Localizing 3D Cuboids in Single-view Images

Jianxiong Xiao Bryan C. Russell* Antonio Torralba

Massachusetts Institute of Technology *Intel Labs

Abstract

In this paper we seek to detect rectangular cuboids and localize their corners in uncalibrated single-view images depicting everyday scenes. In contrast to recent approaches that rely on detecting vanishing points of the scene and grouping line segments to form cuboids, we build a discriminative parts-based detector that models the appearance of the cuboid corners and internal edges while enforcing consistency to a 3D cuboid model. Our model copes with different 3D viewpoints and aspect ratios and is able to detect cuboids across many different object categories. We introduce a database of images with cuboid annotations that spans a variety of indoor and outdoor scenes and show qualitative and quantitative results on our collected database. Our model out-performs baseline detectors that use 2D constraints alone on the task of localizing cuboid corners.

1. Introduction

Extracting a 3D representation from a single-view image depicting a 3D object has been a long-standing goal of computer vision [5]. Traditional approaches have sought to recover 3D properties, such as creases, folds, and occlusions of surfaces, from a line representation extracted from the image [4]. Among these are works that have characterized and detected *geometric primitives*, such as quadrics (or "geons") and surfaces of revolution, which have been thought to form the components for many different object types [1]. While these approaches have achieved notable early successes, they could not be scaled-up due to their dependence on reliable contour extraction from natural images.

In this work we focus on the task of detecting *rectangular cuboids*, which are a basic geometric primitive type and occur often in 3D scenes (e.g. indoor and outdoor man-made scenes [6]). Moreover, we wish to recover the shape parameters of the detected cuboids. The detection and recovery of shape parameters yield at least a partial geometric description of the depicted scene, which allows a system to reason about the affordances of a scene in an object-agnostic fashion. This is especially important when the category of the object is ambiguous or unknown. As shown in Figure 1, we build a 3D cuboid detector to detect individual boxy volumetric structures. We build a discriminative parts-based detector that models the appearance of the corners and internal edges of cuboids while enforcing spatial consistency of the corners and edges to a 3D cuboid model. Our model is trained in a similar fashion to recent work that detects articulated human body joints [8].

Figure 1. Given a single-view input image, our goal is to detect the 2D corner locations of the cuboids depicted in the image.

Our cuboid detector is trained across different 3D viewpoints and aspect ratios. This is in contrast to view-based approaches for object detection that train separate models for different viewpoints, e.g. [2]. Moreover, instead of relying on edge detection and grouping to form an initial hypothesis of a cuboid, we use a 2D sliding window approach to exhaustively evaluate all possible detection windows. Also, our model does not rely on any preprocessing step, such as computing surface orientations. Instead, we learn the parameters for our model using a structural SVM framework. This allows the detector to adapt to the training data to identify the relative importance of corners, edges and 3D shape constraints by learning the weights for these terms. We introduce an annotated database of images with geometric primitives labeled and validate our model by showing qualitative and quantitative results on our collected database. We also compare to baseline detectors that use 2D constraints alone on the tasks of geometric primitive detection and part localization. We show improved performance on the part localization task.

2. 3D Cuboid Detector

We represent the appearance of cuboids by a set of parts located at the corners of the cuboid and a set of internal

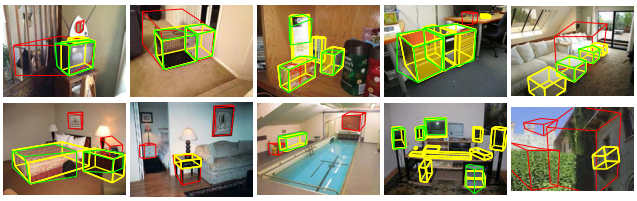

Figure 2. All 3D cuboid detections above a fixed threshold in each image. Notice that our model is able to detect the presence of multiple cuboids in an image (e.g. row 1, columns 2-5) and handles partial occlusions (e.g. row 1, column 4), small objects, and a range of 3D viewpoints, aspect ratios, and object classes. Moreover, the depicted scenes have varying amount of clutter. Yellow - ground truth. Green - correct prediction. Red - false positive. Line thickness corresponds to detector confidence.

