3D Region Proposals For Selective Object Search

Sheetal Reddy[1], Vineet Gandhi[1] and Madhava Krishna[1]
[1]*International Institute of Information Technology, Hyderabad, India*
sheetal.reddy@research.iiit.ac.in, vgandhi@iiit.ac.in, mkrishna@iiit.ac.in

Keywords: RGB-D scene classification, RGB-D semantic segmentation, RGB-D object search

Abstract: The advent of indoor personal mobile robots has clearly demonstrated their utility in assisting humans at various places such as workshops, offices, homes, etc. One of the most important cases in such autonomous scenarios is where the robot has to search for certain objects in large rooms. Exploring the whole room would prove to be extremely expensive in terms of both computing power and time. To address this issue, we demonstrate a fast algorithm to reduce the search space by identifying possible object locations as two classes, namely - Support Structures and Clutter. Support Structures are plausible object containers in a scene such as tables, chairs, sofas, etc. Clutter refers to places where there seem to be several objects but cannot be clearly distinguished. It can also be identified as unorganized regions which can be of interest for tasks such as robot grasping, fetching and placing objects. The primary contribution of this paper is to quickly identify potential object locations using a Support Vector Machine(SVM) learnt over the features extracted from the depth map and the RGB image of the scene, which further culminates into a densely connected Conditional Random Field(CRF) formulated over the image of the scene. The inference over the CRF leads to assignment of the labels - support structure, clutter, others to each pixel.There have been reliable outcomes even during challenging scenarios such as the support structures being far from the robot. The experiments demonstrate the efficacy and speed of the algorithm irrespective of alterations to camera angles, modifications to appearance change, lighting and distance from locations etc.

1 INTODUCTION

The ability to locate a specific object in an indoor environment is a fundamental problem in creating fully autonomous mobile robotic systems. This requires the robot to 1)Locate the object in the exploration environment. 2)Plan a path to reach the object and 3) Perform the desired operation on the object such as servoing to a desired pose, grasping etc.

In indoor scenes, objects are likely to be placed over raised flat surfaces like tables, which we call support structures. Moreover, the objects are often surrounded by several other related articles, which can be termed as clutter. The aim of this work is to locate all support structures and cluttered areas in a given scene. More formally, given a depth and RGB image pair, the proposed method classifies each pixel into one of the three categories i.e. clutter, support structure or other. An example output of the proposed approach is shown in Figure 1, where all the objects (keyboard, mouse, computer screen etc.) are marked as clutter and the rest of the table is marked as support structure. The robot can now move close to the areas marked as clutter to search for the desired object. Furthermore, the obtained result can also be useful for the problem of finding likely locations of placing an object (connected support structure pixels are candidate positions). The motivation behind addressing it as a 3 label problem is to use the labels as a prior for object search. Our work is inspired from the idea that small objects of the order 1cm-5cm, appear very small, making it difficult for the present algorithms to recognize them from far away. A better approach is to guess from far and recognize from near. It is difficult to recognize a single object from far but it is easy to recognize a group of objects placed together.

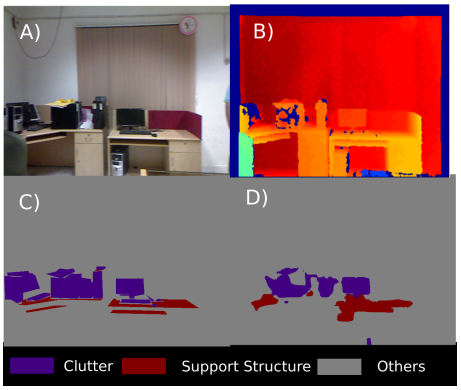

Figure 1: The figure shows sample results of the proposed method. We show an a) input RGB image taken from the LAB dataset. B) shows the input depth image from kinect. C) depicts the ground truth labelling for object search in indoor environments. D) Show the results using our method.

