Object Recognition by Integrating Multiple Image Segmentations

Caroline Pantofaru, Cordelia Schmid, Martial Hebert

▶ **To cite this version:**

Caroline Pantofaru, Cordelia Schmid, Martial Hebert. Object Recognition by Integrating Multiple Image Segmentations. ECCV 2008 - 10th European Conference on Computer Vision, Oct 2008, Marseille, France. pp.481-494, 10.1007/978-3-540-88690-7_36 . inria-00548655

HAL Id: inria-00548655

https://inria.hal.science/inria-00548655v1

Submitted on 20 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Object Recognition by Integrating Multiple Image Segmentations

Caroline Pantofaru[1]*, Cordelia Schmid[2], and Martial Hebert[1]

[1] The Robotics Institute, Carnegie Mellon University, USA
[2] INRIA Grenoble, LEAR, LJK, France
crp@ri.cmu.edu, cordelia.schmid@inrialpes.fr, hebert@ri.cmu.edu

Abstract. The joint tasks of object recognition and object segmentation from a single image are complex in their requirement of not only correct classification, but also deciding exactly which pixels belong to the object. Exploring all possible pixel subsets is prohibitively expensive, leading to recent approaches which use unsupervised image segmentation to reduce the size of the configuration space. Image segmentation, however, is known to be unstable, strongly affected by small image perturbations, feature choices, or different segmentation algorithms. This instability has led to advocacy for using multiple segmentations of an image. In this paper, we explore the question of how to best integrate the information from multiple bottom-up segmentations of an image to improve object recognition robustness. By integrating the image partition hypotheses in an intuitive combined top-down and bottom-up recognition approach, we improve object and feature support. We further explore possible extensions of our method and whether they provide improved performance. Results are presented on the MSRC 21-class data set and the Pascal VOC2007 object segmentation challenge.

1 Introduction

The joint tasks of single-image object class recognition and object segmentation are difficult and important. Deformable objects, however, can take on an intractable number of pixel configurations to explore. Bottom-up image segmentation is one possible method for proposing plausible sets of pixels which may compose an object. Unfortunately, recent extensive experiments in [1] and [2] have shown that a single region generated by an image segmentation can rarely be equated to a physical object or object part. Also, image segmentation quality is highly variable, dependent on both the image data, the algorithm and the parameters used, as is clearly visible in Fig. 1. Most importantly, [1] has argued that a particular algorithm and parameter choice will create segmentations of different quality on different images. Even humans do not agree on a 'correct' image partition [3]. In an effort to address these concerns, we join [4–9] in suggesting the use of multiple segmentations per image.

* The authors thank the INRIA associated team Tethys for support.

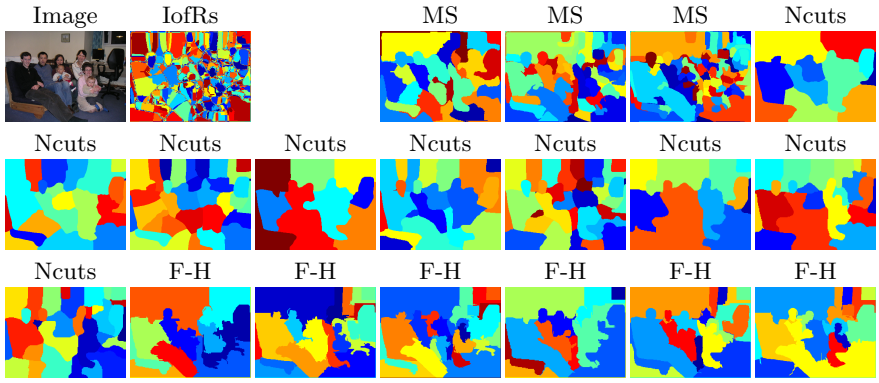

Fig. 1. An example of intersections of regions (IofRs) and the 18 segmentations that generated them: 3 from Mean Shift [10], 9 from Ncuts [11, 12], and 6 from Felzenszwalb and Huttenlocher's method [13].

In this paper, we show that a straightforward approach to integrating the information from multiple image segmentations can provide a more robust basis for object class recognition and object segmentation than one image segmentation alone. Our approach relies on two basic principles: 1) groups of pixels which are contained in the same segmentation region in multiple segmentations should be consistently classified, and 2) the set of regions generated by multiple image segmentations provides robust features for classifying these pixel groups.

The core approach involves generating multiple segmentations of each image, classifying each region in each segmentation, and allowing all of the regions to contribute to the final object map. Using multiple segmentations provides multiple opportunities for discovering object boundaries and creating regions which are appropriate feature supports, thereby providing robustness to outlier poor image segmentations which inevitably occur. This makes it possible to incorporate a segmentation-based approach into a larger system without tedious and potentially futile parameter tuning.

