Semantic Segmentation in Indoor Scenes from Supervised Object-Background Hypotheses

Md. Alimoor Reza and Jana Kosecka
George Mason University
mreza@masonlive.gmu.edu, kosecka@cs.gmu.edu

Abstract

We present a novel framework for semantic segmentation of RGB-D data by effectively combining multiple binary object-background segmentations. The object-background segmentations are learned in a supervised setting, by training binary Conditional Random Field (CRF) models formulated on an image regions of planar and non-planar surfaces. The object hypotheses are combined in a prioritized manner utilizing shape, confidence cues, and object-scene context. The object-scene co-occurrence statistics are exploited both for hard-negative mining for training the data term in the CRF model as well as evaluation methodology.

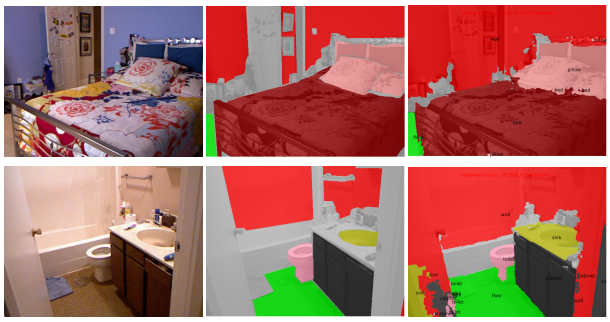

Figure 1. The columns represent the original RGB image, ground truth, ours final semantic segmentation output. Each row represents a case study for a particular scene. The *scene*-types for the first and second rows are *bedroom* and *bathroom* respectively.

1. Semantic Segmentation

The problem of semantic segmentation of an image is to label its each pixel into a set of specific semantic categories. Most methods for semantic segmentation are either grounded on the computational framework of multi-class Conditional Random Field (CRF) that models the contextual relationships among the pixels and superpixels [9][6][8] or on a framework that first generates the bottom-up segmentations learned in unsupervised manner and later on classifies the segments into the semantic categories [3][4][1]. In the following sections we describe the ingredients of our approach. Figure 1 shows final semantic segmentation results from our approach.

Regions: Motivated by the observation that indoor scenes contain many planar structures, we segment an image into regions of planar and non-planar surfaces. Our frameworks starts with identifying the dominant planar surfaces using a RANSAC based approachfrom the depth image. The remaining regions not explained by the planar surfaces are represented as compact SLIC superpixels from the RGB image.

Features: We compute a set of rich and discriminative features over our regions that capture both local and global statistics of the scene utilizing their appearance and geometry. We refer to the work [7] for additional details.

Classifier with negative mining: The computed features are used to learn AdaBoost classifier in a one-vs-all fashion for each of the object e.g. *bed*, *table*, *sofa* etc. The probabilistic output of the classifier is used in the data term of the CRF formulation. Our negative mining strategy utilizes the co-occurrence statistics for selecting important instances from a large pool of negative instances. The important negative instances are sampled according to the distribution of *object*'s co-occurrence with other objects in the training images. We refer to [7] for details.

Object Specific CRF Model: We adopted a standard pairwise CRF framework for generating binary foreground/background segmentation per object category, where the data term is the label likelihood and two smoothness terms using the Potts model with a) color difference and b) 3D position difference based term.

Sequential Combination of Object Specific Segmentation: The final combination of binary CRF proposals is done in a scene dependent manner. We find the most frequently occurring K objects in each scene in training data and sequentially combine their predicted segmentation masks by CRF. To resolve the conflict between an overlapping region we utilize a scheme, where we assign the label

	Wall	Floor	Ceiling	Bed	Pillow	Lamp	Cabinet	Toilet	Sink	Bookshelf	Shelves	Books	Chair	Table	Counter	Sofa	Box	Background
Bedroom	88.22	97.17	48.66	69.94	68.05	50.01	-	-	-	-	-	-	-	-	-	-		43.62
Bathroom	75.24	93.84	0.0	-	-	-	47.46	65.52	62	-	-	-	-	-	-	-	-	29.52
Bookstore	51.52	78.90	89.22	-	-	-	-	-	-	26.95	4.65	28.23	-	-	-	-	-	34.24
Diningroom	70.25	96.14	95.4	-	-	-	45.55	-	-	-	-	-	51.81	61.33	-	-	-	31.04
Homeoffice	81.75	94.08	0.0	-	-	-	-	-	-	57.03	-	-	53.61	43.01	-	-	-	31.95
Kitchen	69.48	97.16	91.07	-	-	-	53.42	-	61.8	-	-	-	-	-	65.35	-	-	47.6
Livingroom	81.42	94.65	95.27	-	-	-	-	-	-	-	-	-	47.48	29.38	-	40.31	-	33.03
Office	75.94	96.94	93.16	-	-	-	30.01	-	-	-	-	-	71.68	-	-	-	30	32.4
Classroom	57.22	97.04	87.94	-	-	-	17.55	-	-	-	-	-	55.53	55.92	-	-	-	35.98
mean	72.34	**93.99**	66.75	**69.94**	68.05	50.01	38.79	65.52	61.9	41.99	4.65	28.23	**56.02**	**47.41**	65.35	**40.31**	30.0	35.49
Couprie[2]	**86.1**	87.3	62.6	38.1	-	-	-	-	-	-	-	13.7	34.1	10.2	-	24.6	-	-
Hermans[5]	71.8	91.5	**83.4**	68.4	-	-	-	-	-	-	-	**45.4**	41.9	27.7	-	28.5	-	-

