Object Detection with Geometrical Context Feedback Loop

Min Sun*
sunmin@umich.edu
Sid Ying-Ze Bao*
yingze@eecs.umich.edu
Silvio Savarese
silvio@eecs.umich.edu

Department of EELS
University of Michigan
Ann Arbor, USA
*indicates equal contributions

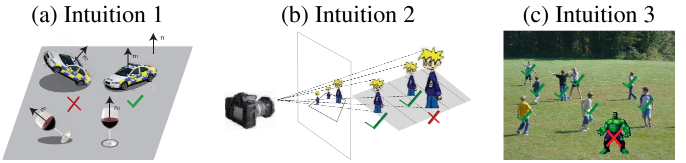

Figure 1: List of intuitions in our paper and comparison with related works. (a) Intuition 1: Rigid objects typically lie up-right on the supporting plane. (b) Intuition 2: Under the perspective camera model, the size of an object in the 2D image is an inversely proportional function of its distance to the camera when the object pose is fixed. (c) Intuition 3: The statistics describing the 2D appearance (features, texture, etc...) of foreground objects are different enough from those describing the 2D appearance of the supporting surfaces.

In this work, we present a new way to establish the contextual relationship between objects and the scene geometric structure. Specifically, we are interested in modeling the relationship between: i) **objects and their supporting surface geometry**: Geometrical configuration of objects in space is tightly connected with the geometry (orientation) of the surfaces holding these objects (Fig. 1 - Intuition 1); ii) **objects and observer geometry**: Object appearance properties such as the scale and pose are directly related to the observer intrinsic (focal length) and extrinsic properties (camera pose and location) (Fig. 1 - Intuition 2); iii) **objects and supporting regions**: The statistics describing the 2D appearance (features, texture, etc...) of foreground objects are different from those describing the 2D appearance of the supporting surfaces (Fig. 1 - Intuition 3).

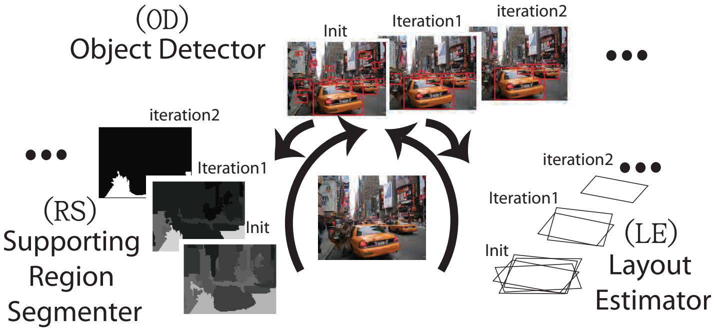

Figure 2: The Context Feedback Loop. We demonstrate that scene layout estimation and object detection can be part of a joint inference process. In this process a supporting region segmentation module (RS) and a scene layout estimation module (LE) provides evidence so as to improve the accuracy of an object detector module (OD). In turn, the OD module enables a more robust estimation of the scene layout (supporting planes orientation, camera viewing angle) and improves the localization of the supporting regions.

Specifically, we propose a new coherent framework for joint object detection, 3D layout estimation, and object supporting region segmentation from a single image. Our approach is based on the mutual interactions among three novel modules (see Fig. 2): i) object detector (OD); ii) scene 3D layout estimator (LE); iii) object supporting region segmenter (RS). The interactions between such modules capture the contextual geometrical relationship between objects, the physical space including these objects, and the observer. An important property of our algorithm is that the object detector module, developed based on [3], is capable of adaptively changing its confidence in establishing whether a certain region of interest contains an object (or not) as new evidence is gathered about the scene layout. This enables an iterative estimation procedure where the detector becomes more and more accurate as additional evidence about a specific scene becomes available (Fig. 3).

