Supplementary Material: Three-Dimensional Object Detection and Layout Prediction using Clouds of Oriented Gradients

Zhile Ren   Erik B. Sudderth
Department of Computer Science, Brown University, Providence, RI 02912, USA

1. Proposing Layout Candidates

We predict floors and ceilings as the 0.001 and 0.999 quantiles of the 3D points along the gravity direction. After that, we only need to propose wall candidates that follows Manhattan structure.

We discretize orientation into 18 evenly spaced angles between 0 and 180°. For each orientation, the frontal wall is bounded by the 0.99 quantiles of the farthest points and the back wall is bounded by camera location. Because the layout candidates should follow a Manhattan structure, the left/right wall should be orthogonal to the frontal wall and should be bounded by the 3D points. We further discretize along the width and depth direction by 0.1m, and propose candidates that capture at least 80% of all 3D points. For typical scenes, there are 5,000-20,000 layout hypotheses.

2. Contextual Features

Here, we give a detailed explanation of the contextual features we use to model object-object and object-layout relationships in the second stage cascaded classifier.

For completeness, we define the notations again. For an overlapping pair of detected bounding boxes B_i and B_j, we denote their volumes as $V(B_i)$ and $V(B_j)$, their volume of their overlap as $O(B_i, B_j)$, and the volume of their union as $U(B_i, B_j)$. We characterize their geometric relationship via three features: $S_1(i,j) = \frac{O(B_i,B_j)}{V(B_i)}$, $S_2(i,j) = \frac{O(B_i,B_j)}{V(B_j)}$, and the IOU $S_3(i,j) = \frac{O(B_i,B_j)}{U(B_i,B_j)}$. To model object-layout context [1], we compute the distance $D(B_i, M)$ and angle $A(B_i, M)$ of cuboid B_i to the closest wall in layout M.

The first-stage detectors provide a most-probable layout hypothesis, as well as a set of detections (following non-maximum suppression) for each category. For each bounding box B_i with confidence score z_i, there may be several bounding boxes of various categories $c \in \{1, 2, ..., C\}$ that overlap with it. We let i_c be the instance of category c with the maximum confidence score z_{i_c}. The features ψ_i for bounding box B_i are then as follows:

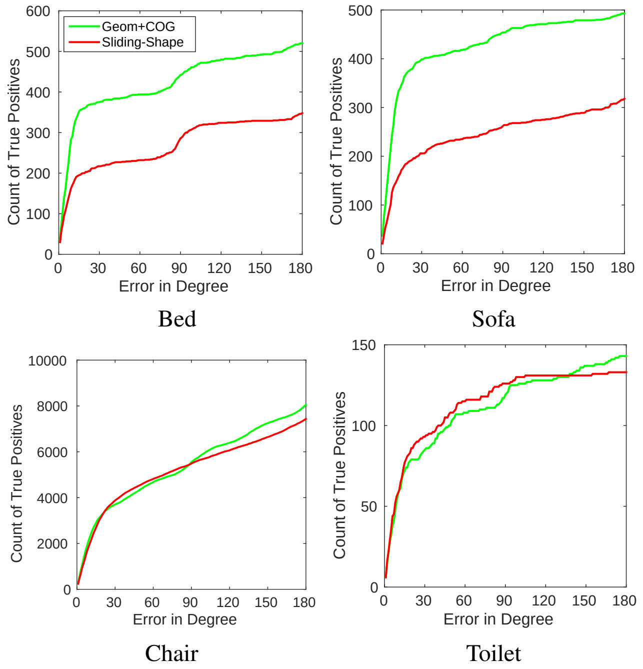
1. Constant bias feature, and confidence score z_i from the first-stage detector.
2. For $m \in \{1, 2, 3\}$ and $c \in \{1, 2, ..., C\}$, we calculate $S_m(i, i_c)$, $S_m(i, i_c) \cdot z_{i_c}$, $S_m(i, i_c) \cdot z_i$ and concatenate those numbers.
3. For $c \in \{1, 2, ..., C\}$, we calculate the difference in confidence score from each first-stage detector, $z_i - z_{i_c}$, and concatenate those numbers.
4. For $D(B_i, M)$, we consider radial basis functions of the form
$$f_j(x) = \exp(-\frac{(x-\mu_j)^2}{2\sigma^2}) \tag{1}$$
For a typical indoor scene, the largest object-to-wall distance is usually less than 5m, therefore we space the basis function centers μ_j evenly between 0 and 5 with step size 0.5, and choose $\sigma = 0.5$. We expand $D(B_i, M)$ using this radial basis expansion.
5. The absolute value of the cosine of $D(B_i, M)$: $|\cos(D(B_i, M))|$

To model the second-stage layout candidates, we select the bounding box i_c with the highest confidence score z_{i_c} from the first-stage classifier in each category $c \in \{1, 2, ..., C\}$, and use the following features for layout M_i with confidence score z'_i:

1. All the features used in the first-stage to model M_i using *Manhattan Voxels*.
2. For $c \in \{1, 2, ..., C\}$, we calculate the radial basis expansion for $D(B_{i_c}, M_i)$, and its product with z'_i and z_{i_c}.
3. For $c \in \{1, 2, ..., C\}$, we calculate the absolute value of the cosine of $D(B_{i_c}, M_i)$: $|\cos(D(B_{i_c}, M_i))|$, $|\cos(D(B_{i_c}, M_i))| \cdot z'_i$ and $|\cos(D(B_{i_c}, M_i))| \cdot z_{i_c}$.
4. For $c \in \{1, 2, ..., C\}$, we calculate the difference in confidence score from each first-stage detector, $z'_i - z_{i_c}$, and concatenate those numbers.