In addition to our core object recognition and object segmentation approach, we explore a number of intuitive extensions, questioning whether they provide worthwhile performance gains. The core approach considers all segmentation-generated regions to be equally useful, so the first extension we attempt is to learn the reliability of a region to predict its contents. Second, we attempt to go beyond independent region classification by modeling adjacent regions and utilizing a random field formulation for global consistency.

The above system is trained using fully supervised data; images with each object carefully masked by a person. Given the expense of creating such a data set, our third extension considers using additional data with noisy, weaker supervision. Finally, since significant work exists on image classification without object localization, and object detection with bounding boxes (or other fixed shapes), we look at whether our approach improves such object information.

2 Related Work

The idea of using unsupervised image segmentation to obtain good spatial support is not new. In practice, however, approaches which use this idea have made strong and questionable assumptions. Russell et al. [4] assume that the entire object falls within one image segmentation region, which is unlikely given object complexity and the simplicity of bottom-up features. In fact, [1] argues that segmentation is rarely 'correct', and [2] shows that often an object encompasses multiple regions. The approaches in [14, 15], as well as others, enforce spatial constraints on object parts that are too rigid for highly deformable objects. Many of the existing approaches to using bottom-up segmentation for recognition have higher complexity and are less intuitive than our own [16, 8, 14].

The object segmentation problem can also be approached by pixel or patch-based methods that do not use image segmentation regions, as in [17–19] and others. These approaches can be useful for repetitive textures like grass, or somewhat rigid objects like faces or cars, but are difficult to apply to deformable objects. They often provide very coarse segmentations by overlapping small patches [20].

The segmentation-integration method we advocate does not make any of these assumption and so is successful over a wide range of object classes.

3 Evaluation

The comparisons in this paper are performed on two difficult data sets, the MSRC 21-class data set [21] and the PASCAL Visual Object Challenge 2007 segmentation competition [22] data set. Each data set contains multiple object classes with extreme variation in deformation, scale, illumination, pose, and occlusion. All results are reported with respect to pixel-level performance, requiring that exact object masks be obtained. On the MSRC 21-class data set, we use the same training and test sets as Shotton et al. [21]. We also compare to the more recent work by Verbeek and Triggs [17], although it uses a different data split. On the PASCAL segmentation set, training and testing sets are as in the challenge, using only the 422 fully segmented images to train our core approach.

4 Core Approach

In this section, we describe the details of our core approach to the joint object recognition and object segmentation problem. The process involves three steps: generating multiple segmentations, describing and classifying each region, and combining the region classifications into an object map indicating which pixels belong to each object. We show that using a single segmentation to generate such object maps produces results of varying quality, while using all of the segmentations in concert produces comparable or improved object map accuracy.

4.1 Generating Multiple Segmentations

To capture the variety in color, edge contrast, texture, image size and noise that images possess, we produce multiple segmentations of each image. We assume (although not guarantee) that all of the object edges will be contained in the union of region outlines. We also assume that each pixel is contained in at least one region which has large enough spatial support for feature computation. Any method for generating multiple segmentations of an image could be used provided it satisfies these assumptions. Here, we describe the particular segmentation algorithms used to create the 18 segmentations for our system.

The first three segmentations are generated by the mean shift-based segmentation algorithm [10] using pixel position, color (in the L*u*v* color space), and a histogram of quantized textons as features [5]. We perform segmentation of images with dimensions scaled to 0.5, 0.75 and 1 times their original lengths. The second set of nine segmentations is generated using the normalized cuts algorithm with the 'probability of boundary' features as in [11, 12]. For each image size, segmentations with 9, 21 and 33 regions are generated (as suggested in [2]). The final six segmentations are generated using the graph-based method by Felzenszwalb and Huttenlocher (F-H) [13]. For each image size we use two values for the parameter $k = \{200, 500\}$, affecting the scale of the final regions.

Examples of the segmentations we generate can be seen in Fig. 1. The granularity of the regions changes with the parameters. The regions created by each algorithm also have different natures. The mean shift segmentation regions are slightly rounded (due to the texture features), smaller, and with accurate boundaries. The normalized cuts regions are also rounded and tend to be of similar size at the cost of subdividing homogeneous regions or joining different textures. The F-H method more easily captures corners and thin, wiry objects, but also produces imprecise boundaries.

4.2 Describing and Classifying Regions from a Single Segmentation

An object map can be created from a single segmentation by classifying each region with one of many available classification algorithms. To instantiate this method, support vector machines as implemented in LIBSVM [23] work well. LIBSVM provides $P(c_r = k|r)$, the probability that the label of region c_r is k, as in [24], and our classification of the region is $\text{argmax}_k P(c_r = k|r)$.

