High-Resolution 3D Layout from a Single View

M. Zeeshan Zia[1], Michael Stark[2], and Konrad Schindler[1]

[1] Photogrammetry and Remote Sensing, ETH Zürich, Switzerland
[2] Stanford University and Max Planck Institute for Informatics

Abstract

We explore 3D layout estimation from a monocular image using detailed 3D object class models. We leverage the high geometric resolution provided by such models to reason about interactions among multiple object instances and demonstrate superior performance to using coarse 3D bounding box level hypotheses.

1. Introduction

Recent advances in computer vision and allied fields have enabled researchers to revisit attempts at 3D scene-level understanding [4, 5] from the early days of computer vision, leveraging on advances in local shape features, discriminative classifiers, and efficient approximate inference. In addition to 3D layout estimates, they also demonstrate superior 2D recognition performance compared to 2D only approaches. However, these attempts have been limited to rather coarse models leaving open the question of whether even finer-grained models can be beneficial to scene-level understanding. Here we investigate the utility of using a detailed 3D object class model in estimating 3D scene layout. Specifically we demonstrate how a deformable wireframe model of object classes allows reasoning about occlusions among modeled object instances at the level of detail of individual wireframe vertices. Further, we enforce all object instances in a given scene to lie on a common ground plane, and show that this constraint gives further improvements in accuracy. Both these constraints are compared against coarser 3D bounding box level estimates and shown to outperform it by significant margins.

2. 3D Scene Model

Our 3D scene model consists of a set of 3D deformable objects, augmented with occluder masks, and a common ground plane. What distinguishes the model from previous work[8] is a much more expressive solution space that allows one to reason about the locations, shapes and interactions of objects, at the level of individual vertices and faces (Fig. 2). We express the likelihood of a particular scene hypothesis in that space as a combination of per-object likelihoods, computed with an existing model [10], and perform sample-based inference to find the best hypothesis.

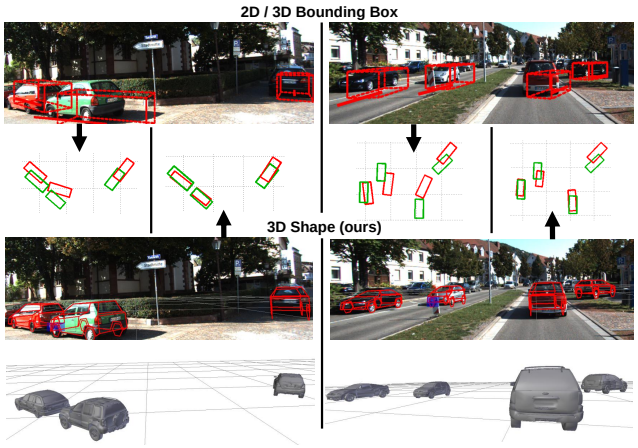

Figure 1. *Top:* Coarse 3D object bounding boxes derived from 2D bounding box detections. *Bottom:* our fine-grained 3D shape model fits improve 3D localization (see bird's eye views).

2.1. Detailed Object Class Models

We utilize explicit representations of global object geometry [7, 8] that are better suited for estimating 3D object shape and pose than coarser models. Specifically, in the tradition of *active shape models*[2] we learn a deformable 3D wireframe from annotated 3D CAD models as in [8]. The wireframe model is defined through an ordered collection of n vertices in 3D-space, chosen at salient points on the object surface in a fixed topological layout. The local appearance model comprises of a multi-class Random Forest, trained on rendered multi-view patches of 3D CAD models at the annotated vertex locations, as well as random background examples. These patches are encoded using a dense variant of the shape context descriptor. We employ the Random Forest classifier in a sliding-window fashion, searching over image locations and scale to detect the locations of object parts in test images.

