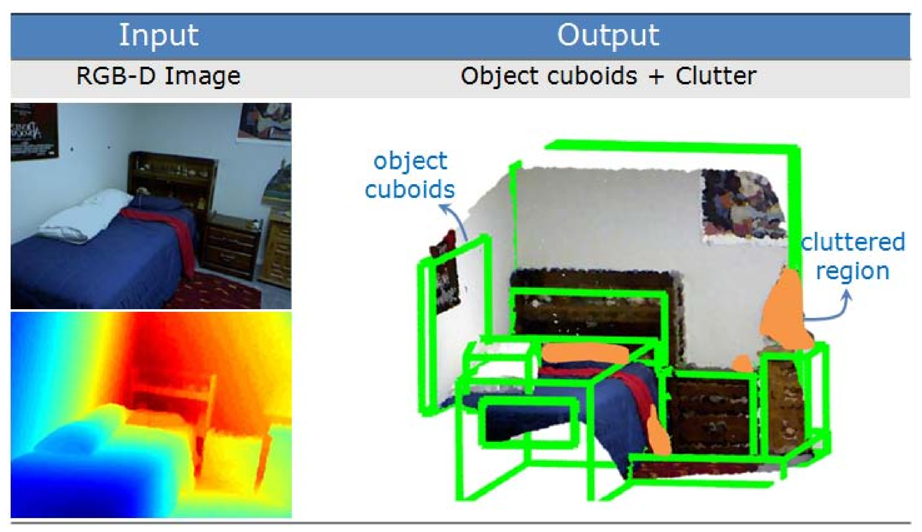
Separating Objects and Clutter in Indoor Scenes

Salman H. Khan[1], Xuming He[2], Mohammed Bennamoun[1], Ferdous Sohel[1], Roberto Togneri[3]

[1] School of CSSE, [3] School of EECE, University of Western Australia, [2] NICTA, Australian National University

Introduction

Problem Definition:



Highlights:

- We perform *'fine-grained structure categorization'* by predicting all the major objects and structures and simultaneously labeling the cluttered regions.
- The proposed CRF model incorporates a rich set of local appearance, geometric features and interactions between the scene elements.
- We take a structural learning approach with a loss of 3D localisation to estimate the model parameters from a large annotated RGBD dataset, and a mixed integer linear programming formulation for inference.

Our Approach

- The goal is to describe an RGBD image with an optimal set of cuboids and pixel-level labeling of cluttered regions.
- We represent an indoor scene as an overlay of the cluttered regions (modeled as local surfaces) and the non-cluttered regions (modeled using 3D cuboids).
- Our method for initial cuboid hypothesis generation is based on a bottom-up clustering and fitting procedure, which generates both *'object cuboids'* and *'scene bounding cuboids'*.
- We build a CRF model on the superpixel clutter variables (***m***) and the object variables (***c***) to describe the properties of clutter, objects and their relationship in the scene.

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

- The first term is defined as a combination of unary and pair-wise terms on cuboids:

$$E_{obj}(\mathbf{c}) = \sum_{k=1}^{K}\left[\psi^{u}_{obj}(c_k) + \psi^{h}_{obj}(c_k)\right] + \sum_{i<j}\psi^{p}_{obj}(c_i,c_j)$$

Cuboid feature set include:
- Volumetric occupancy feature, color consistency feature, normal consistency feature, tightness feature, support feature, geometric plausibility, cuboid size feature, cuboid MDL potential.

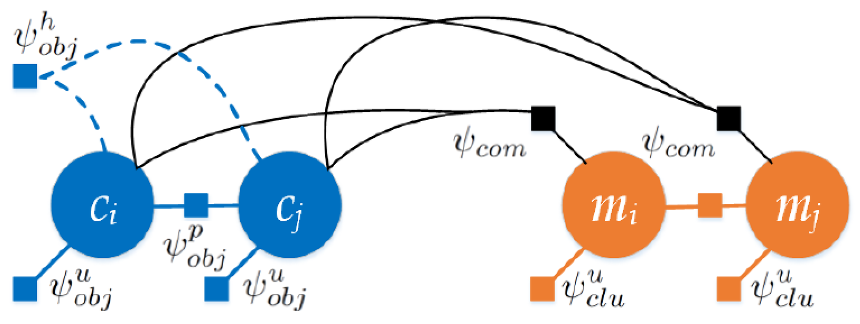


Figure: Graph structure representation for the potentials defined on the object cuboids and the cluttered/non-cluttered regions.

- View obstruction potential, cuboid intersection potential.

•Similarly at the superpixel level, the second term consists of two potential functions:

$$E_{sp}(\mathbf{m}) = \sum_{j=1}^{J}\psi^{u}_{sp}(m_j) + \sum_{(i,j)\in N_s}\psi^{p}_{sp}(m_i,m_j)$$

- The unary potential is based on the rich kernel descriptors (Bo et al., 2011).
- The pairwise potential is a contrast based Potts model.