In most of the scenarios small objects are placed on support structures. If support structure is not visible , it is mostly due to the support structure occluded by non distinguishable objects which we define as clutter. Clutter can act as a clue in adding the small object within the search space. These image regions give a strong prior for object search for a robot in an indoor environment.The primary problem is the computation time that previous approaches take. A vision-based autonomous robot needs to tackle the problem of object search efficiently in the shortest possible time. Our proposed method demonstrates simple yet efficient strategy for object class segmentation exploiting the rich geometric information from the 3D point cloud. In summary, our main contributions are:

- We propose a method for segmenting clutter and support structures from RGBD data using a dense CRF formulation over appearance and SVM features extracted from geometry of the scene.
- We use clutter as a clue for recognizing areas where objects could be present despite support structure completely absent in a scene. The absence can be mainly due to two reasons. 1)The support structure being occluded. 2)Unreliable depths after 3m from kinect. Clutter also depicts the presence of an assortment of objects, which can be included in the search space.
- Our quantitative results on NYU and LAB show that our model is reliable across datasets without training on every dataset unlike ALE. We show considerable improvements over ALE on both the datasets.

2 RELATED WORK

Scene labeling, aiming to densely label everything in a scene, has been extensively studied in computer vision. Single color image based methods have been extremely successful, especially in outdoor scenes (Shotton et al., 2009),(Gould et al., 2009),(Ladicky et al., 2010), (Zheng et al., 2015). (Shotton et al., 2009) proposed a segmentation method incorporating a boosting based unary classifiers into a conditional random field (CRF). (Ladicky et al., 2010) showed that global potentials like co-occurrence statistics can be defined over all variables in the CRF to obtain significant improvement in accuracy. More recently, the methods combining CRF's with convolutional neural nets (CNN) have been shown to obtain effective results (Zheng et al., 2015). But purely image based approaches do not perform equally well in the harder case of indoor scenes (Quattoni and Torralba, 2009). They tell only a little about the physical relationships between objects, possible actions that can be performed, or the geometric structure of the scene.

The work by Silberman and Fergus (Silberman and Fergus, 2011) was one of the extensively tested method to demonstrate that incorporating depth data gives a significant performance gain over methods limited to intensity information for the task of indoor scene labeling. A large variety of other RGB-D based segmentation works have been proposed (Ren et al., 2012), (Reza and Kosecka, 2014), (Koppula et al., 2011), (Gupta et al., 2015), (Kim et al., 2013). The work by (Reza and Kosecka, 2014) combines a Adaboost classifier on combined RGB-D features with a CRF framework, to obtain binary segmentation (particular object vs background). Ren et al. (Ren et al., 2012) uses kernel features and solves a standard MRF over superpixels. Gupta et al. (Gupta et al., 2015) propose a framework to exploit depth data in multiple related task of contour detection, object detection to semantic segmentation. In contrast to these approaches, where large number of scene labels have been considered, our approach focuses only on predicting the areas where the objects are more likely to be present (clutter) or the areas where new object could be placed (support structure). This allows us to avoid using large number of complicated features.

In one such work by koppula et al. (Koppula et al., 2011) the point clouds obtained from the Kinect sensor are merged together using RGBDSLAM and are segmented using Euclidean clustering. These segments are the underlying basic structures for MRF and are labelled to different categories. The idea of 3D geometry has also been exploited in the voxel based approach proposed by Kim et al. (Kim et al., 2013). Although, these approaches have shown impressive results, the algorithm can take upto 18 minutes to run on a single stitched point cloud (Koppula et al., 2011), which is unacceptable in most robotic tasks.

In this paper, we extend the framework by Krähenbühl and Koltun (Krähenbühl and Koltun, 2012) to incorporate the depth information. Previous methods have used basic CRF or MRF methods composed of unary potentials on individual pixels or superpixels and pairwise potentials on neighboring pixels or superpixels. It has been shown in the past (Toyoda and Hasegawa, 2008), that fully connected CRF's can improve the accuracy of semantic labelling over standard CRF's. (Hermans et al., 2014) and (Wolf et al., 2015) use a similar segmentation pipeline as ours but unlike us, they train a Random forest classifier on appearance features. They do not use height or normal pairwaise kernels in Dense CRF.