We use three types of features to describe each region. Region position is given by the centroid normalized by the image dimensions. Color is described by a 100-dimensional histogram of quantized hue features [25]. The image structure within and *near* a region is captured by a 300-dimensional region-based context feature (RCF) [5], which is based on a distance-weighted histogram of quantized SIFTs [26]. This yields a 402-dimensional region-specific feature. Since overall image context is often informative, we also aggregate the color and RCF histograms over the entire image for a final set of 802 features.

Examples of good and poor results of using this single-segmentation method can be seen for both data sets in Figs. 3 and 5, columns 5 and 6 respectively.

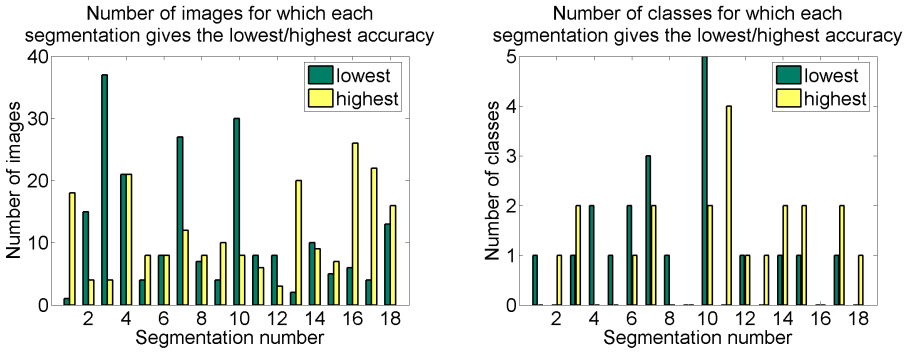

Fig. 2. Histograms of the number of PASCAL 2007 images (left) or object classes (right) for which each single segmentation provides the best or worst pixel accuracy. Each segmentation is the best or worst on at least one image, and most are the best or worst on at least one object class. This suggests that all of the segmentations are useful, and none should be discarded nor used exclusively.

Fig. 4 displays all 18 results for one image. It is visually evident that the object map quality is extremely variable. Tables 1 and 2 confirm this quantitatively. The per-pixel accuracy of the best and worst-performing single segmentations are given for each class, and as an average of the classes. The disparities in class-averaged performance between the best and worst single segmentations are large, 10.2% on the MSRC data set, and 5.5% on the PASCAL data set.

Is there one segmentation/parameter/feature combination which would give the 'best' partition for every image? As shown in [1], the answer is 'no'. Our results on the PASCAL 2007 data suggest the same conclusion. In Table 2, the best overall segmentation has lower classification accuracy than the worst overall segmentation on some classes. This suggests that using only one segmentation is disadvantageous. Fig. 2 shows the number of images (on the left) and the number of object classes (on the right) for which each segmentation gives the worst or best accuracy (for the PASCAL data set). Every segmentation is the best or worst on at least one image, and most of the segmentations are the best or worst on at least one object class. This suggests that none of the segmentations should be discarded as they can all produce useful results, but no one segmentation dominates. Instead of trying to choose one segmentation algorithm, we need to combine the strengths of all the algorithms. We next explore how to combine the individual segmentation results into a more robust object delineation.

4.3 Integrating Multiple Segmentations

Our approach to combining multiple segmentations revolves around two principles. First, pixels which are grouped together by every segmentation should be classified consistently. So, the 'basic units' of our approach are intersections of regions (IofRs), pixels which belong to the same region in every segmentation, as in Fig. 1. Region intersections differ from superpixels [29] as they are con-