• The third term is the compatibility constraint which enforces the consistency of the cuboid activations and the superpixel labeling:

$$E_{com}(\mathbf{m},\mathbf{c}) = \sum_{j=1}^{J}\psi_{com}(m_j,\mathbf{c})$$

It consists of two terms:
- superpixel membership potential
- superpixel cuboid occlusion potential

Inference

- We minimize the CRF energy to get optimal labeling:

$$\{\mathbf{m}^*,\mathbf{c}^*\} = \operatorname*{argmin}_{\mathbf{m},\mathbf{c}} E(\mathbf{m},\mathbf{c}|\mathcal{I}).$$

- We adopt the relaxation method in [Geiger'11, Jiang'13] and transform the minimization into a Mixed Integer Linear Program (MILP) with linear constraints.

$$\begin{aligned} &\min_{\mathbf{m},\mathbf{c},\mathbf{x},\mathbf{y},\mathbf{w},\mathbf{z}} E(\mathbf{m},\mathbf{c},\mathbf{x},\mathbf{y},\mathbf{w},\mathbf{z}|\mathcal{I}) \\ \text{s.t.}\quad & \text{linear inequality constraints in } \mathcal{LC}, \\ & m_j, c_k \in \{0,1\}, \qquad \forall j,k \\ & w_{i,j}, x_{i,j}, y_{i,j}, z_{j,k} \geq 0, \qquad \forall i,j,k \end{aligned}$$

- The MILP formulation can be solved much faster compared to the original ILP, using the branch and bound method.

- On average, 81948% variables are involved in each inference and the final MILP gap is zero for 98:5% of the cases on the whole dataset.

	Small gap	Large gap	Cuts	LP relax.
Time (sec)	$1.84 \pm 31\%$	$1.31 \pm 24\%$	$0.45 \pm 13\%$	$0.001 \pm 0.4\%$
Det. Rate	26.8%	26.1%	24.4%	19.9%

Table: Inference running time comparisons for variants of MILP formulation.

Cuboid Detection Results

Figure: Comparison of our results (*3rd row*) with the state of the art technique [Jiang et al., 2013] (*2nd row*) and Ground Truth (*1st row*). The detected cuboids are shown in color on top of the original RGB image.

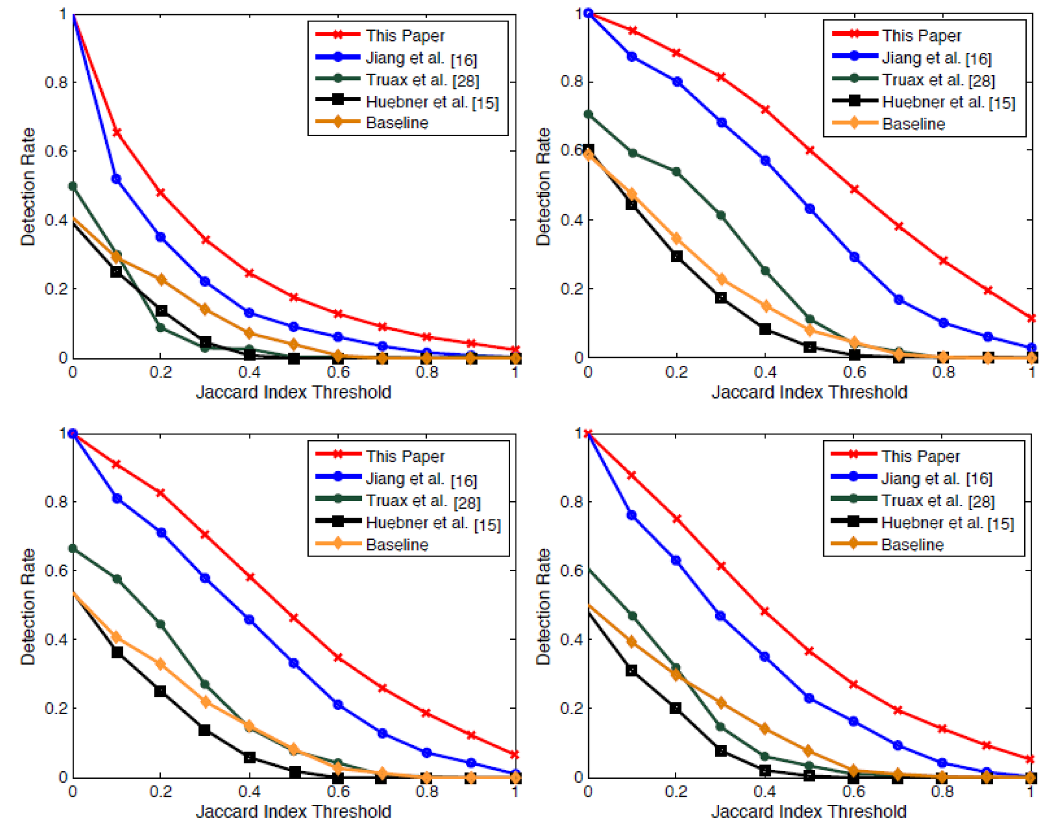


Figure: Jaccard Index comparisons for all annotated cuboids (*top left*), for the most salient cuboid (*top right*), for top two salient cuboids (*bottom left*) and top three salient cuboids (*bottom right*).