3 SYSTEM OVERVIEW

The motivation behind our model is to speed up the scene labelling process for selective search rather than resorting to an exhaustive search. The system architecture that we follow is explained in the figure 2. The input for the model are the RGB image and depth image of a indoor scene containing multiple support structures and grouped objects from Microsoft kinect sensor. The RGB image is first super-

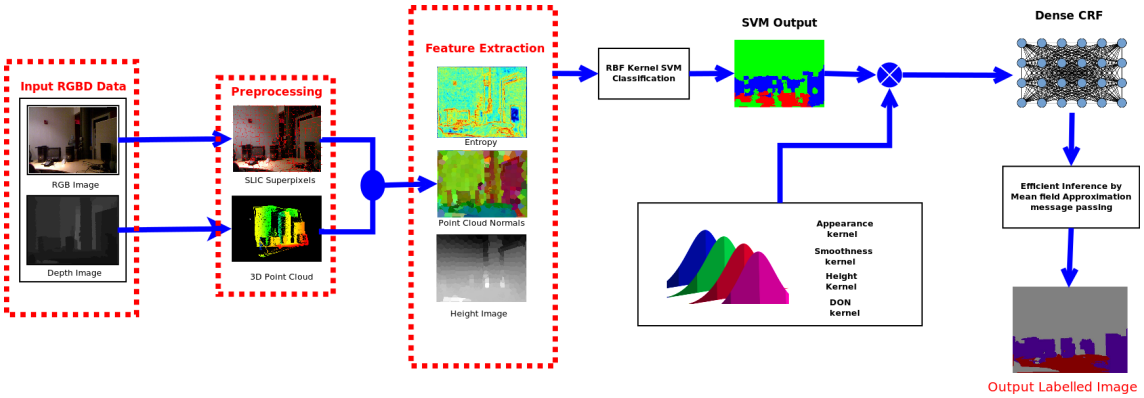

Figure 2: The Flow chart shows the stages of our system pipeline. The Input RGBD data contains the RGb image and depth image from the kinect sensor. We do a preprocessing on the input data to give the super-pixels using the SLIC algorithm and the 3D point cloud. Using the above preprocessed information, we do a feature extraction to give the entropy, point cloud normals and the height image. Once the features are extracted, we run a RBF kernel classification algorithm to gain probability of each class to be input to the CRF. We create a dense fully connected crf with multiple pairwise terms and run an mean-field based inference algorithm to accurately segment the scene into support structure and clutter.(Figure Best viewed in color and enlarged)

pixelled using SLIC algorithm(Achanta et al., 2012) and the corresponding depth data for each segment is extracted. We use the RGB and depth data to compute the normals of each segment. We train a Support Vector Machine (SVM) over these images for support structure and clutter detection. The SVM probabilities are taken as the initialization into the fully connected CRF model and inferred using the mean-field approximation.

The structure of the paper is as follows. In section 4 we give the formulation for detection of support structure using input RGB and depth images. In section 5 we use probabilities estimated from the Section 4 to formulate a CRF. The evaluation algorithm on the dataset has been explained in Section 6.

4 OBJECT CLASS DETECTION

This section explains computation of the features for the object class segmentation and its classification using the Kernel SVM. We first preprocess the input RGBD image to reduce the computational time of the algorithm.

4.0.1 Superpixels

In our approach, rather than performing classification on every pixel, we consider small regions or patches called superpixels as the basic units of classification to speed up the process. We compute superpixels over the image using Simple Linear Iterative Clustering(SLIC) method (Achanta et al., 2012). Over segmenting the image allows us to work on a few hundred data points per image rather than working with 640X480 pixels per image. Figure 3 shows an example of a super pixelated image

Figure 3: A sample scenario. (a) RGB image. (b) SLIC superpixelled image

4.1 Feature Computation

3D point cloud is extracted using the depth map and the RGB image from the kinect sensor. We use the PCL library (Rusu and Cousins, 2011) for the computation of the point cloud. We compute 3D features, which capture the geometry, shape and texture of the support structures on which objects can be placed. To support an object, we exploit the constraint that the surface should always be horizontal and ideally parallel to the ground. We have not used appearance features to avoid sensitivity to color parameters. The features we have used are listed in the table 2. We considered small feature set to enable speed in the algorithm.

4.1.1 Entropy Map

Entropy is a statistical measure of randomness that can be used to characterize the texture of an image. It is defined as

$$-\sum_{p=k_1}^{p=k_n} plog_2(p) \quad (1)$$

where k_1 k_2....k_n are the histogram counts. We take a 9X9 neighbourhood around each pixel and compute the histogram counts for each window. Secondly we compute the entropy values at every pixel in all the

three channels R,G,B and also on the depth image. Entropy map gives high values at inconsistent depth changes. We then compute the average entropy value for each superpixel.