Table 1. Performance comparison on the NYUD-V2 dataset in pixelwise percentage recall.

	Wall	Floor	Ceiling	Bed	Pillow	Lamp	Cabinet	Toilet	Sink	Bookshelf	Shelves	Books	Chair	Table	Counter	Sofa	Box	Background
Bedroom	56.49	71	31.76	53.86	20.69	9.54	-	-	-	-	-	-	-	-	-	-		39.84
Bathroom	42.07	52.96	0.0	-	-	-	22.3	30.57	22.96	-	-	-	-	-	-	-	-	26.43
Bookstore	15.7	63.45	18.07	-	-	-	-	-	-	17.88	1.14	11.74	-	-	-	-	-	30.02
Diningroom	39.27	65.41	54.98	-	-	-	20.68	-	-	-	-	-	39.86	43.16	-	-	-	24.9
Homeoffice	45.97	66.87	0.0	-	-	-	-	-	-	27.57	-	-	21.64	5.25	-	-	-	29.96
Kitchen	37.44	77.62	23.15	-	-	-	39.58	-	14.24	-	-	-	-	-	40.2	-	-	39.21
Livingroom	44.25	62.19	28.27	-	-	-	-	-	-	-	-	-	14.09	9.23	-	30.95	-	30.02
Office	34.86	70.12	38.4	-	-	-	10.73	-	-	-	-	-	38.98	-	-	-	6.48	29.83
Classroom	26.3	56.55	44.59	-	-	-	8.81	-	-	-	-	-	23.86	20.82	-	-	-	32.34
mean	38.04	65.13	26.58	53.86	20.69	9.54	20.42	30.57	18.6	**22.73**	1.14	**11.74**	27.69	19.62	40.20	30.95	**6.48**	31.39
Gupta[3]	67.6	81.2	**61.1**	57	30.3	16.3	44.8	46.5	35.7	19.5	**4.5**	5.5	36.7	28	**52**	40.8	2.1	-
Gupta[4]	**68**	**81.3**	60.5	**65**	**34.4**	**34.8**	**44.9**	**55.1**	**37.5**	18.1	3.5	6.4	**47.9**	**29.9**	51.3	**47.9**	2.1	-

Table 2. Performance comparison on the NYUD-V2 dataset in pixelwise percentage Jaccard Index.

of object whose overlap region is larger. For example if C is an overlapped region between segments B and A, which have different labels. We compute the ratios $\frac{C}{A}$, and $\frac{C}{B}$ then assign C to labels of the segment that has the larger ratio value.

2. Experiments

To evaluate the performance of our method, we experimented using the 9 most common scene categories as defined in [3] on NYUD-V2 dataset [8] using the standard train/test split 795/654 images in [3] [8] . In Table 1 we report the pixelwise percentage of recall accuracy. The tenth row represents the mean of each per-class accuracy across different scene experiments. Here we report the 3-most frequent occurring objects plus the common background objects (wall, floor, and ceiling) for all of our experiments. We compare against the methods of Couprie et al. [2] and Hermans et al. [5] and our method performs better than the common eight categories that were reported in their settings. On average across eight categories ours (59.37) perform 14 percent better than [2] (44.59) and 2 percent better than [5] (57.33). In Table 2 we compare our performance against methods that reports the performance in pixelwise percentage Jaccard index. Our method performs better on few categories against the methods of [4].

References

[1] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

[2] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[3] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[4] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision (ECCV)*. 2014.

[5] A. Hermans, G. Floros, and B. Leibe. Dense 3d semantic mapping of indoor scenes from rgb-d images. In *International Conference on Robotics Automation (ICRA)*, 2014.

[6] H. Koppula, A. A., J. T., and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*, 2011.

[7] M. Reza and J. Kosecka. Object recognition and segmentation in indoor scenes from rgb-d images. In *Robotics Science and Systems (RSS) conference - 5th workshop on RGB-D: Advanced Reasoning with Depth Cameras*, 2014.

[8] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012.

[9] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pages 702–709, 2012.