Physically Grounded 3D Scene Interpretation with Detailed Object Models

M. Zeeshan Zia\textsuperscript{1,3}, Michael Stark\textsuperscript{2}, and Konrad Schindler\textsuperscript{1}

\textsuperscript{1} Photogrammetry and Remote Sensing, ETH Zürich, Switzerland
\textsuperscript{2} Stanford University and Max Planck Institute for Informatics
\textsuperscript{3} Robot Vision, Imperial College London

Abstract

We explore 3D scene reconstruction from a single view building on detailed 3D representations of object classes. These representations allow precise estimation of 3D shape and pose of individual objects, as well as explicit modeling of occlusions. We use the rich detail afforded by these models to reason about physical interactions among multiple object instances, and empirically demonstrate superior performance to models where no such reasoning is performed.

1. Introduction

Over the last few decades, efforts in semantic computer vision have focused on 2D recognition: bounding box level object detection, pixel labeling, and 2D contextual reasoning. Such methods provide the layout of different scene elements (compact, well-delineated objects and “stuff” such as walls, sky, ground) in 2D image space which is only a weak proxy for 3D scene layout. Intuitively, since images are projections of 3D scenes, reasoning about interactions among different elements in 3D should provide stronger constraints and thus better interpretation of visual data. Only recently have scene understanding approaches been demonstrated which perform such coarse reasoning in 3D space, and succeed in providing a physical interpretation of the scene as well as superior detection performance to 2D only approaches \cite{3, 4}. Still, such approaches represent object estimates rather coarsely as 3D bounding boxes and discretized viewpoint estimates. Here, we demonstrate that finer 3D object models allow modeling of object-object interactions at a higher-resolution, yielding better scene interpretations.

We lift a rich object class model that we have developed previously \cite{5} to true 3D, and use it as a building block in our scene representation. We leverage the fine-grained geometry estimates provided by the model to enforce a common ground plane and reason about mutual occlusions at the resolution of individual wireframe vertices. We demonstrate that such detailed, physically grounded interpretation outperforms coarser reasoning.

2. 3D Geometric Object Class Model

We represent an object class as a deformable 3D wireframe, as in the classical “active shape model” formulation. The vertices of the wireframe are defined manually, and wireframe exemplars are collected by annotating a set of 3D CAD models. Principal Components Analysis (PCA) is applied to obtain the mean configuration of the vertices in 3D as well as the principal modes of their relative displacement \cite{1, 5}. For the present work the training exemplars are scaled to their real-world dimensions, to allow for reasoning in a metric scene space. We establish the connection between the 3D geometric object model and its object instances in an image, by means of a set of \textit{parts}, one for each wireframe vertex. For each part, a multi-view appearance model is learned, by generating training patches with non-photorealistic rendering of 3D CAD models from a large number of different viewpoints, and training a view-invariant random forest classifier on these patches.

2.1. Explicit Occluder Representation

The 3D wireframe model allows one to represent partial occlusion at the level of individual parts: each part has an associated binary variable that stores whether the part is visible or occluded. Note that, in theory, this results in an exponential number of possible combinations of occluded and unoccluded parts, hindering efficient inference over occlusion states. We therefore take advantage of the fact that partial occlusion in not entirely random, but tends to follow reoccurring patterns that render certain joint occlusion
states of multiple parts more likely than others: the joint occlusion state depends on the shape of the occluding physical object(s). Here we approximate the shapes of (hypothetical) occluders as a finite set of occlusion masks. This set of masks constitutes a (hard) non-parametric prior over possible occlusion patterns. Occlusions from unknown causes are modeled by searching for the occluder mask that best explains the part detection scores; whereas mutual occlusion is considered by reasoning about the physical interactions among the detected object instances (Sect. 3.2).

3. 3D Scene Model

We proceed by extending the object model of Sect. 2 to entire scenes, where we can jointly reason about multiple objects and their geometric relations, placing them on a common ground plane and taking into account mutual occlusions (Fig. 1). Our 3D scene model comprises a common ground plane and a set of 3D deformable wireframes with corresponding occlusion masks (Sect. 2.1). Our hypothesis space is more expressive than the 2.5D representations used in our previous work [3], as it allows physical reasoning about locations, spaces, and interactions of objects, at the level of individual wireframe vertices and faces.

3.1. Common ground plane

We constrain all object instances to lie on a common ground plane, as often done for street scenes. This assumption usually holds and 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 inference.

3.2. Deterministic occlusion reasoning

Apart from searching for an occluder mask that best explains missing part evidence, we also utilize physical interactions between detected object instances to improve occlusion estimates [4]. Since all object hypotheses reside in the same 3D coordinate system, mutual occlusions can be derived deterministically from their depth ordering (Fig. 2).

3.3. Inference

All evidence in our model comes from object part detection, and the prior for allowable occlusions is given by per-object occlusion masks and relative object positions. We formulate a scene-level likelihood by combining object-level likelihoods (sum of visible part detection scores and a constant score for occluded parts) and ground plane. To do inference in this model we resort to a sampling based optimization scheme, initialized with 2D pre-detections lifted to 3D [5].

4. Evaluation

We perform experiments on a subset of the KITTI dataset [2] of street scenes, which also provides 3D locations of objects (cars and vans) in the images, as measured by a Velodyne laser scanner. 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.

5. Conclusions

We discuss how a 3D deformable object class representation, which includes an explicit occluder model, can be used to obtain high quality scene-level estimates from a monocular image. Modeling fine-grained physical interactions between such object instances yields improved 3D scene estimates.

References