Towards Scene Understanding with Detailed 3D Object Representations

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 detailed 3D representations of rigid, deformable 3D object classes, amenable to both estimating the 3D shape and pose of individual objects, and to scene understanding tasks such as reasoning jointly about multiple object instances or about scene-level interactions such as occlusions.

1. Introduction

Visual scene understanding requires good quality object detections as input so that higher-level reasoning about interactions among objects and between objects and the scene can be performed. Over the last decade, object class detectors have attained reasonable efficiencies at finding instances of a variety of object classes in images [2, 1]. However, the object hypotheses these detectors provide as output, *i.e*. 2D bounding boxes along with viewing angles discretized into a few bins, are overly crude. We believe that such simplistic representations hamper subsequent higher-level reasoning about objects and their relations, since they convey very little information about the objects' geometry. Recent research has revisited a number of classic ideas w.r.t. fine-grained 3D object modeling [8, 7, 5], ranging in detail from about a dozen planar segments used as parts [7] to over thirty surface vertices in a wireframe representation [8, 5]. While a number of recent works have revived (rough) 3D geometric models in the context of scene level understanding [4, 3], we are unaware of any attempts to employ detailed 3D models for scene-level reasoning. In this work we describe a detailed 3D object model together with an explicit occluder representation [9], and use it for estimating the 3D layout of the scene which allows benefiting from the interactions between the modeled 3D objects, all in a common camera-centered coordinate frame.

2. 3D Geometric Object Class Model

We split 3D object detection and modeling into two layers. The first layer is a representation in the spirit of the *poselet* framework [1], whereas the second layer is a 3D active shape model (ASM) based on local $parts$, augmented with a collection of explicit occlusion masks. The ASM tightly constrains the geometry to plausible shapes, and thus can more robustly predict object shape when parts have only weak evidence or are occluded. The inference for the second layer relies on explicit 3D model fitting, adjusting the object hypotheses (in shape, 3D location, and pose) such that their projection best matches image evidence. Thus we inherently reconstruct the 3D layout of the scene while searching for the best deformable models fits for the individual objects.

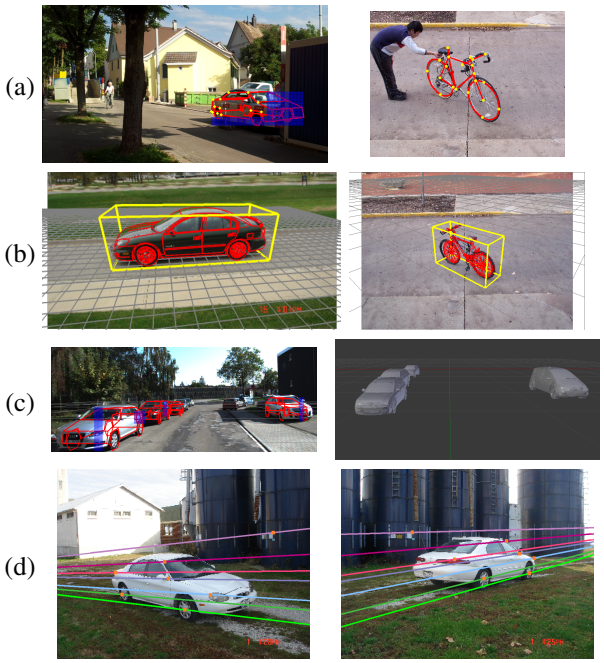
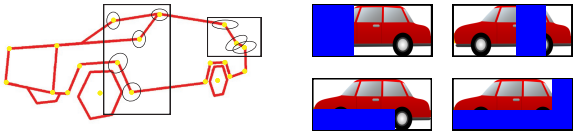


Figure 1. Different outputs of the proposed 3D model – see Sec. 3 for details.