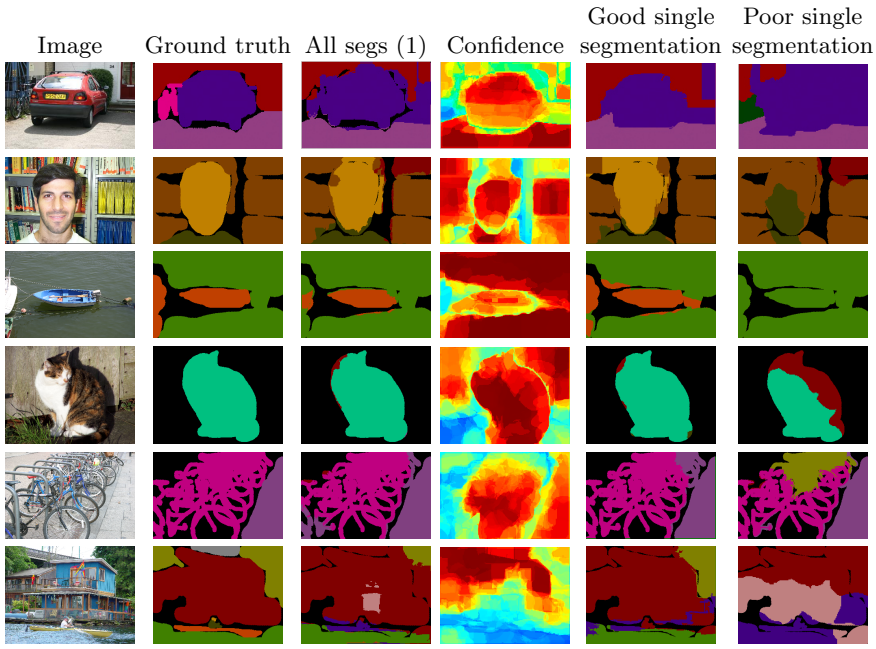

Fig. 3. Object map results from the MSRC 21-class data set. Each map shows the most likely object at each pixel. The third column results from the core multiple segmentation method in (1), with corresponding confidence maps in column four. For comparison, one high and one low-accuracy result of using single segmentations is given for each image. The black pixels in all maps are 'void' in the ground truth. The top five rows show promising results, the last less accurate.

	class avg	**pixel avg**	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bike	flower	sign	bird	book	chair	road	cat	dog	body	boat
Shotton [14]	57.7	72.2	62	98	86	58	50	83	60	53	74	63	75	63	35	19	92	15	86	54	19	62	7
Verbeek [17]	64.0	73.5	52	87	68	73	84	94	88	73	70	68	74	89	33	19	78	34	89	46	49	54	31
Worst seg	49.6	63.3	48	80	69	51	61	87	73	71	57	47	56	34	28	15	75	16	76	28	17	40	11
Best seg	59.8	72.2	61	89	79	57	66	92	81	80	67	63	66	52	31	26	88	27	80	52	32	45	30
All segs (1)	60.3	74.3	68	92	81	58	65	95	85	81	75	65	68	53	35	23	85	16	83	48	29	48	15
All segs (2)	59.9	74.2	68	92	81	57	63	95	82	81	76	65	67	54	34	23	84	16	83	47	30	46	14

Table 1. Pixel accuracy results for the MSRC 21-class data set in the form of the class-averaged pixel accuracy, overall pixel accuracy, and pixel accuracy for each class. The class-averaged and overall accuracies of the multiple segmentation approaches are comparable to using only the best single segmentation, and more importantly they are robust to the worst single segmentation. Using multiple segmentations out-performs the Textonboost approach of Shotton et al. [14], and is comparable to that of Verbeek and Triggs [17] (however [17] uses a different split of the data.)

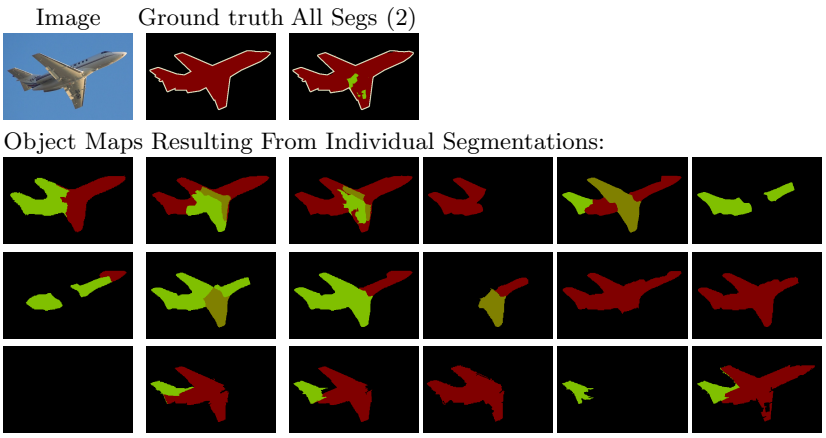

Fig. 4. Example of object segmentation results for a PASCAL VOC2007 image generated using single and multiple image segmentations. The top-left image is the original, the top-middle is the ground truth labeling and the top-right shows the most likely class using all of the segmentations combined with (2). The beige pixels are denoted 'void' in the ground truth, they are *not* generated by our method. The last three rows show the object maps generated using each individual segmentation.

