Bayesian inference of indoor scenes using composite object models

Luca Del Pero
University of Arizona
delpero@email.arizona.edu

Kobus Barnard
University of Arizona
kobus@sista.arizona.edu

Indoor scene understanding from monocular images has received much recent interest [4, 6, 8], as advancements in recovering the 3D indoor geometry have enabled several interesting applications, such as predicting human activities [2] and identifying objects [5] in the context of the estimated 3D scene. Recent methods achieved promising reconstruction results by using a simple 3D model [4, 6, 8], where both the scene layout and the objects in it are approximated with right-angled parallelepipeds (3D boxes). Typically, a single box is used to approximate the walls, floor and ceiling enclosing the scene (**room box**), and also to model an object inside it, such as a bed or a table.

While gross geometry simplifies inference, we advocate that more topologically specific models, such as tables with legs and top or couches with seat and backrest, hold several advantages over a single box representation. First, bounding boxes of concave objects projected into images tend to include much background, which is confusing evidence for inference (Fig. 1). Second, a more realistic representation is also more discriminative for object recognition, since different object types have now different geometry, as opposed to having each object approximated with a bounding box. Third, non-convex models enable more complex configurations, such as sliding a chair under a table (Fig. 1).

Hence we are developing a comprehensive **3D Bayesian generative model for indoor scenes**, where we use 3D composite object models created from a set of re-usable geometric primitives (parts). Our goal is to provide a **comprehensive and realistic parsing** of the indoor environment, where individual elements are estimated in the context of the overall 3D scene, by jointly inferring the **objects**, the **camera** and the **room box**.

This work relates to that of Lee [6], where the joint inference of room box and objects in it (approximated by single boxes) improves estimating the room box, as the objects explain occlusions. However, we advocate a more top-down approach and unified representation, along the lines of the work and Hoiem [3] and our previous work [7], where individual elements in the scene are estimated while also capturing the interplay among them. An important step with respect to previous work is the use of a more detailed and realistic 3D geometric representation for objects, which allows us to improve both object recognition and the global 3D reconstruction of the scene. In what follows, we provide an overview of the proposed model, the inference strategy, and a few quantitative results. For a more extended discussion, we refer to our 2013 paper [1].

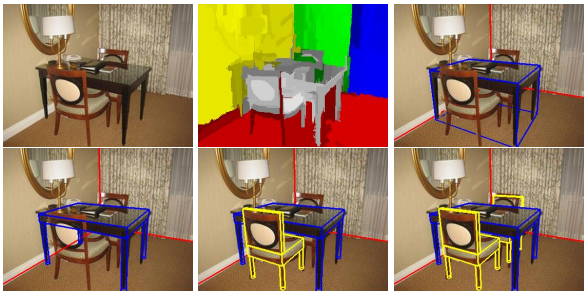

Figure 1. Motivation for using more specific geometric models. A 2D classifier [4] finds a reasonable map for object pixels shown in gray (top middle). However, fitting a single convex block for the table (top right) is hampered by the confusing evidence from the background, while a table with legs and top provides a better fit (bottom left). Further, non-convex models allow for complex configurations, such as a chair under a table (bottom middle). Finally, using context to propose (but not evaluate) object locations allows us to infer a second chair that would be even more difficult to handle using a block representation (bottom right).

A Bayesian model for indoor scenes. We assume that images are generated by the projection of the 3D objects in the scene. We partition model parameters, θ, into scene parameters, s, encoding the 3D geometry, and camera parameters c, modeling the perspective transformation. We define the posterior distribution as

$$p(\theta|D) \propto p(D|\theta)p(\theta) \quad (1)$$

where D are features detected on the image plane and $p(\theta)$ is a prior distribution, which constrains the parameter space to realistic configurations. This includes priors on the typical height of the camera from the floor, and on the size and position of an object given the room box — for example beds tend to be quite short and against a wall.

For the camera parameters, c, we use a simplified perspective camera model [1]. Scene parameters s =

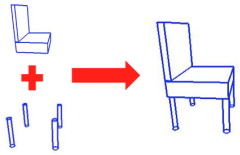

Figure 2. Object models are built by vertically stacking re-usable parts. Here, a chair model is created from a set of four symmetric legs, and an L-component approximating seat and backrest.