2.2. Occluder Masks

We assume occluders to block the view onto a spatially connected region of the object. Due to the object being modeled as a sparse collection of parts, occluders can only be distinguished if the visibility of at least one part changes, which further reduces the space of possible occluders. Thus, one can well approximate the set of all occluders by a discrete set of occlusion masks. The set of occluder masks can uniformly model both types of occlusions

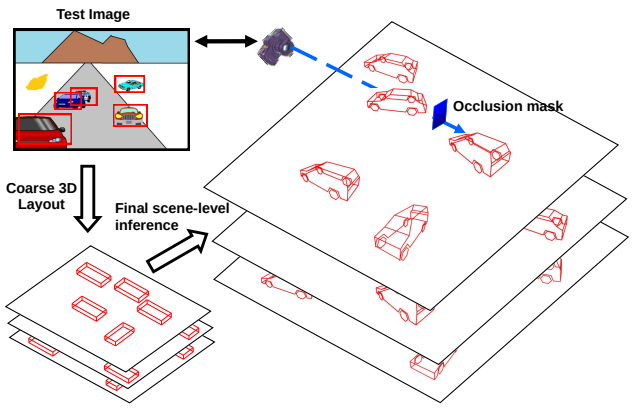

Figure 2. 3D Scene Model

that we are concerned with here: (i) occlusions corresponding to missing image evidence, and (ii) occlusion due to other modeled objects. In our inference, we search through the set of occlusion masks, assigning a fixed low score to all parts falling inside the occluded portion of the mask instead of the detection score. The low score corresponds to a weak uniform prior that prefers parts to be visible and counters the bias to "hide behind the occluder".

2.3. Deterministic Occlusion Reasoning

Since by construction we recover the 3D locations and shapes of multiple objects in a common frame, we can calculate whether a certain object instance is occluded by any other modeled object instance in our scene. This is calculated efficiently by casting rays to all (not self-occluded) vertices of the object instance, and checking if a ray intersects any other object in its path before reaching the vertex. This deterministically estimates which parts of the object instance are occluded by another modeled object in the scene, allowing us to choose an occluder mask that best represents the occlusion.

2.4. Common Ground Plane

We constrain all the object instances to lie on a common ground plane, as often done for street scenes. This assumption usually holds and drastically reduces the search space for possible object locations (2 degrees of freedom for translation and 1 for rotation, instead of $3 + 3$). Moreover, the consensus for a common ground plane stabilizes 3D object localization.

2.5. Initialization, 3D Lifting, and Inference

We formulate the 3D layout estimation problem as an objective function. This function [9] is highly non-convex including discrete variables and high dimensional. We thus employ a sampling based optimization procedure [6] to reach a reasonable local maxima. However, the optimization requires a good initialization.To obtain this initialization,we employ a clustering based procedure to merge together activations [1] of partial object detectors, obtaining coarse 2D bounding box and discrete viewpoint estimates. Since, we reason in a fixed, camera-centered 3D coordinate frame, the initial detections are directly lifted to 3D space, by casting rays through 2D bounding box centers and instantiating objects on these rays, such that their reprojections are consistent with the 2D boxes and discrete viewpoint estimates, and reside on a common ground plane. The inference then amounts to refining this coarse 3D layout by finding good local maxima of the objective function.

3. Evaluation

We evaluate object localization in 3D metric space as well as 3D pose estimation on the challenging KITTI dataset [3] of street scenes.

Quantitatively, our full system offers superior 3D pose estimation, correctly localizing 44% of the detected cars up to 1 m, compared to 40% when deterministic occlusion reasoning is disabled, and 26% when common ground plane is also turned off. If only coarse 3D estimates are used, without any physical interaction reasoning, we are only able to correctly localize 21% of the detected cars. Fig. 1 visualizes the 3D pose estimates from our full system as compared to coarse 3D bounding box detections. These results support physically grounded reasoning on the basis of detailed object class models for 3D scene understanding, beyond coarse estimates of independent objects.

4. Conclusions

We investigated how detailed 3D object class models can be used to perform accurate reasoning about 3D scene layout. We used fine-grained estimates of object shapes to model interactions among multiple object instances and demonstrated superior performance to coarse 3D bounding box hypotheses.

References

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. *ICCV 2009*.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models, their training and application. *CVIU*, 61(1), 1995.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. *CVPR'12*.

[4] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. *ECCV'10*.

[5] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. *ECCV'10*.

[6] M. Leordeanu and M. Hebert. Smoothing-based optimization. *CVPR 2008*.

[7] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. *CVPR 2012*.

[8] M. Z. Zia, M. Stark, B. Schiele, and K. Schindler. Detailed 3d representations for object recognition and modeling. *PAMI*, 2013.

[9] M. Z. Zia, M. Stark, and K. Schindler. Are Cars Just 3D Boxes? – Jointly Estimating the 3D Shape of Multiple Objects. *CVPR'14*.

[10] M. Z. Zia, M. Stark, and K. Schindler. Explicit occlusion modeling for 3d object class representations. In *CVPR*, 2013.