Geometry Driven Semantic Labeling of Indoor Scenes

Salman H. Khan[1], Mohammed Bennamoun[1], Ferdous Sohel[1] and Roberto Togneri[2]
[1]School of CSSE, [2]School of EECE, The University of Western Australia
{salman.khan,mohammed.bennamoun,ferdous.sohel,roberto.togneri}@uwa.edu.au

The task of indoor scene labeling is a relatively difficult problem compared to its outdoor counterpart. Indoor scenes have a large number of categories that are significantly different from each other (e.g., corridors, bookstores and kitchens). They also contain illumination variations, clutter, significant appearance variations and imbalanced representation of object categories [6]. Recently, inexpensive structured light sensors (e.g., Microsoft Kinect) are proving to be a rich source of information for indoor scenes. They provide co-registered color (RGB) and depth (D) images in real-time. Efficient use of this information for indoor scene labeling problems is a critical opportunity.

Several recent works focus on the use of RGBD images for scene labeling of indoor scenes. Koppula *et al.* [5] used Kinect fusion to create a 3D point cloud and then densely labeled it using a Markov Random Field (MRF) model. Silberman and Fergus [10] achieved a reasonable semantic labeling performance using a Conditional Random Field (CRF) with SIFT features and 3D location priors. Couprie *et al.* [2] used ConvNets to learn feature representations from RGBD data to label the images while Ren *et.al* [8] employed kernel descriptors to capture the distinctive features. These works are focused on extracting discriminative features from RGBD data and have shown that the depth information can certainly improve the scene labeling performance. However, the question of how to adequately incorporate depth information to model local, pairwise and higher order interactions has not been fully addressed.

In this work, we propose a novel depth-based geometrical CRF model to more efficiently utilize the depth information along side the RGB data. *First*, we incorporate the geometrical information in the most important potential of our CRF model, namely the appearance potential. At the appearance level, we encode both the intensity and depth based characteristics in the feature space. These features are used to predict the unary potentials in a discriminative fashion. Likewise, planes, which are the fundamental geometric units of indoor scenes, are extracted using a new smoothness constraint based *region growing algorithm* (see Sec. 5 in [4]). Compared to other plane detection methods (e.g.,[7, 11]), our method is robust to outer-boundary

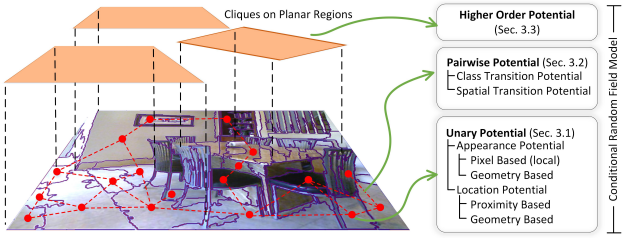

Figure 1: Our approach combines geometrical information with low-level cues with in a CRF model. Only limited graph nodes are shown for the purpose of clear illustration. The detailed sections can be found in [4].

holes present in Kinect's depth maps. The geometric as well as the appearance based characteristics of these planar patches are learned and used to provide unary estimates. We propose a novel *hierarchical fusion scheme* to combine the pixel and planar based unary potentials. This hierarchical scheme first uses a number of contrasting opinion pools and finally combines them using a Bayesian framework (see Sec. 3.1 in [4]).

Next, we turn our attention towards the *location potential*, which encodes the possible spatial locations of all classes. In contrast to the conventional 2D location prior (e.g., in [9, 10]), we propose to integrate the rough geometry of planar regions along with their location in each scene (see Sec. 3.1, 4.1 [4]). We also propose a novel *spatial discontinuity potential* (SDP) in the pairwise smoothness model. It combines a number of different boundaries (such as depth edges, contrast based edges and superpixel edges) and learns a balanced combination of these using a quadratic cost function minimization procedure based on the manually segmented images of the training set (see Sec. 4.2). *Finally*, we add a higher order potential (HOP) in our CRF model which is defined on cliques that encompass planar patches. The proposed HOP increases the expressivity of the random field model by assimilating the geometric context. This encourages all pixels inside a planar patch to take the same class label (see Sec. 3.3).

In short, we have proposed a new random field formulation which elegantly combines the geometric information with the appearance information at various levels of the model hierarchy (Fig. 1).

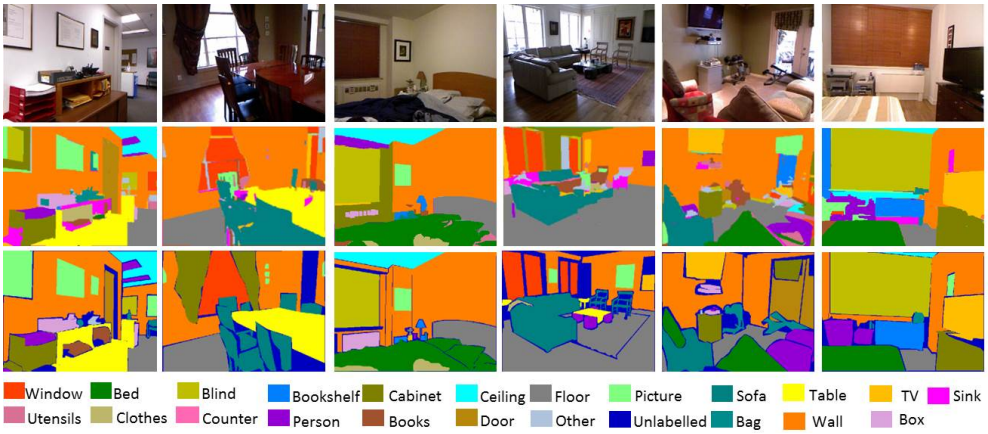

Figure 2: Examples of semantic labeling results on the NYU-Depth v2 dataset. Figure shows intensity images (*top* row), ground truths (*bottom* row) and our results (*middle* row). Our framework performs well in many cases including some unlabeled regions.

