Separating Objects and Clutter in Indoor Scenes

Salman H. Khan[1], Xuming He[2], Mohammed Bannamoun[1], Ferdous Sohel[1], Roberto Togneri[3]
[1]School of CSSE, UWA. [2]NICTA and ANU. [3]School of EECE, UWA.

Understanding real-world scenes requires reasoning both semantic and 3D structures of objects as well as the rich relationship among them. An important step towards this goal is the volumetric reasoning about generic 3D objects and their 3D spatial layout. Much progress has been made based on representing objects as 3D geometric primitives, such as cuboids. Some of the first efforts focus on the 3D spatial layout and cuboid-like objects in indoor scenes from monocular imagery [4, 7]. Owing to the complex structure of the scenes, additional depth information has recently been introduced to obtain more robust estimation [3, 5]. However, real-world scenes are composed of not only large regular-shaped structures and objects (such as walls, floor, furniture), but also irregular shaped objects and cluttered regions which cannot be represented well by object-level primitives. The overlay of different types of scene elements makes the procedure of localizing 3D objects fragile and prone to misalignment [1].

We aim to address the problem of 3D object cuboid detection in a cluttered scene. In this work, we propose to jointly localize generic 3D objects (represented by cuboids) and label cluttered regions from an RGBD image. Unlike the recent cuboid detection techniques, which consider such regions as background, our method explicitly models the appearance and geometric property of the fine-grained cluttered regions. We incorporate scene context (in the form of object and clutter) to better model the regular-shaped objects and their interaction with other types of regions in a scene.

We adopt the approach in [3] for representing an indoor scene, which models a room as a set of hypothesized cuboids and local surfaces defined by superpixels. To cope with clutters, we formulate the joint detection task using a higher-order Conditional Random Field model (CRF) on superpixels and cuboid hypotheses generated by a bottom-up grouping process. Our CRF approach extends the linear model of [3] in several aspects. First, we introduce a random field of local surfaces (superpixels) that captures the local appearance and spatial smoothness of cluttered and noncluttered regions. In addition, we improve the cuboid representation by generating two types of cuboid hypotheses, one of which corresponds to regular objects inside a scene and the other is for the main structures of a scene, such as floor and walls. Furthermore, we incorporate both the consistency between superpixel labels and cuboid hypotheses and the occlusion relation between cluttered regions and cuboid objects.

Given an RGBD image, we decompose it into a number of contiguous partitions, i.e., super-pixels: $\mathcal{S} = \{s_1, \cdots, s_J\}$. We associate a binary membership variable $m_j \in \mathbf{m}$ with each super-pixel s_j to indicate whether it belongs to the cluttered or non-cluttered regions. Similarly for each cuboid, we introduce a binary variable $c_k \in \mathbf{c}$ to indicate whether the k^{th} cuboid hypothesis is active or not. We build a CRF model on the superpixel clutter variables $\mathbf{m}$ and the object variables $\mathbf{c}$ to describe the properties of clutter, objects and their relationship in the scene. Formally, we define the Gibbs energy of the CRF as follows,

$$E(\mathbf{m},\mathbf{c}|\mathcal{I}) = E_{obj}(\mathbf{c}) + E_{sp}(\mathbf{m}) + E_{com}(\mathbf{m},\mathbf{c}), \quad (1)$$

where $E_{obj}(\mathbf{c})$, $E_{sp}(\mathbf{m})$ captures the object level and the super-pixel level properties respectively, and $E_{com}(\mathbf{m},\mathbf{c})$ models the interactions between the two levels.

We take a structural learning approach to estimate the CRF parameters from an annotated indoor dataset, which enables us to systematically incorporate more features into our model and to avoid tedious manual tuning. We use a max-margin based objective function that minimizes a loss defined on cuboid detection. Similar to [3], the (loss-augmented) MAP inference of our CRF model can be formulated as a mixed integer linear programming (MILP) formulation. We empirically show that the MILP can be globally optimized with the Branch-and-Bound method within a time of seconds to find a solution in most cases.