Extensive quantitative and qualitative experiments are conducted on a new in-house dataset [3] and two publicly available datasets [1, 2], and demonstrate competitive object detection, 3D layout estimation, and object supporting region segmentation results. Here we use the results from the in-house dataset as examples. Fig. 4(a) shows the overall Precision Recall curve of our system. Fig. 4(b) and Table 1 demonstrate that the feedback loop is effective in improving i) 3D layout estimation and supporting region segmentation, ii) the object detection performance. Similar results are observed in other two datasets [1, 2].

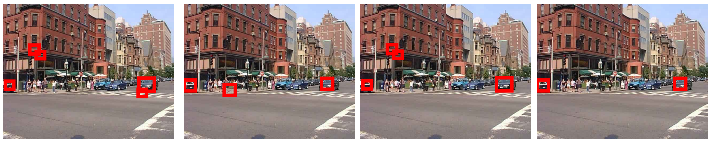
(a) OD (b) OD + LE (c) OD + RS (d) Full system

Figure 3: Interactions between different modules contribute to improve the detection performance. Panels show the results of the baseline detection (a), joint detection and 3D layout estimation (b), joint detection and supporting region segmentation (c), and our full system (d).

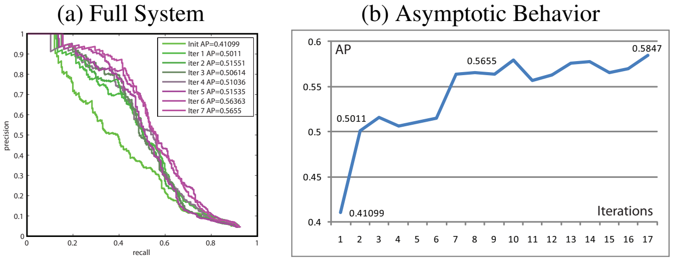

Figure 4: Detection performance using precision-recall measurement. Panel (a) shows the results when all modules (OD,LE,RS) are used in the loop from iteration 1 to 7. Panel (b) shows that the performance of our full system asymptotically converges to a steady state.

Loop Iterations	e_n (radius)	e_η (%)	e_f (%)	e_{seg}^{FA} (%)	e_{seg}^{MS} (%)
First Loop	0.125	25.9	13.2	2.07	51.2
Final Loop	0.118	21.2	11.7	1.79	56.9

Table 1: Estimation errors of surface layout parameters (n, η, f), and supporting regions. The first three columns show the errors of the estimated surface normal e_n, camera height e_η, and surface normal e_f. Each of the errors are defined as follows: $e_n = \arccos(n_{est} n_{gt})$, $e_\eta = \frac{\|\eta_{est} - \eta_{gt}\|}{\eta_{est}}$, and $e_f = \frac{\|f_{est} - f_{gt}\|}{f_{est}}$, where subscript labels *est* and *gt* indicate estimated and ground truth values respectively. The last two columns report two types of segmentation errors: e_{seg}^{FA} and e_{seg}^{MS} are the amount to which the segmenter mistakenly predicts a foreground region as supporting region and the segmenter misses the truth supporting region, respectively. In detail, let I_P denotes the supporting region predicted by our model, I_{SR} denotes the ground-truth supporting region, and I_F denotes the ground truth foreground objects. We define $e_{seg}^{FA} = \frac{|I_P \cap I_F|}{|I_F|}$ and $e_{seg}^{MS} = \frac{|I_P \cap I_{RS}|}{|I_{RS}|}$, where $|\bullet|$ counts the pixel number. The smaller e_{seg}^{FA}, the lower the false alarm rate is for confusing foreground pixels as background. The higher e_{seg}^{MS}, the larger the area our algorithm can classify as supporting region. All five types of errors are further reduced as the number of iterations increases (the table reports results for the 1st and 7th iteration).

[1] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Putting objects in perspective. In *CVPR*, 2006.

[2] Erik B. Sudderth, Antonio Torralba, William T. Freeman, and Alan S. Willsky. Describing visual scenes using transformed objects and parts. In *IJCV*, 2008.

[3] Min Sun, Gary Bradski, Bing-Xin Xu, and Silvio Savareses. Depth-encoded hough voting for coherent object detection, pose estimation, and shape recovery. In *ECCV*, 2010.