Variants of Our Method	NYU-Depth v2		SUN3D	
	Pixel Accuracy	Class Acc.	Pixel Accuracy	Class Acc.
Feature Ensemble (FE)	$44.4 \pm 15.8\%$	39.2%	$41.9 \pm 11.1\%$	40.0%
FE + Planar Appearance Model (PAM)	$52.5 \pm 15.5\%$	42.4%	$48.3 \pm 11.5\%$	42.6%
FE + PAM + Planar Location Prior (PLP)	$55.3 \pm 15.8\%$	43.1%	$51.5 \pm 11.9\%$	43.3%
FE + PAM + PLP + CRF (Regular Potts Model)	$55.5 \pm 15.8\%$	43.2%	$51.8 \pm 12.0\%$	43.5%
FE + PAM + PLP + CRF (SDP + HOP)	$\mathbf{58.3 \pm 15.9\%}$	**45.1%**	$\mathbf{54.2 \pm 12.2\%}$	**44.7%**

Table 1: Semantic Labeling Performance: We report the results of our proposed framework when only variants of unary potentials were used (top 3 rows), a CRF with regular Potts model was used (second last row) and the improvements observed when more sophisticated priors and HOPs (last row) were added. Accuracies are reported for 22 and 13 class semantic labeling for NYU v2 and SUN3D datasets respectively.

Method	Semantic Classes				Pixel Accuracy	Class Accuracy
	Floor	Structure	Furniture	Props		
Supp. Inf. [11]	68	59	**70**	**42**	58.6	59.6
ConvNet [3]	68.1	87.8	51.1	29.9	63	59.2
ConvNet + D [2]	87.3	86.1	45.3	35.5	64.5	63.5
Im $\cup$ 3D [1]	**87.9**	79.7	63.8	27.1	67.0	64.3
This paper	87.1	**88.2**	54.7	32.6	**69.2**	**65.6**

Table 2: Comparison of results on the NYU-Depth v2 (4-class labeling task): Our method achieved best performance in terms of average pixel and class accuracies. We also get the best classification performance on structure class.

We evaluated our framework on the New York University (NYU) Depth dataset (v2) and a recent SUN3D dataset. The NYU dataset [10] consists of 1449 labeled images. SUN3D is a large scale indoor RGBD dataset [12], however it is still under development and only a small portion has been labeled. We extracted keyframes from SUN3D which amounted to 83 labeled images. In our evaluations, we exclude all unlabeled regions. For both datasets, roughly 60%/40% train/test split was used. A relatively small validation set consisting of 50 random images was extracted from NYU-Depth v2. This validation set was used with the genetic search algorithm for the selection of useful features and for the choice of the initial estimates of the parameters which gave the best performance (for SUN3D we used the same parameters). Afterwards, these parameters were optimized during the learning process as described in Sec. 4.2.

We used two popular evaluation metrics to assess our results, '*pixel accuracy*' and '*class accuracy*' (see Table 1). Pixel accuracy accounts for the average number of pixels which are correctly classified in the test set. Class accuracy measures the average of the correct class predictions which is essentially equal to the mean of the values occurring at the diagonal of the confusion matrix. We extensively evaluated our approach on both the NYU-Depth and SUN3D datasets. Our experimental results are shown in Table 1. The comparisons with state-of-the-art techniques are shown in Tables 2. Sample labelings for NYU-Depth v2 are presented in Fig. 2. Although the unlabeled portions in the annotated images are not considered during our evaluations, we observed that the labeling scheme mostly predicts accurate class labels (see Fig. 2).

References

[1] C. Cadena and J. Košecká. Semantic segmentation with heterogeneous sensor coverages. 2014.

[2] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. *ICLR*, 2013.

[3] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *TPAMI*, 35(8):1915–1929, 2013.

[4] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Geometry driven semantic labeling of indoor scenes. In *Computer Vision–ECCV 2014*, pages 679–694. Springer, 2014.

[5] H. S. Koppula, A. Anand, et al. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*, pages 244–252, 2011.

[6] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009.

[7] T. Rabbani, F. van Den Heuvel, and G. Vosselmann. Segmentation of point clouds using smoothness constraint. *Intl. Archives of PRSSIS*, 36(5):248–253, 2006.

[8] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, pages 2759–2766. IEEE, 2012.

[9] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, 2009.

[10] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *ICCV Workshops*, pages 601–608. IEEE, 2011.

[11] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.

[12] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*. IEEE, 2013.