Figure: An ablation study on the model potentials/features for the cuboid detection task at the 40% JI threshold.

Method	Accuracy
Unary cuboid cost of Jiang [16]	6.5%
Our unary cuboid cost only	8.8%
Our unary + pairwise cuboid cost only	19.4%
Our full model	26.1%

Figure: Examples of detection errors.

Parameter Learning

- For parameter learning, we apply the structural SVM framework with margin re-scaling and 3D localization loss given by:

$$\Delta(\mathbf{t}^{(n)},\mathbf{t}) = \sum_i\left(1 - \frac{|o_i^{(n)} \cap o_i|}{|o_i^{(n)} \cup o_i|}\right)$$

- The formulation uses the cutting plane algorithm [Joachims'07] to search the optimal parameter setting.

Segmentation Results

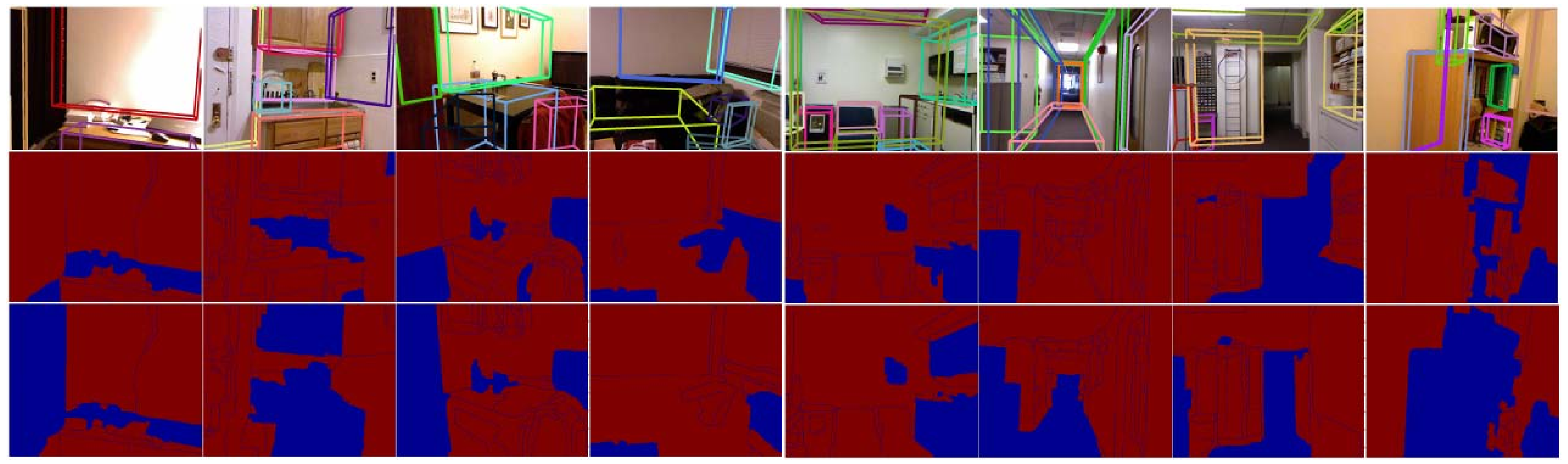


Figure: Our method is able to accurately detect cuboids in the case of cluttered indoor scenes (1st row). The 2nd and 3rd rows show our clutter labelling and the ground-truth labelling on superpixels, respectively. In the bottom two rows, red color represents non-clutter while blue color represents clutter.

Method	Precision	Recall	F-Score
Superpixel 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: Evaluation on Clutter/Non-Clutter Segmentation Task. Precision signifies the accuracy of clutter classification.

Eval. Criterion	CPMC [6]		This Paper	
	Pre.	Rec.	Pre.	Rec.
Most salient obj.	$0.83 \pm 11\%$	0.79 ± 12	$0.85 \pm 15\%$	$0.82 \pm 15\%$
Top 2 salient obj.	$0.77 \pm 12\%$	0.73 ± 14	$0.81 \pm 16\%$	$0.79 \pm 16\%$
Top 3 salient obj.	$0.69 \pm 15\%$	0.66 ± 17	$0.79 \pm 21\%$	$0.76 \pm 19\%$
All objects	$0.54 \pm 17\%$	0.51 ± 20	$0.73 \pm 23\%$	$0.69 \pm 21\%$

Table: Evaluation on Foreground/Background Segmentation Task. Precision signifies the accuracy of clutter classification.