$(r, o_1, ..., o_n)$ include the room box r and the unknown number of objects. As in previous work [4, 6], we approximate the room itself with a 3D box. Each object is defined by its type (e.g. table, bed), and by a collection of contiguous 3D geometric parts, which is a function of the object type

$$o_i = (t_i, p_{i1}, ..., p_{in}) \quad . \tag{2}$$

While the list of parts for each object type t is fixed, we allow details of each part to vary among instances of that type. For example, while all objects of type t ="table" share a topology consisting of a top supported by four legs, details such as the leg width or the top height can vary among different table instances. Further, all objects have to lie on the floor, with the exception of objects attached to a wall, which are called frames (e.g. doors and windows) [7].

We create the list of parts defining an object type manually from a set of re-usable geometric primitives, such as a set of four symmetric cylinders, used for modeling the legs of tables and chairs, or an L-shaped structure, approximating a seat with backrest used for both chairs and couches (Fig. 2). In this work, we managed to build object models for eight different categories (bed, cabinet, chair, couch, table, door, picture frame, window) from a set of four parts. Each part is defined by a minimal set of parameters, for example, a backrest is a 3D block defined by its height, width and length. Conditioned on the object type, each part parameter is constrained within a range of plausible values, which is set from training data. For example, the backrest of a couch is usually much thicker than that of a chair. All object parameters are defined as ratios with respect to the total scene size, since we do not have access to absolute size and position when reconstructing from a single image [1].

Priors and likelihood. We use priors on the overall size and position of an object conditioned on its type [7]. Such priors are set from text in online furniture catalogs. For example, we set the mean and variance of the height of a table from the online Ikea catalog. The image likelihood is proportional to the distance between the features generated by projecting the model and those detected on the image plane. We use three standard features in this domain (edges [1], oriented surfaces [6] and geometric context [4]), and introduce a fourth one that encourages 3D object hypotheses whose 2D projection is more uniform in color distribution.

Inference. We use MCMC sampling to search the output space, defined by the parameters of room box, camera, and objects, which we infer jointly. We use the reversible jump modification of the Metropolis Hastings acceptance formula for evaluating discrete changes in the model, which include adding/removing an object, or changing its type, for example turning a table into a couch. We use Hamiltonian Dynamics to sample over subsets of the continuous parameters [1], which include the room box parameters, the camera parameters, and the parameters of each object and its parts.

We rely on data-driven techniques to speed up the inference. For example, we initialize the camera parameters from a triplet of orthogonal vanishing points [6], and propose new objects from 2D corners detected on the image plane [1]. We also introduce part-specific data-driven mechanisms, such as proposing legs from pairs of contiguous vertical segments. These mechanisms are shared by all objects containing that part (tables and chairs in the case of legs), thus **making inference issues transparent** to potential new object models created from the available parts. Finally, we use top-down information about contextual relationships among objects to inform the inference. For example, we propose chairs around the tables in the current scene hypothesis, and this allowed us to find chairs despite heavy occlusions, such as the chair behind the table in Fig. 1.

Results. Our method performs slightly better than the state-of-the-art [8] on the standard room layout error, which measures the accuracy of the predicted room box (13.7% error versus 13.8% on the Hedau dataset [4]). For object recognition, we outperform our previous results [7] on furniture (precision: 53.9% versus 32.5%, recall: 35.7% vs. 20.7%) and frame recognition (precision: 44.9% vs. 33.1%, recall: 41.8% vs. 18.7%). A qualitative example of a scene reconstruction is shown in Fig. 1 (bottom right).

References

[1] L. Del Pero, J. Bowdish, B. Kermgard, E. Hartley, and K. Barnard. Understanding bayesian rooms using composite 3d object models. In *CVPR*, 2013. 1, 2

[2] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011. 1

[3] M. Hebert, A. Efros, and D. Hoiem. Putting objects in perspective. pages II: 2137–2144, 2006. 1

[4] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009. 1, 2

[5] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010. 1

[6] D. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010. 1, 2

[7] L. D. Pero, J. Bowdish, D. Fried, B. Kermgard, E. Hartley, and K. Barnard. Bayesian geometric modeling of indoor scenes. In *CVPR*, 2012. 1, 2

[8] A. Schwing, T. Hazan, M. Pollefeys, and U. R. Efficient structure prediction with latent variables for general graphics models. In *CVPR*, 2012. 1, 2