Data-Driven 3D Voxel Patterns for Object Category Recognition

Yu Xiang[1,2], Wongun Choi[3], Yuanqing Lin[3], and Silvio Savarese[1]
[1]Stanford University, [2]University of Michigan at Ann Arbor, [3]NEC Laboratories America, Inc.

Despite the great progress achieved in recognizing objects as 2D bounding boxes in images, it is still very challenging to detect occluded objects and estimate the 3D properties of multiple objects from a single image. Consider Fig. 1-top for instance, where cars occupy just a small portion of the image and most of them are heavily occluded by other cars. Except for a few exceptions [5, 9], most of the recent object detection methods have hard time in parsing out the correct configuration of objects from this kind of imagery.

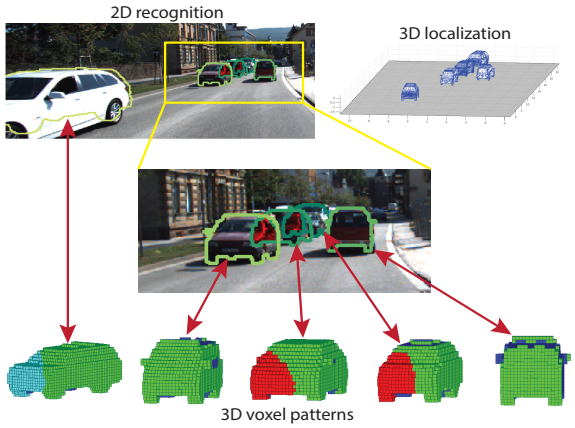

Figure 1: By introducing the 3D voxel patterns, our recognition framework is able to not only detect objects in images, but also segment the detected objects from the background, estimate the 3D poses and 3D shapes, localize them in the 3D space, and even infer the occlusion relationship among them. Green, red and cyan voxels are visible, occluded and truncated respectively.

In this paper, we present a novel recognition pipeline that addresses the key challenges: i) it goes beyond 2D bounding box detection and is capable of estimating 3D properties of multiple detected objects such as 3D pose as well as their depth ordering from the observer; ii) it is designed to handle situations where objects are severely occluded by other objects or truncated because of a limited field of view; iii) it is capable of accurately estimating the occlusion boundaries of each objects as well as inferring which portions of the object are occluded or truncated and which are not (see Fig. 1).

At the foundation of our recognition pipeline is the newly proposed concept of *3D Voxel Pattern* (3DVP). A 3DVP is a novel object representation that jointly captures key object properties which relates: i) appearance – the RGB luminance values of the object in the image; ii) 3D shape – the 3D geometry of the object expressed as a collection of 3D voxels; iii) occlusion masks – the portion of the object that is visible or occluded because of self-occlusions, mutual occlusions and truncations. Our approach follows the idea that luminance variability of the objects in the image due to intra-class changes and occlusions can be effectively modeled by learning a large dictionary of such 3DVPs whereby each 3DVP captures a specific shared "signature" of the three properties listed above (appearance, 3D shape and occlusions).

In our recognition pipeline, we train a bank of ACF detectors [1] using our dictionary of 3DVPs. Because the 3DVPs retain shared properties about the object (specifically, 3D shape and occlusion masks), these can be transferred during the detection regime so as to recover the 2D segmentation mask of the object, its 3D pose as well as which portions of the objects are occluded and which are visible. Finally, and most critically, we use these properties to reason about object-object interactions and infer which object is an "occluder" and which is an "occludee". This in turn helps adjusting the confidence values of the detectors

	Object Detection (AP)			Orientation (AOS)		
Methods	Easy	Moderate	Hard	Easy	Moderate	Hard
ACF [1]	55.89	54.74	42.98	N/A	N/A	N/A
DPM [2]	71.19	62.16	48.43	67.27	55.77	43.59
DPM-VOC+VP [8]	74.95	64.71	48.76	72.28	61.84	46.54
OC-DPM [9]	74.94	65.95	53.86	73.50	64.42	52.40
SubCat [6]	81.94	66.32	51.10	80.92	64.94	50.03
AOG [5]	84.36	71.88	59.27	43.81	38.21	31.53
SubCat [7]	84.14	75.46	59.71	83.41	74.42	58.83
Regionlets [10]	84.75	**76.45**	59.70	N/A	N/A	N/A
Ours	**87.46**	75.77	**65.38**	**86.92**	**74.59**	**64.11**

Table 1: Car detection and orientation estimation results of different methods on the KITTI test set. We show the results of 227 3DVP clusters for **Ours**. More comparisons are available at [3].

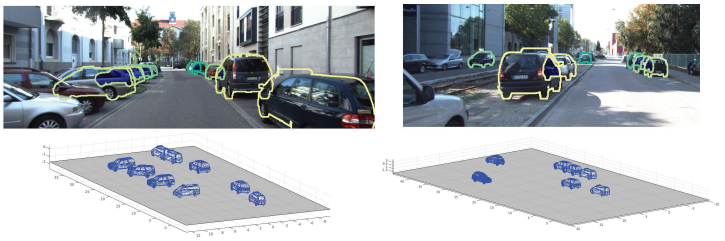

Figure 2: 2D recognition and 3D localization results on the KITTI test set. Blue regions in the images are the estimated occluded areas.

We trained and tested our approach using the KITTI detection benchmark [4] and focused on recognizing cars and estimating their 3D properties. Table 1 shows our car detection and orientation estimation results on the KITTI dataset and the comparison with the state-of-the-art methods. Fig. 2 shows some 2D recognition and 3D localization results from our method on the KITTI test set. We also evaluated our method using the outdoor-scene dataset proposed in [11]. Our extensive experimental evaluation shows that: i) our approach based on 3D voxel patterns produces significant improvement over state-of-the-art results for car detection and 3D pose estimation on KITTI; ii) our approach allows us to accurately segment object boundaries and infer which areas of the objects are occluded and which are not; we demonstrate that our segmentations results are superior than several baseline methods; iii) our approach allows us to localize objects in 3D and thus infer the depth ordering of the object from the camera's viewpoint.

[1] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *TPAMI*, 36(8):1532–1545, 2014.

[2] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645, 2010.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Kitti object detection benchmark. http://www.cvlibs.net/datasets/kitti/eval_object.php. Accessed: 2015-03-18.

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[5] Bo Li, Tianfu Wu, and Song-Chun Zhu. Integrating context and occlusion for car detection by hierarchical and-or model. In *ECCV*, pages 652–667. 2014.

[6] Eshed Ohn-Bar and Mohan M Trivedi. Fast and robust object detection using visual subcategories. In *CVPRW*, pages 179–184, 2014.

[7] Eshed Ohn-Bar and Mohan M. Trivedi. Learning to detect vehicles by clustering appearance patterns. *T-ITS*, 2015.

[8] Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Multi-view and 3d deformable part models. *TPAMI*, 2015.

[9] Bojan Pepikj, Michael Stark, Peter Gehler, and Bernt Schiele. Occlusion patterns for object class detection. In *CVPR*, pages 3286–3293, 2013.

[10] Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. In *ICCV*, pages 17–24, 2013.

[11] Yu Xiang and Silvio Savarese. Object detection by 3d aspectlets and occlusion reasoning. In *ICCVW*, pages 530–537, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.