4.1.2 Point Cloud Normals

For each superpixel, surface normal is computed at the centroid. The normal at a point is computed by approximating it to the problem of computing normal of a plane located tangent to the surface.

4.1.3 Height

We consider height(h_x) of the centroid as one of the 3D cues. Our locations of interest are flat surfaces and grouped objects raised to a certain height. The height features give us a good demarcation in segmenting the regions of interest. We can justify the selection of these features by the intuition that objects cannot be found close to the ceiling of an indoor scene and similarly support structures can not lie on the ground.

4.2 Classification

Based on the above features the support vector machine assigns a label and score to every superpixel. We use the SVM implementation from libSVM and libLinear(Chang and Lin, 2001). We tested our model on both the linear and kernel SVM. The RBF kernel performed well compared to the linear kernel. We used LibSVM (Chang and Lin, 2001) with Radial basis function kernel. The training set is small and imbalanced as the positive samples in a image will be far less than the negative samples. Appropriate cost paramater C and γ were found by searching over the grid with the cross validation data set. SVM predicts the class labels without probability information. To incorporate the SVM output into a conditional random field we follow the method given in(Wu et al., 2004), where they have extended SVM to give probability estimates. The probability estimates are given to the Conditional random fields as Unary potentials as shown in section 5.

5 CONDITIONAL RANDOM FIELD

We formulate the labelling problem as a Conditional Random field(CRF) in the image space. CRF is a graph based method generally used for segmentation problems. This is implemented using a model constituting of set of random variables $X = \{X_1, X_2, ..., X_N\}$ each taking a state from the label space $\zeta = l_1, l_2, l_3$. These random variables belong to all image pixels $i \in \nu = 1, 2, ..., N$, Let η be the neighbourhood system of the random field defined by the sets $\eta_i, \forall i \in \nu$, where η_i denotes the neighbourhood of

Table 1: Quantitative evaluation for LAB dataset. The figure shows the percentage of correctly classified support structures. The first, second and third columns presents results with the ALE, fully connected CRF and our method proposed.

ALE	FULLY-C	OURS
36%	46%	69%

Table 2: The above table shows the features used and their corresponding count

No	Feature set of the superpixel	Count
F1	Vertical position of the centroid c_z	1
F2	Vertical, Horizontal and z components of the normal: n_x, n_y, n_z	3
F3	Entropy on RGB (3 channels)	3
F4	Entropy on Depth	1

the variable X_i. Here l_1 belongs to the support structure, l_2 belongs to the clutter and the l_3 belongs to the regions not belonging to either the support structure or clutter. These energies take the form

$$E(X) = \sum_{i \in \nu} \psi_u(x_i) + \sum_{i \in \nu, j \in \eta_i} \psi_p(x_i, x_j) \qquad (2)$$

Here ψ_u is defined as the unary potential. This potential represents whether the pixel belongs to the support structure or clutter or neither. Here ψ_p represents the pairwise potential, which exploits the consistency of the label in the image space.

5.1 Unary Potential

We compute the unary potentials using the probability estimates from Section 4.2. Given k classes of the data, the goal of the classification algorithm as proposed by (Wu et al., 2004) is to estimate $p_i = P(y = i|x), i = 1, ..., k$. The algorithm follows the one-against-one approach for multi-class classification. We train the SVM for the three classes proposed earlier and input the probabilities to the CRF unary potential as follows:

$$\psi_u(x_i) = p_i \qquad (3)$$

p_i represent the probability of each label. The unary potential gives a probability of the pixel belonging to each class.

5.2 Pairwise Potential

The pairwise potential exploits the consistency of the label in the image space. We use multiple constraints to exploit the continuity of the label. The formulation of the pairwaise potential is given in (Krähenbühl and Koltun, 2012)

$$\psi_p(x_i, x_j) = \mu(x_i, x_j)[\sum_{m=1}^{k} w^{(m)} k^{(m)}(f_i, f_j)] \qquad (4)$$

Table 3: Recall accuracy on lab dataset

Method	Clutter	Support Structure	Others
ALE	**43.28**	36.71	**98.63**
FULLY-C	20.28	51.13	93.99
OURS	32.63	**59.4**	96.5

Table 4: Intersection over union accuracy on lab dataset

Method	Clutter	Support Structure	Others
ALE	**30.2**	29.99	**94.19**
FULLY-C	15.11	**39.87**	94.9
OURS	19.10	39.6	93.99