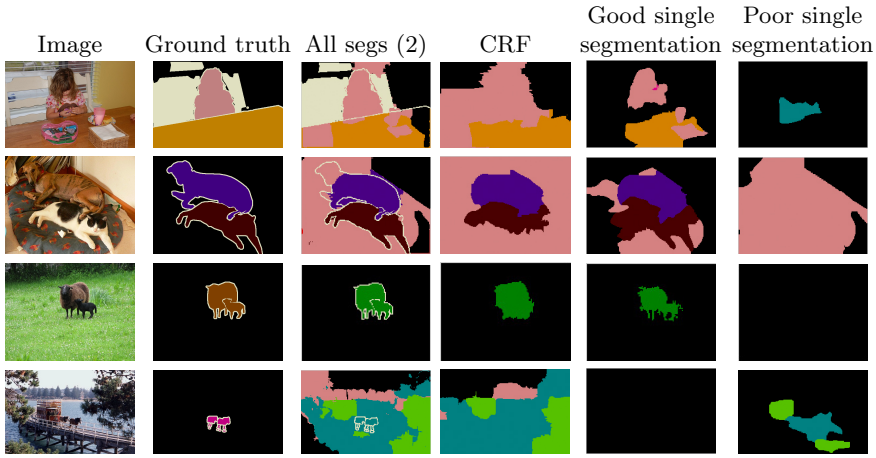

Fig. 5. Object map results from the PASCAL VOC2007 segmentation data set. Each map shows the most confident class at each pixel. The third column was generated using multiple segmentations with (2) and the fourth column with the random field method ($\beta = 0.5$). For comparison, a good and bad single segmentation result is given for each image. The beige pixels in columns two and three are 'void' in the ground truth and not considered in the pixel accuracy results. The first result is promising, the girl and most of the table are correctly labeled. The second result is promising for the difficult dog and cat classes, however the background is misclassified. The third row shows a perfect segmentation but misclassified as 'cow', likely due to the relative scarcity of brown sheep. The final result is complete failure. (Best viewed in color.)

	class avg	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tv/monitor
Brookes [27]	**8.5**	78	6	0	1	1	0	9	5	10	1	2	11	1	6	6	29	2	2	1	11	1
TKK [28]	**30.4**	23	19	21	5	16	3	1	78	1	3	1	23	69	44	42	0	65	30	35	89	71
Worst seg	**12.7**	71	10	7	1	1	8	29	2	14	3	1	7	0	13	20	50	0	5	9	11	8
Best seg	**18.2**	60	15	1	12	1	2	29	11	18	4	4	28	7	23	13	79	8	16	1	21	32
All segs (1)	**19.1**	55	28	1	9	2	1	33	13	17	3	8	31	9	23	16	80	8	19	1	28	17
All segs (2)	**19.6**	59	27	1	8	2	1	32	14	14	4	8	32	9	24	15	81	11	26	1	28	17
CRF	**19.3**	47	25	1	12	1	1	34	15	16	3	7	34	6	23	14	87	8	27	1	28	18

Table 2. Pixel accuracy results for the PASCAL VOC2007 segmentation data set. Given for each approach are the class-averaged pixel accuracy and pixel accuracy for each class. We compare our approach with that of the Oxford Brookes [27] entry into the segmentation competition [27]. The TKK [28] entry had higher performance, but as an entry into the detection challenge it was trained using a much larger data set of thousands of images. Our overall accuracy is much higher than that of Brookes. Overall, the combined segmentation methods both out-perform the single segmentations.

structed by intersecting larger regions, not by image segmentation with small kernel bandwidths or enforcing many regions. Thus IofRs may in fact be quite large in homogeneous image sections (such as the wall in Fig. 1), or small in heterogeneous image sections (such as the people).

The second principle is that the original regions provide better support for extracting features than the IofRs. The IofRs may too small for computing features. Also, the variation in the segmentation-generated regions provides multiple features of different scales and content, increasing the information available.

Thus, our approach is to classify each IofR by combining the information from all of the individual segmentations. Let i be an IofR, and r_i^s the region which contains i in segmentation s. Let c_i be the class label of i, k a specific class label, and I the image data. Then we define segmentation integration method 1 to be:

$$P(c_i = k|I) \propto \sum_s P(c_i^s = k|r_i^s, I) \tag{1}$$

This average over the individual regions' confidences amounts to marginalizing over the regions containing i, assuming they are each equally likely. As before, the class assigned to an IofR is $\text{argmax}_k P(c_i = k|I)$.

Fig. 3 shows selected results for the MSRC 21-class data set. Qualitatively, the results of this multiple segmentation approach are comparable to using the best single segmentation, and robust to the poor segmentation. This conclusion is confirmed quantitatively on both data sets in Tables 1 and 2. Using multiple segmentations gives slightly higher class-averaged accuracy than using the best single segmentation, and the results are robust to the poor performance of the worst single segmentation. For the MSRC data set, our class-averaged and overall

pixel accuracy results out-perform those of [14], and are comparable to those of [17]. For the PASCAL 2007 data set, we out-perform the Oxford Brookes entry [27]. The TKK entry [28] does out-perform ours, however it is not directly comparable as it was an entry in the detection challenge and so trained on thousands of additional images not in the 422-image segmentation training set.

5 Extensions

We have shown that a straightforward method for combining multiple segmentations can lead to robust and accurate object recognition and object segmentation. There are many extensions which could be suggested for our system, here we explore a number of them and ask whether the added complexity is worthwhile.