This is an extended abstract. The full paper is available at the CVF webpage.

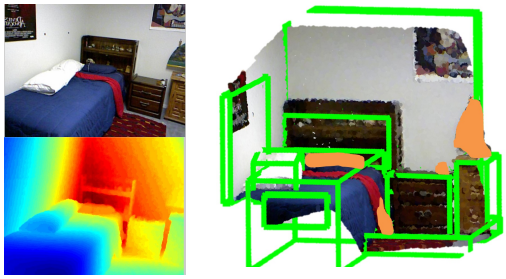

Figure 1: With a given RGBD image (*left column*), our method explores the 3D structures in an indoor scene and estimates their geometry using cuboids (*right image*). It also identifies cluttered/unorganized regions in a scene (*shown in orange*) which can be of interest for tasks such as robot grasping.

During testing, the MAP estimate of our CRF not only detects cuboid objects but also identifies the cluttered regions. For extensive evaluation we assess the performance of our approach on three tasks, namely the cuboid detection, clutter-nonclutter labelling and foreground-background segmentation. We evaluate our method on the NYU Kinect v2 dataset with augmented cuboid and clutter annotations, and demonstrate that the proposed approach achieves superior performance to the state of the art. Fig. 2 shows the performance of our approach compared to a baseline approach and the state of the art techniques [2, 3, 6]. An ablative analysis for clutter segmentation (Tab. 1) indicates that both the newly introduced features and the joint modeling contribute to the overall improvement in the clutter/non-clutter segmentation accuracy.

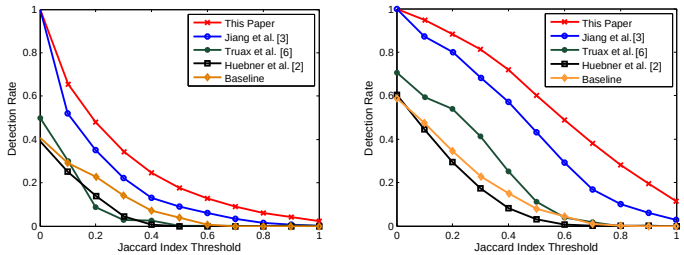

Figure 2: Jaccard Index comparisons for all annotated cuboids (*left*), for the most salient cuboid (*right*).

Method	Precision	Recall	F-Score
Super-pixel unary only	$0.43 \pm 13\%$	$0.45 \pm 11\%$	$0.44 \pm 16\%$
Unary + pairwise	$0.46 \pm 12\%$	$0.48 \pm 10\%$	$0.47 \pm 16\%$
Full model (all classes)	$0.65 \pm 9\%$	$0.68 \pm 8\%$	$0.66 \pm 12\%$
Full model (only object classes)	$0.75 \pm 6\%$	$0.71 \pm 8\%$	$0.73 \pm 10\%$

Table 1: Evaluation on Clutter/Non-Clutter Segmentation Task. Precision signifies the accuracy of clutter classification.

[1] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*. IEEE, 2009.

[2] Kai Huebner, Steffen Ruthotto, and Danica Kragic. Minimum volume bounding box decomposition for shape approximation in robot grasping. In *ICRA*, pages 1628–1633. IEEE, 2008.

[3] Hao Jiang and Jianxiong Xiao. A linear approach to matching cuboids in rgbd images. In *CVPR*. IEEE, 2013.

[4] David C Lee, Abhinav Gupta, et al. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010.

[5] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. *ICCV*, 2013.

[6] Robert Truax, Robert Platt, and John Leonard. Using prioritized relaxations to locate objects in points clouds for manipulation. In *ICRA*, pages 2091–2097. IEEE, 2011.

[7] Jianxiong Xiao, Bryan C Russell, and Antonio Torralba. Localizing 3d cuboids in single-view images. In *NIPS*, 2012.