Where each $k^{(m)}$ is a Gaussian kernel , the vectors f_i and f_j are feature vectors for pixels i and j in an arbitrary feature space, $w^{(m)}$ are linear combination weights, and μ is a label compatibility function. Since our problem is a multi-class image segmentation and labelling problem, we follow (Krähenbühl and Koltun, 2012) and use contrast-sensitive kernel potentials as:

$$k(f_i,f_j) = w^{(1)}exp(\underbrace{-\frac{|p_i-p_j|^2}{2\theta_\beta^2}-\frac{|I_i-I_j|^2}{2\theta_v^2}}_{appearancekernel}) + w^{(2)}exp(\underbrace{-\frac{|p_i-p_j|^2}{2\theta_p^2}}_{smoothnesskernel}) + w^{(3)}exp(\underbrace{-\frac{|p_i-p_j|^2}{2\theta_\beta^2}-\frac{|h_i-h_j|^2}{2\theta_q^2}}_{heightkernel}) + w^{(4)}exp(\underbrace{-\frac{|p_i-p_j|^2}{2\theta_\beta^2}-\frac{|n_i-n_j|^2}{2\theta_n^2}}_{normalkernel})$$

where p_i , p_j are the positions, I_i,I_j are the intensity vectors, n_i,n_j are normal vectors, h_i,h_j are heights. $w^{(1)}$,$w^{(2)}$,$w^{(3)}$,$w^{(4)}$ are the corresponding weights for each kernel. We have fine tuned the CRF parameters by empirical evaluation of qualitative results. The appearance kernel is inspired by the observation that nearby pixels with similar appearance are likely to have same class. The smoothness kernel removes small isolated regions. In the case of support structure and clutter we exploit the constraint of height in segmenting the image. The height kernel exploits the difference of height between the labels i.e. pixels belonging to the same label need to have the same height in the depth image. Similarly pixels belonging to the same label need to have same normal orientation and is exploited using the normal kernel.

Figure 4: Qualitative Results: Our method is able to detect clutter in case of cluttered indoor scenes where support sturctures are not visible. The 2nd and 3rd columns shows the ground-truth labelling and our method labelling respectively(Figure best viewed in color and enlarged).

5.3 Inference

We follow (Krähenbühl and Koltun, 2012), which uses a mean field approximation approach for inference. In this approach we try to find a mean field approximation $Q(x)$ that minimizes the KL-divergence $D(Q||P)$ among all the distributions Q that can be expressed as a product of independent marginals, $Q(x) = \prod_i Q_i(x_i)$.

$$Q_i(x_i = l) = 1/Z_i exp\{-\psi_u(x_i) - \sum_{l' \in \mathcal{L}} \sum_{j \neq i} Q_j(x_j = l')\psi_p(x_i, y_i)\}$$

where Z_i is a constant which normalizes the marginal at pixel i. If the updates are made in sequence across pixels, the KL-Divergence is guaranteed to decrease.

6 EXPERIMENTAL RESULTS

6.1 Dataset

LAB Dataset: We consider labeling support structure and clutter in a 3D scene using a Kinect sensor. Data has been collected from 5 labs with varying compositions of the three labels, which we will be addressing as lab dataset. Each scene in the lab dataset had a image resolution of 640x480 and contained about 300,000 points of depth points. These scenes are challenging as they contained objects, which cannot be grouped or classified using existing computer vision algorithms. Manual annotation of each scene from the LAB dataset with the 3 classes is performed as shown in the second column Fig 5. We have classified the dataset into 35 traning examples and 60 testing examples.

NYU Dataset:To test the algorithm on publicly available dataset, we used the NYUv2 RBG-D dataset (Silberman et al., 2012) which comprises of 795 training and 654 testing images. NYU is one of the most widely used RGBD dataset for semantic segmentation. These images were semantically labelled to contain multiple classes. We have selected 65 testing