5.1 Determining the Reliability of a Region's Classification

The core approach assumes that all of the segmentations should have an equal vote in the final classification. Since segmentations differ in quality, another reasonable assumption is that the reliability of a region's prediction corresponds to the number of objects it overlaps. Hoiem et al. [6] suggest learning a classifier to predict the 'homogeneity' of a region with respect to the class labels. If we consider the homogeneity as a measure of the likelihood of a particular region, $P(r_i^s|I)$, then we can write segmentation integration method 2 as:

$$P(c_i = k|I) \propto \sum_s P(r_i^s|I)P(c_i^s = k|r_i^s, I) \tag{2}$$

The classifier used to determine $P(r_i^s|I)$ is a set of boosted decision trees, trained using logistic AdaBoost [30, 31]. We use 20 trees with 16 leaf nodes each to avoid over fitting. The region features used are normalized average position (2D), RCF (300D), color histogram (100D), region size divided by image size (1D), and the number of IofRs encompassed (1D).

Figs. 4 and 5 show qualitative results on the PASCAL 2007 images. We can see once again that using multiple segmentations produces robust object maps. Quantitatively, Tables 1 and 2 show the same conclusion for the overall and class-averaged pixel accuracies. Compared to our first method of integrating segmentations, however, the results are mixed. On the MSRC data set, the original method was slightly better, but on the PASCAL data incorporating homogeneity provides a slight improvement. Whether the expense of computing the homogeneity score for a region is justified is questionable.

Although the official metric for the PASCAL challenge was class-averaged accuracy, examining the overall pixel accuracy provides a very different picture with Brookes achieving 58.4%, our method achieving 50.1%, and TKK achieving 24.4% accuracy. This order reversal is due to the tradeoff between performance on the background class versus other objects, with our approach being the most balanced. These results also demonstrate the importance of using multiple relevant evaluation metrics.

5.2 Incorporating Contextual Information

Our approach thus far has classified regions independently, incorporating contextual and spatial information implicitly by using RCFs, and by using large regions from some segmentations to smooth the labeling of smaller regions in others. One extension is to use explicit spatial information, specifically through a random field formulation of our problem. We can redefine the image labeling problem as an energy minimization, considering potentials of single and pairs of adjacent IofRs in the following manner:

$$E(C) = \sum_i E(c_i) + \sum_{i,j} E(c_i, c_j) \tag{3}$$

Where C is the labeling of the entire image and i, j are neighboring IofRs. The unary potentials are defined as $E(c_i) = -\log P(c_i|I)$ to penalize uncertainty, computed using (2). The binary potentials penalize discontinuity between adjacent labels as suggested by [6]:

$$E(c_i, c_j) = \begin{cases} 0 & \text{if } c_i = c_j, \\ \beta \left(\log p_{ij} - \log \left(1 - p_{ij} \right) \right) & \text{otherwise.} \end{cases} \tag{4}$$

We enforce that $E(c_i, c_j) \geq 0$ and use graph cuts with alpha-expansion to minimize the energy [32]. The p_{ij} reflect the likelihood that the parent regions of adjacent IofRs belong to the same object in each segmentation:

$$p_{ij} \propto \sum_s p_{ij}^s, \quad p_{ij}^s = \begin{cases} 1 & \text{if } r_i^s = r_j^s \\ P\left(c_i^s = c_j^s | r_i^s, r_j^s, I \right) & \text{otherwise} \end{cases} \tag{5}$$

Classifiers for $P\left(c_i^s = c_j^s | r_i^s, r_j^s, I \right)$ are learned by logistic AdaBoost [6] with the following features. The union of the two regions is described using normalized average position, RCF, and color histogram. To compare two regions we use the smaller region size divided by larger region size, the symmetrical KL-divergence between the individual regions' RCFs scaled between 0 and 1, and between the two color histograms, and the normalized difference in region positions.

From Table 2, we can see that the random field results are mixed. We hypothesize that the use of multiple segmentations of various scales, plus the RCFs which model the image surrounding a region, causes most of the label smoothing to occur without the random field. Also, the random field formulation can produce undesired over-smoothing, as in Fig. 5. Finally, the pairwise potentials are difficult to learn due to inaccurate ground truth labeling on object boundaries, as seen throughout this paper. Despite these difficulties, use of the random field does increase the certain class accuracies (bird, person, sheep, etc.), so its use warrants further study.