3. Additional Experimental Results

Orientation Besides evaluating detection based on intersection-over-union score, we can also evaluate the accuracy of the predicted orientations. We plot the cumulative counts of true positive detections whose orientation error in degrees is less than certain thresholds in Fig. 1. Using geometric feature and COG, our detector is able to detect more true positives than sliding-shape [3], while maintaining comparable accuracy in orientation estimation. Those curves are not related to the precision-recall curves as we are only evaluating on true positives.



Bed

Sofa

Chair

Toilet

Figure 1. Quantification of orientation estimation accuracy for the four object categories considered by [2].

Additional Layout Prediction results Besides the two examples for layout prediction shown in the paper, in figure 2 we show some additional results to demonstrate the effectiveness of *Manhattan Voxels*.

References

[1] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, pages 1417–1424. IEEE, 2013. 1

[2] S. Song, L. Samuel, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*. IEEE, 2013. 2

[3] S. Song and J. Xiao. Sliding shapes for 3d object detection in depth images. In *ECCV*, pages 634–651. Springer, 2014. 2

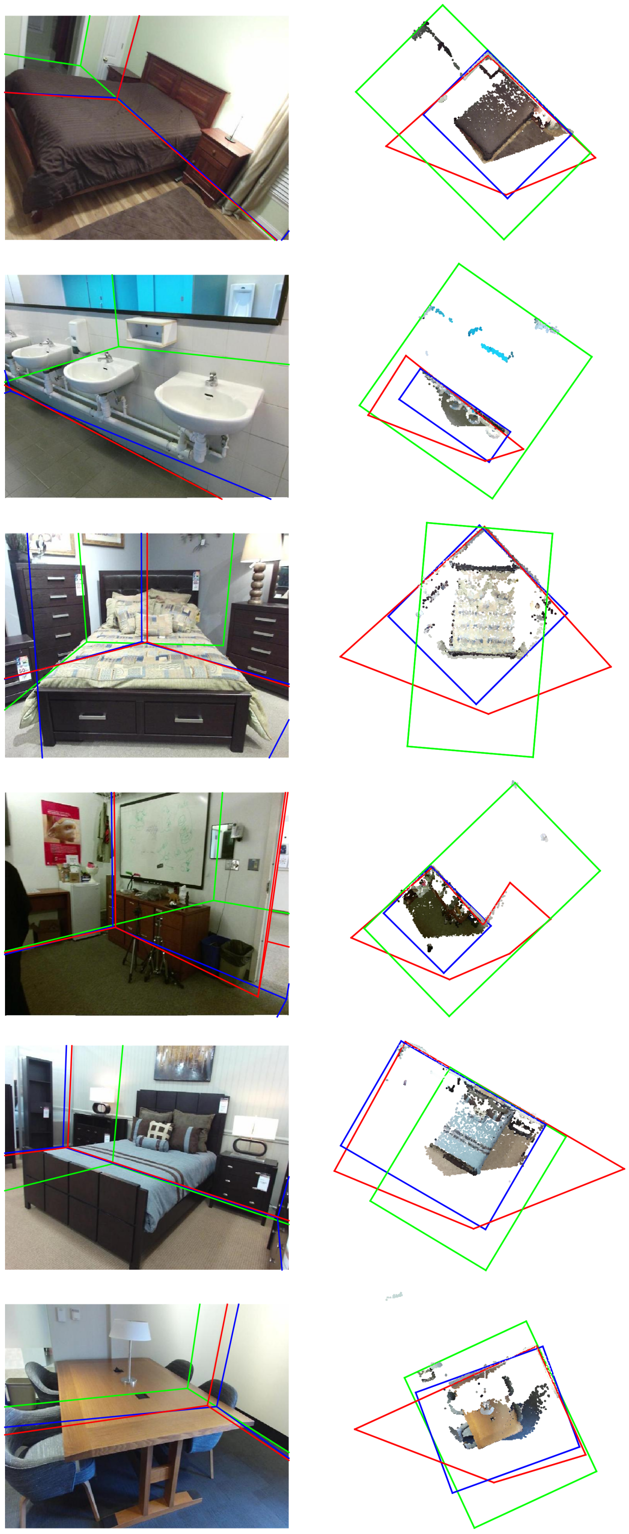


Figure 2. Comparison of our Manhattan voxel 3D layout predictions (blue) to the SUN RGB-D baseline ([2], green) and the ground truth annotations (red). Our learning-based approach is less sensitive to outliers and degrades gracefully in cases where the true scene structure violates the Manhattan world assumption.