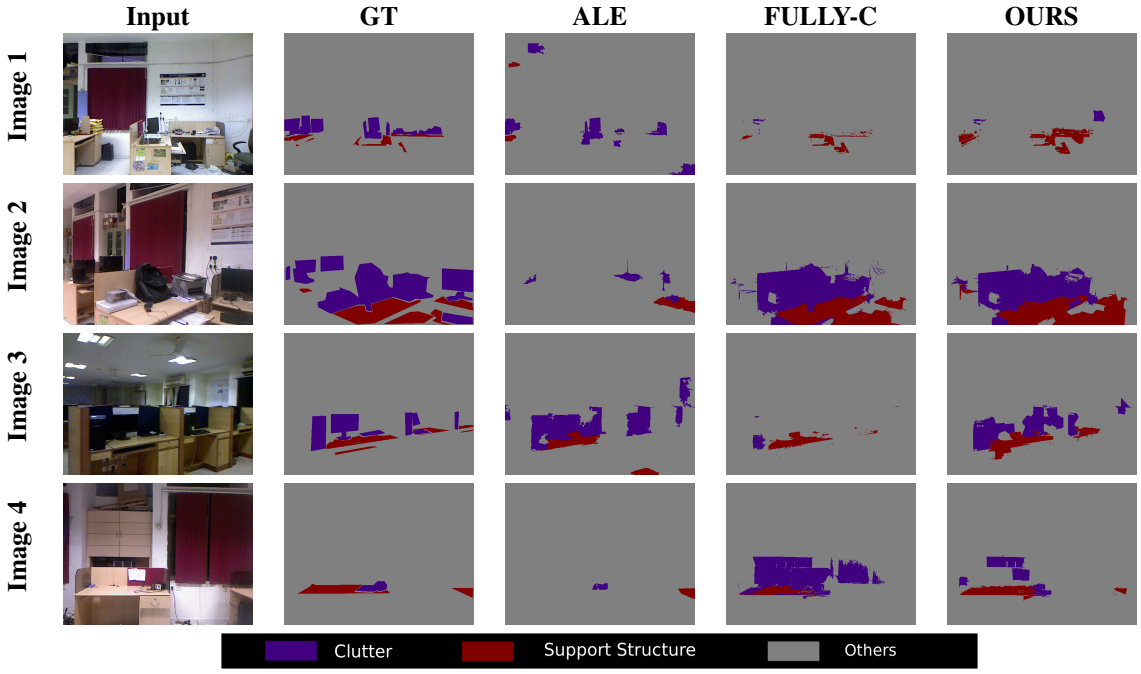

Figure 5: The columns from left to right represents the original RGB image from our lab data set with its corresponding manually annotated ground truth, predicted labels by the Super-pixel clique CRF(ALE), predicted labels by the fully connected CRF (FULLY-C). The rightmost column represents the predicted labels by our proposed CRF (OURS) based learning approach. The locations of interest here are the support structures, and clutter.

Table 5: Intersection over union accuracy on NYU

Method	Clutter	Support Structure	Others
ALE	7.40	5.18	**85.12**
FULLY-C	30.04	36.01	84.45
OURS	**31.68**	**38.14**	83.13

Table 6: Recall accuracy on NYU

Method	Clutter	Support Structure	Others
ALE	7.34	5.41	**99.0**
FULLY-C	51.61	35.68	88.99
OURS	**57.17**	**37.57**	87.14

samples containing challenging scenarios for support structure detection. The ground truth provided by NYU cannot be used for our problem as we are trying to segment support structure tops rather than the whole support struture to reduce computation. Therefore manual annotation was performed on these selected images for all the three classes.

6.2 Evaluation

In this section, we show an extensive evaluation of our algorithm on the datasets mentioned in section 6.1. We quantitatively compare our results with other state-of-the-art algorithms in scene understanding. All the currently available datasets like NYU V2, contained scenes where support structures are closer to the camera. We created the LAB dataset with challenging scenes where the support structures and clutter are relatively far away and difficult to segment compared to other publicly available datasets.