5.3 Incorporating Weakly Labeled Training Data

So far, all of the training data has been fully labeled with object masks. Generating such ground truth for very large data sets is prohibitively expensive. Even

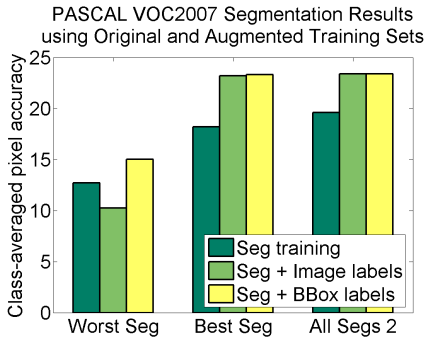

Fig. 6. Class-averaged pixel accuracy on the PASCAL VOC2007 using the 422 fully segmented training images, and augmenting the training set with 400 images with weak image labels (no localization), and weak bounding box labels. Using a relatively small amount of additional, weakly labeled data, the results improve by almost 4%.

the small number of web-based efforts to label data, such as LabelMe [33] and Peekaboom [34], produce inaccurate labels like the 'void' labels throughout this paper. A possible solution to this problem is the use of weakly labeled data to increase training set size. In this section, we increase the size of the PASCAL 2007 segmentation training set from the original 422 fully labeled images by augmenting it with 400 random, weakly labeled images from the larger PASCAL set. The weak labels will take two forms: bounding boxes as in the PASCAL ground truth, and image-level object labels which contain no localization information.

The weakly labeled data is incorporated into our approach by assuming that the weak ground truth labels are noisy object masks. If multiple bounding boxes or image labels exist for one pixel, they are all considered 'correct' for training. We use the augmented training sets to learn the individual region classification probabilities. Since the noise in the bounding box labels lies around the object outlines, the extra images are not used to relearn the homogeneity measure. The procedure is otherwise unchanged. The results of this process can be seen in Fig. 6. Using either augmented data set, the overall class-averaged accuracy increases by nearly 3-5% for both the best single segmentation and the multiple segmentations methods. These results show that a relatively small amount of additional, weakly-labeled training data can improve recognition performance.

5.4 Using Object Detection to Guide Object Segmentation

The final extension we explore is the use of other object recognition systems to provide priors for our object segmentation. Specifically, many object recognition algorithms provide image classification or bounding boxes around identified objects. The other two PASCAL challenges were exactly these tasks. We ask how much using these systems' outputs could potentially improve our results.

We perform a preliminary study by using the ground truth bounding boxes for the PASCAL segmentation challenge test images. Let W_k be a map of pixels

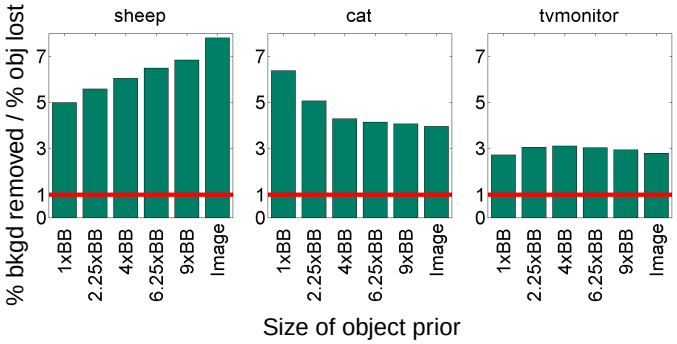

Fig. 7. The effects of using bounding boxes of various sizes as priors for object segmentation on three representative objects from the PASCAL VOC2007 data set. The bars show object mask accuracy improvement using our approach versus the boxes alone. Bar height shows the ratio between the percent of background pixels correctly removed from the boxes by our object segmentation, and the percent of object pixels incorrectly removed. The bars are all above 1; object segmentation always improves upon the bounding boxes.

inside the bounding boxes of object k. Our confidence in class k at pixel q will be: $T(c_q = k|I) = W_k(q)P(c_i = k|I)$, where i is the IofR containing q. We repeat the experiment with increasingly larger bounding boxes until they fill the image, generating an image classification. Using the ground truth bounding boxes as a prior improves the class-averaged pixel accuracy to 79.4%, while using image classification improves the accuracy to 58.9%. The accuracy improvement decreases monotonically between these extremes. This is the best performance we can expect with perfect object detection or image classification.

Of the pixels labeled 'object' by a bounding box mask, some are actually object pixels, while others are actually background. Object segmentation labels only a subset of the bounding box pixels as 'object'. We evaluate the amount our object segmentation improves the bounding boxes by computing the percent of background pixels in the boxes we correctly remove versus the percent of object pixels we incorrectly remove. Fig. 7 shows three trends seen in the behavior of this measure with increasing box size. There are more background pixels incorrectly contained in larger bounding boxes than in smaller, so the intuitive trend would be for object segmentation to offer increasing improvement with increasing box size, as for the sheep class. However, objects such as the cat class show the opposite trend. We speculate that this is due to confusion between classes; as the boxes of other objects increase in size they overlap the true cat pixels and the cat is misclassified. The third trend is for minimal change between box sizes and is seen in more difficult objects. These patterns are interesting and suggest that both strong (bounding box) and weak (image classification) priors can provide large overall improvement, but the effects on individual classes are varied. Most importantly, the bars on plots as in Fig. 7 for all classes and box sizes were above 1, showing that our method always improves the ground truth

boxes. As image classification and object detection systems improve it will be important to compare future results to the 'ideal' situation here.

