and segmentations. We have quantitatively and qualitatively demonstrated that our method: i) produces better segmentation results than state-of-the art on the Stanford dataset; ii) improves the recall of object instances on Stanford dataset.

Acknowledgements. We acknowledge the support of the Gigascale Systems Research Center and NSF CPS grant #0931474.

References

1. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: ICCV (2009)
2. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
3. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV Workshop on Statistical Learning in Computer Vision (2004)
4. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: CVPR (2009)
5. He, X., Zemel, R.S., Carreira-Perpiñán, M.Á.: Multiscale conditional random fields for image labeling. In: CVPR (2004)
6. Kohli, P., Ladicky, L., Torr, P.H.: Robust higher order potentials for enforcing label consistency. In: CVPR (2008)
7. Shotton, J., Blake, A., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: CVPR (2008)
8. Gould, S., Gao, T., Koller, D.: Region-based segmentation and object detection. In: NIPS (2009)
9. Heitz, G., Gould, S., Saxena, A., Koller, D.: Cascaded classification models: Combining models for holistic scene understanding. In: NIPS (2008)
10. Sun, M., Bao, S.Y., Savarese, S.: Geometrical context feedback loop. IJCV (2012)
11. Hoiem, D., Efros, A.A., Hebert, M.: Closing the loop on scene interpretation. In: CVPR (2008)
12. Ladický, L., Sturgess, P., Alahari, K., Russell, C., Torr, P.H.S.: What, Where and How Many? Combining Object Detectors and CRFs. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 424–437. Springer, Heidelberg (2010)
13. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Graph Cut Based Inference with Co-occurrence Statistics. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 239–253. Springer, Heidelberg (2010)
14. Shotton, J., Winn, J., 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, Part I. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
15. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector learning for interdependent and structured output spaces. In: ICML (2004)
16. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI (2001)
17. Kim, B., Sun, M., Kohli, P., Savarese, S.: Relating things and stuff by high-order potential modeling. Technical report (2012), `http://www.eecs.umich.edu/vision/ACRFproj.html`
18. Boros, E., Hammer, P.: Pseudo-boolean optimization. Discrete Applied Mathematics (2002)
19. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts. PAMI (2004)

20. Tighe, J., Lazebnik, S.: SuperParsing: Scalable Nonparametric Image Parsing with Superpixels. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 352–365. Springer, Heidelberg (2010)
21. Munoz, D., Bagnell, J.A., Hebert, M.: Stacked Hierarchical Labeling. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 57–70. Springer, Heidelberg (2010)
22. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. In: CVPR (2006)
23. Bosch, X.B., Gonfaus, J.M., van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzàlez, J.: Harmony potentials for joint classification and segmentation. In: CVPR (2010)
24. Gould, S., Russakovsky, O., Goodfellow, I., Baumstarck, P., Ng, A., Koller, D.: The stair vision library (v2.3) (2009)