To find the best labelling algorithm over the SVM trained potentials, we have experimented with multiple CRF formulations and inference. We have used ALE (Ladicky et al., 2009) to test and train over the RGBD images using the texton features as the unary potential which we will call as ALE. We evaluated the algorithm with the fully connected CRF model(FULLY-C) from (Krähenbühl and Koltun, 2012) using the SVM potentials from section 4.2 without the additional height and normal kernels in the pairwise term, which we will be calling as FULLY-C. Finally we compared the accuracies of our proposed method with respect to the above mentioned methods and show an improvement in the accuracy of segmentation. We would like to emphasize that ALE and our system are both trained on the LAB dataset and tested on both the datasets. As ALE is trained on texture features, It performed well on LAB dataset but not on NYU dataset. Our algorithm is not trained on appearance cues to avoid sensitivity to color. This would allow us to test on variety of datasets without needing to train on all the datasets. Our algorithm is aimed at labelling only the part of the structure which supports objects for example table-top but not the whole structure. The annotations provided in the NYU dataset for structure cannot be used for the

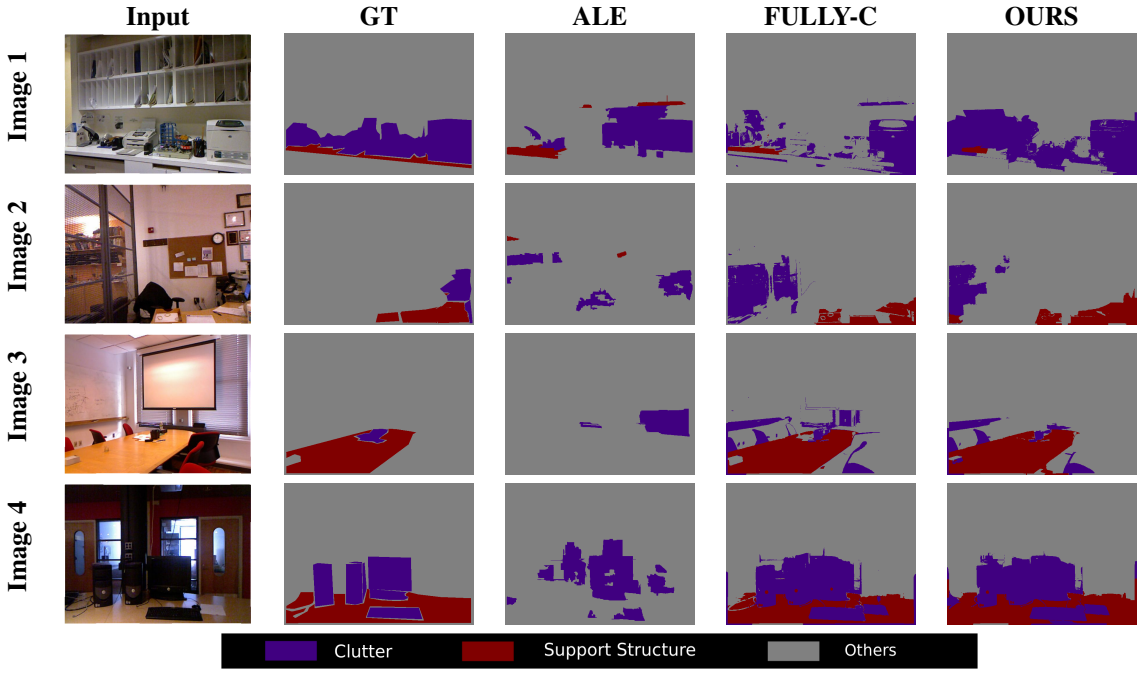

Figure 6: The columns from left to right represents the original from the NYU V2 dataset with its corresponding ground truth, predicted labels by the Super-pixel clique CRF(ALE), the dense CRF output. The rightmost column of images represent the predicted labels of our proposed method(OURS) for different NYU scenes. The locations of interest here are the support structures and clutter.

problem we are trying to address due to the aforementioned reason. Labeling only the support surfaces of the whole structure can be used as a prior for a faster object search in robotic applications. We show a quantitative evaluation of the proposed algorithm on the LAB dataset in Fig. 5. Image 2 of Fig. 5 shows how we improve upon FULLY-C by adding normal and height kernels.We show support structure level accuracies in Table 1 which gives us the information about the percentage of support structures correctly classified. We show an improvement of 10% on the support structure detection with our proposed method compared to FULLY-C and 33% increase compared to ALE.All the methods aforementioned are trained on LAB dataset. The models were tested on both LAB dataset and NYU. The standard algorithms like ALE have performed poorly on the NYU dataset and also on LAB in support structure detection because of dependency of the algorithm on the texture features. Our proposed method has scaled well on both the datasets because of its exclusive features which are not dependent on the texture of the image.