6 Conclusions

We have presented an intuitive method for recognizing and segmenting objects. Our approach relies on multiple bottom-up image segmentations to support top-down object recognition and allows us to use well-established methods for classification. Aggregating knowledge across image scales and features through multiple segmentations smooths our image labeling in a data-driven manner, increasing robustness. We have presented results on the MSRC 21-class data set and the PASCAL VOC2007 segmentation challenge data set which show that the segmentation combination method not only performs well, but is able to cope with large variation in segmentation quality.

In addition to our core approach, we have suggested extensions and studied whether they are beneficial. Modeling region reliability proved difficult, although class-specific performance improvement warrants further study. On the other hand, increasing the training set size with a relatively small amount of weakly labeled data significantly improved results, and image-level weak labels were sufficient. We also concluded that explicitly incorporating spatial information in a random field was not worthwhile given the implicit spatial information captured in our approach. Finally, we took a preliminary look at using image classification and object detection as a prior for object segmentation.

In conclusion, we believe that this paper stresses two important issues: the importance of algorithm robustness, and the importance of examining whether algorithm extensions reward their added complexity with improved performance.

References

1. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. PAMI **29** (2007)
2. Malisiewicz, T., Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: BMVC. (2007)
3. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV. (2001)
4. Russell, B., Efros, A., Sivic, J., Freeman, W., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR. (2006)
5. Pantofaru, C., Dorkó, G., Schmid, C., Hebert, M.: Combining regions and patches for object class localization. In: Beyond Patches Workshop, CVPR. (2006)
6. Hoiem, D., Efros, A., Hebert, M.: Recovering surface layout from an image. IJCV **75** (2007)
7. Azran, A., Ghahramani, Z.: Spectral methods for automatic multiscale data clustering. In: CVPR. (2006)
8. Tu, Z., Chen, Z., Yuille, A.L., Zhu, S.C.: Image parsing: Unifying segmentation, detection, and recognition. IJCV (2005)

9. Borenstein, E., Malik, J.: Shape guided object segmentation. In: CVPR. (2006)
10. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. PAMI (2002)
11. Fowlkes, C., Martin, D., Malik, J.: Learning affinity functions for image segmentation: Combining patch-based and gradient-based approaches. In: CVPR. (2003)
12. Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using local brightness, color and texture cues. PAMI (2003)
13. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV **59** (2004) 167–181
14. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV. (2006)
15. Winn, J., Jojic, N.: Locus: Learning object classes with unsupervised segmentation. In: ICCV. (2005)
16. Tu, Z., Zhu, S.C.: Image segmentation by data-driven markov chain monte carlo. PAMI **24** (2002) 657–673
17. Verbeek, J., Triggs, B.: Region classification with markov field aspect models. In: CVPR. (2007)
18. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: CVPR. (2006)
19. Kumar, M., Torr, P., Zisserman, A.: Obj cut. In: CVPR. (2005)
20. Leibe, B., Schiele, B.: Interleaved object categorization and segmentation. In: BMVC. (2003)
21. Shotton, J., Winn, J., Rother, C., Criminisi, A.: The MSRC 21-class object recognition database (2006)
22. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The PASCAL VOC2007. http://www.pascal-network.org/challenges/VOC/voc2007 (2007)
23. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
24. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research **5** (2004) 975–1005
25. van de Weijer, J., Schmid, C.: Coloring local feature extraction. In: ECCV. (2006)
26. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
27. Ladicky, L., Kohli, P., Torr, P.: Oxford Brookes entry, PASCAL VOC2007 Segmentation Challenge. http://www.pascal-network.org/challenges/VOC/voc2007 (2007)
28. Viitaniemi, V.: Helsinki University of Technology, PASCAL VOC2007 Challenge. http://www.pascal-network.org/challenges/VOC/voc2007 (2007)
29. Ren, X., Malik, J.: Learning a classification model for segmentation. In: ICCV. (2003)
30. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics (2000)
31. Collins, M., Schapire, R., Singer, Y.: Logistic regression, Adaboost and Bregman distances. Machine Learning (2002)
32. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI **23** (2001) 1222–1239
33. Russell, B., Torralba, A., Murphy, K., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. IJCV (2007)
34. von Ahn, L., Liu, R., Blum, M.: Peekaboom: A game for locating objects in images. In: ACM CHI. (2006)