Data-Driven Scene Understanding from 3D Models

Scott Satkin
ssatkin@ri.cmu.edu

Jason Lin
jasonli1@andrew.cmu.edu

Martial Hebert
hebert@ri.cmu.edu

Carnegie Mellon University
The Robotics Institute
Pittsburgh, Pennsylvania

http://cmu.satkin.com/bmvc2012/

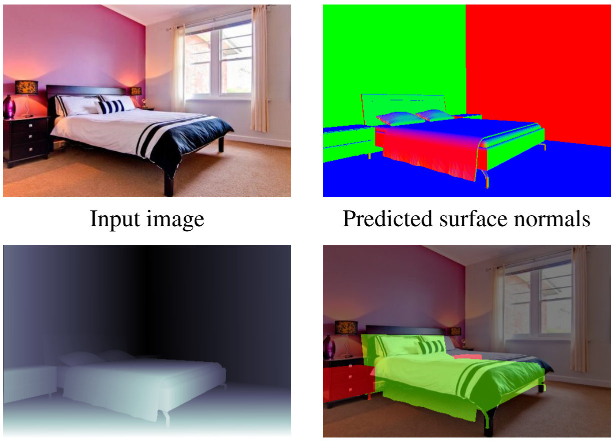
Input image
Predicted surface normals
Predicted depth map
Predicted object labels

Figure 1: From a single image, we estimate detailed scene geometry and object labels.

In this paper, we propose a data-driven approach to leverage repositories of 3D models for scene understanding. Our ability to relate what we see in an image to a large collection of 3D models allows us to transfer information from these models, creating a rich understanding of the scene. We develop a framework for auto-calibrating a camera, rendering 3D models from the viewpoint an image was taken, and computing a similarity measure between each 3D model and an input image. We demonstrate this data-driven approach in the context of geometry estimation and show the ability to find the identities and poses of object in a scene. Additionally, we present a new dataset with annotated scene geometry. This data allows us to measure the performance of our algorithm in 3D, rather than in the image plane.

Recently, large online repositories of 3D data such as Google 3D Warehouse have emerged. These resources, as well as the advent of low-cost depth cameras, have sparked interest in geometric data-driven algorithms. At the same time, researchers have (re-)started investigating the feasibility of recovering geometric information, *e.g.*, the layout of a scene. The success of data-driven techniques for tasks based on appearance features, *e.g.*, interpreting an input image by retrieving similar scenes, suggests that similar techniques based on *geometric* data could be equally effective for 3D scene interpretation tasks. In fact, the motivation for data-driven techniques is the same for 3D models as for images: real-world environments are not random; the sizes, shapes, orientations, locations and co-location of objects are constrained in complicated ways that can be represented given enough data. In principle, estimating 3D scene structure from data would help constrain bottom-up vision processes. For example, in Figure 1, one nightstand is fully visible; however, the second nightstand is almost fully occluded. Although a bottom-up detector would likely fail to identify the second nightstand since only a few pixels are visible, our method of finding the best matching 3D model is able to detect these types of occluded objects. This is not a trivial extension of the image-based techniques. Generalizing data-driven ideas raises new fundamental technical questions never addressed before in this context: What features should be used to compare input images and 3D models? Given these features, what mechanism should be used to rank the most similar 3D models to the input scene? Even assuming that this ranking is correct, how can we transfer information from the 3D models to the input image? To address these questions, we develop a set of features that can be used to compare an input image with a 3D model and design a mechanism for finding the best matching 3D scene using support vector ranking. We show the feasibility of these techniques for transferring the geometry of objects in indoor scenes from 3D models to an input image.

Naturally, we cannot compare 3D models directly to a 2D image. Thus, we first estimate the intrinsic and extrinsic parameters of the camera and use this information to render each of the 3D models from the same view as the image was taken from. We then compute similarity features between the models and the input image. Lastly, each of the 3D models is ranked based on how similar its rendering is to the input image using a learned feature weighting. See Figure 2 for an overview of this process. Please read our full paper for a detailed explaination of our data-driven geometry estimation algorithm and results.

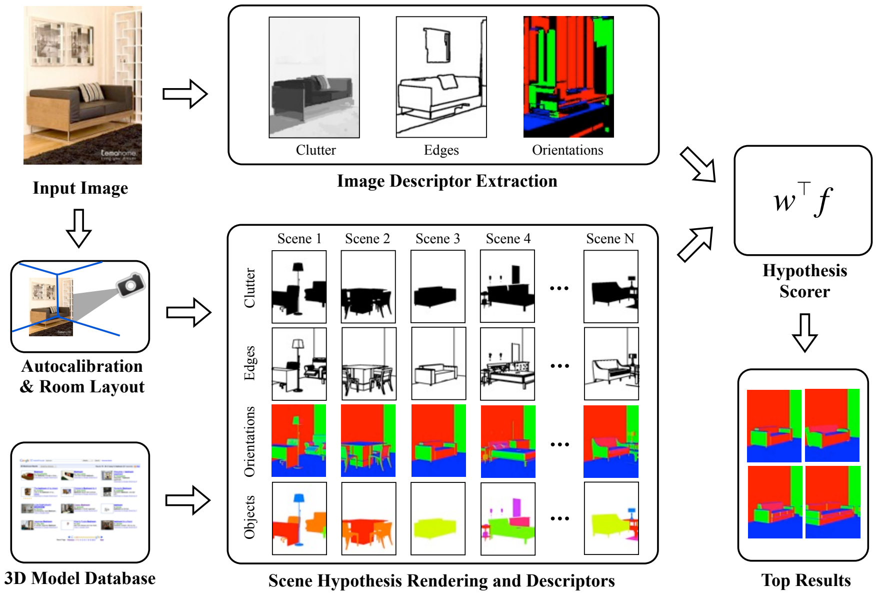

Figure 2: Overview of our approach for matching a 3D model with a monocular image.