We summarize the Intersection over Union accuracies of object class segmentation on the LAB Dataset in Table 4 and on the NYU dataset in Table 5. The Intersection over union measure is defined as $TP/(TP+FP+FN)$, where TP represents True positive, FP represents False positive and FN represents the false negatives. Similarly we evaluate on the recall accuracies for each label and summarize them in Table 6 for the NYU dataset and in Table 3 for the LAB datasets. Here recall is defined as $TP/(TP+FN)$, which defines the probability of retrieval of a specific label with respect to its query. We observe that our algorithm performs better than the standard dense CRF based method in both the datasets.

We show 3D reconstruction of a lab environment using RTAB MAP((Real-Time Appearance-Based Mapping), a RGB-D SLAM approach using visual odometry in the supplementary video. We use a Kinect mounted on a P3DX robot. The system is built on ROS(Robotic Operating system). We run our algorithm on the live 3-D stream and label the support structures present in the scene. The segmented regions can help the robot for the task of object search and can also be used for path planning for faster area coverage while search for objects.

7 CONCLUSION

We have proposed an algorithm which uses geometric 3d cues and texture cues to classify the scene into support structures and clutter which will be a prior for reducing the object search space. We are proposing a generic method which will work for a variety of scenes without training on every dataset. In figure 4 we show that clutter can be used as a feature to locate regions of interest when support structures

are absent or occluded. The experiments performed on NYU show the robustness of the algorithm to drastic change in appearance of the support structures and clutter in the scene. We show 7% and 2% increase in pixelwise recall accuracies for support structure on LAB and NYU. The performance can be attributed to the consideration of geometric features from the 3-D point cloud which would otherwise not be possible if only texture cues were considered as features. Since the algorithm is fast, it is possible to implement it in a multi processor architecture for real time performance which makes it easy to use in robotic environments for region proposals. From the evaluation we conclude that our proposed method scales well across datasets. As part of the future research, We intend to segment clutter to individual objects for object recognition and formulate an Optimized path planning strategy for the robot to simultaneously explore and navigate in large rooms efficiently depending on the task it is assigned. Further by assigning the confidence for each pixel being a support structure or clutter, a more robust and optimal search strategy can be derived.

8 FUTURE WORK

We would like to extend this work further and use these region proposals for a faster object search in indoor environment. Further, we would like to investigate the performance of convolution neural networks for the same task.

REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282.

Chang, C. and Lin, C. (2001). *LIBSVM: a library for support vector machines*.

Gould, S., Fulton, R., and Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1–8. IEEE.

Gupta, S., Arbeláez, P., Girshick, R., and Malik, J. (2015). Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112(2):133–149.

Hermans, A., Floros, G., and Leibe, B. (2014). Dense 3d semantic mapping of indoor scenes from rgb-d images. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2631–2638. IEEE.

Kim, B.-s., Kohli, P., and Savarese, S. (2013). 3d scene understanding by voxel-crf. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1425–1432.

Koppula, H. S., Anand, A., Joachims, T., and Saxena, A. (2011). Semantic labeling of 3d point clouds for indoor scenes. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 244–252. Curran Associates, Inc.

Krähenbühl, P. and Koltun, V. (2012). Efficient inference in fully connected crfs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*.

Ladicky, L., Russell, C., Kohli, P., and Torr, P. H. (2009). Associative hierarchical crfs for object class image segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 739–746. IEEE.

Ladicky, L., Russell, C., Kohli, P., and Torr, P. H. (2010). Graph cut based inference with co-occurrence statistics. In *Computer Vision–ECCV 2010*, pages 239–253. Springer.

Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE.

Ren, X., Bo, L., and Fox, D. (2012). Rgb-(d) scene labeling: Features and algorithms. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2759–2766. IEEE.

Reza, M. A. and Kosecka, J. (2014). 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*.

Rusu, R. B. and Cousins, S. (2011). 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China.

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

Silberman, N. and Fergus, R. (2011). Indoor scene segmentation using a structured light sensor. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 601–608. IEEE.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, pages 746–760, Berlin, Heidelberg. Springer-Verlag.

Toyoda, T. and Hasegawa, O. (2008). Random field model for integration of local information and global information. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(8):1483–1489.

Wolf, D., Prankl, J., and Vincze, M. (2015). Fast semantic segmentation of 3d point clouds using a dense crf with learned parameters. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4867–4873. IEEE.

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. (2015). Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537.