2.1. Parts and part configurations

The atomic units of our representation are $parts$, which are small square patches located at salient points of the object (yellow dots in Fig. 2a). We encode patches with densely sampled shape-context descriptors, and learn a multi-class Random Forest to recognize them, trained on synthetic renderings of 3D CAD models rather than on real data, which greatly reduces the annotation effort. The basic unit of the first layer are larger part $configurations$ ranging in size from 25% to 60% of the full object extent (two examples shown in Fig. 2a). These are defined in the spirit of [1] and found with k-means clustering. The spatial vari-



(a) (b)

Figure 2. (a) 3D model with two $configurations$ of multiple $parts$ (yellow dots) and their distributions within the $configurations$; (b) example occlusion masks.

ability within a $configuration$ is accounted for by training a single-component DPM detector [2] for each. We found that for these detectors real training data is needed.

2.2. Geometric model

We employ two different geometric models for the initial detection and the subsequent 3D modeling. The first layer follows the philosophy of the ISM/poselet method. The second layer utilizes a more explicit representation of global object geometry that is better suited for estimating detailed 3D shape and pose. In the tradition of *active shape models* we learn a deformable 3D wireframe model (through PCA of n salient vertices in 3D-space [8]) from CAD data. The $parts$ described above are defined as small windows around the 2D projection of such a vertex ($\approx 10\%$ in size of the full object width). They allow for fine-grained estimation of 3D geometry and continuous pose, as well as part-level reasoning about occlusion relations.

2.3. Explicit occlusion modeling

The second layer includes an explicit representation of occluders, which are assumed to block the view onto a spatially connected region of the object. Since occluders can only be distinguished if the visibility of at least one part changes, one can approximate the space of all possible occluders by a small, discrete set of masks (Fig. 2(b)). With that set, we aim to explicitly recover the occlusion pattern during second-layer inference, by selecting one of the masks. All parts falling inside the occlusion mask are considered occluded. Their detection scores are not considered in the objective function (Sec. 2.4), instead they are assigned a fixed low score, corresponding to a weak uniform prior that prefers parts to be visible.

2.4. Shape, pose, and occlusion estimation

During inference, we attempt to find instances of the 3D shape model whose 2D projections into the image plane along with occlusion masks best explain the observed image evidence. We devise an objective function which comprises the image evidence from part and $configuration$ detectors, for given viewpoint, 3D location, shape and occlusion mask. Since inference over the non-convex, high-dimensional objective is difficult, we employ a sample-based maximization scheme [6]. This sampling-based approach, where one hypothesizes plausible shapes, poses, and locations of multiple objects in the same camera-centered 3D coordinate frame, further allows one to reason about object-object interactions (*e.g.* determine occlusions from depth-ordering) and object-scene interactions (*e.g.* by explicitly modeling a ground plane).

3. Applications and Experiments

We have evaluated different aspects of our approach and found it to perform en par with or better than state-of-the-art methods for continuous viewpoint estimation, part localization, part occlusion estimation, fine-grained categorization, and even ultra-wide baseline matching.

Fig. 1(a) shows 3D deformable wireframe detections of individual objects. In Fig. 1(b) we additionally retrieve 3D CAD models most similar to these wireframes from the training set (fine-grained object categorization). Fig. 1(c) shows an image with multiple objects (left), and the estimated 3D layout of the scene (right).

Given two images (with very wide baseline) of a static scene with the same object(s), we can also recover the relative camera pose from 3D predictions of those objects, or equivalently find (part) correspondences independent of local appearance. Fig. 1(d) visualizes corresponding epipolar lines.

4. Conclusions

We discuss a detailed 3D representation for deformable object classes which includes an explicit occluder model, so as to enable part-level reasoning about multiple objects in a camera-centric frame. In the future, we intend to include more object interactions and further explore 3D scene understanding using this model.

References

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

[2] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI 2010*.

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

[4] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. *ECCV10*.

[5] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. *NIPS 2012*.

[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. Explicit occlusion modeling for 3d object class representations. *CVPR 2013*.