edges. We enforce spatial consistency among the corners and edges by explicitly reasoning about its 3D shape. Let I be the image and $p_i = (x_i, y_i)$ be the 2D image location of the ith corner on the cuboid. We define an undirected loopy graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over the corners of the cuboid, with vertices $\mathcal{V}$ and edges $\mathcal{E}$ connecting the corners of the cuboid. We define a scoring function associated with the corner locations in the image:

$$S(I,p) = \sum_{i \in \mathcal{V}} w_i^H \cdot \text{HOG}(I, p_i) + \sum_{ij \in \mathcal{E}} w_{ij}^D \cdot \text{Displacement}^{\text{2D}}(p_i, p_j) + \sum_{ij \in \mathcal{E}} w_{ij}^E \cdot \text{Edge}(I, p_i, p_j) + w^S \cdot \text{Shape}^{\text{3D}}(p) \quad (1)$$

where $\text{HOG}(I, p_i)$ is a HOG descriptor computed at image location p_i and $\text{Displacement}^{\text{2D}}(p_i, p_j) = -[(x_i - x_j)^2, x_i - x_j, (y_i - y_j)^2, y_i - y_j]$ is a 2D corner displacement term that is used in other pictorial parts-based models [2, 8]. By reasoning about the 3D shape, our model handles different 3D viewpoints and aspect ratios. Notice that our model is linear in the weights w. Moreover, the model is flexible as it adapts to the training data by automatically learning weights that measure the relative importance of the appearance and spatial terms. Please refer to [7] for the details on the definition of Edge and Shape^{3D}. This model is trained using standard Structured SVM learning. The inference is initialized by tree-based dynamic programming and distance transform [8], followed by hill climbing [7].

3. SUN Primitive Database

To develop and evaluate any models for 3D cuboid detection in real-world environments, it is necessary to have a large database of images depicting everyday scenes with 3D cuboids labeled. In this work, we seek to build a database by manually labeling point correspondences between images and 3D cuboids. We have built a labeling tool that allows a user to select and drag key points on a projected 3D cuboid model to its corresponding location in the image. Given the corner correspondences, the parameters for the 3D cuboids and camera are estimated. The cuboid and camera parameters are estimated up to a similarity transformation via camera resectioning using Levenberg-Marquardt optimization. For our database, we have 785 images with 1269 cuboids annotated. We have also collected a negative set containing 2746 images that do contain any cuboid like objects.

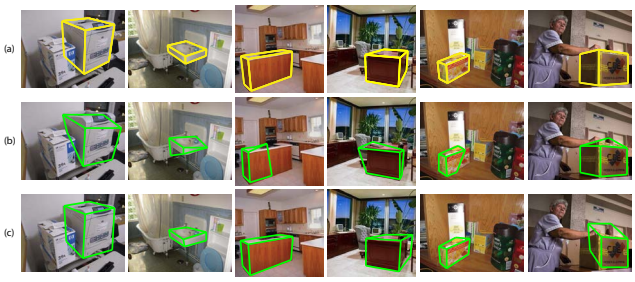

Figure 3. Corner localization comparison for detected geometric primitives. (a) Input image and ground truth annotation. (b) 2D tree-based initialization. (c) Our full model. Notice that our model is able to better localize cuboid corners over the baseline 2D tree-based model, which corresponds to 2D parts-based models used in object detection and articulated pose estimation [2, 8]. The last column shows a failure case where a part fires on a "cuboid-like" corner region in the image.

4. Conclusion

We have introduced a novel model that detects 3D cuboids and localizes their corners in single-view images. Moreover, we have constructed a dataset with ground truth cuboid annotations. In [3], we further develop a system to improve the performance by making use of depth map from RGB-D sensors. Our dataset is publicly available at http://SUNprimitive.csail.mit.edu.

References

[1] I. Biederman. Recognition by components: a theory of human image interpretation. *Pyschological review*, 94:115–147, 1987. 1

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 32(9), 2010. 1, 2

[3] H. Jiang and J. Xiao. A linear approach to matching cuboids in rgbd images. In *CVPR*, 2013. 2

[4] J. L. Mundy. Object recognition in the geometric era: A retrospective. In *Toward Category-Level Object Recognition, volume 4170 of Lecture Notes in Computer Science*, pages 3–29. Springer, 2006. 1

[5] L. Roberts. Machine perception of 3-d solids. In *PhD. Thesis*, 1965. 1

[6] J. Xiao and Y. Furukawa. Reconstructing the world's museums. In *ECCV*, 2012. 1

[7] J. Xiao, B. Russell, and A. Torralba. Localizing 3d cuboids in single-view images. In *NIPS*. 2012. 2

[8] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In *CVPR*, 2011. 1, 2