Inferring 3D Layout of Building Facades from a Single Image

Jiyan Pan[1], Martial Hebert[2], Takeo Kanade[2]
[1]Google Inc. [2]Robotics Institute, Carnegie Mellon University.

Given a single image of an urban scene, automatically inferring the underlying 3D layout of building facades in the scene would significantly benefit many tasks in fields such as autonomous navigation and augmented reality. It goes beyond depth map estimation [4, 5, 6, 7], because it provides a richer understanding of the scene, such as camera pose, locations of planes and blocks, and how they are related with each other in 3D [1].

In this paper, we model building facades as a set of planes with continuous orientations, and then quantitatively reason over their 3D locations using inter-planar geometric constraints. This approach would produce a richer interpretation of the facade scene than existing pixel/segment based approaches [2, 3, 5] and block-based approaches [1]. More specifically, our approach is able to provide critical scene understanding information (*e.g.* quantitative orientation, depth, and relationships of facade planes) that existing algorithms do not provide. Please see Figure 1 for an example.

The main contributions of this work are as follows. 1) We propose a plane-based fully quantitative model to infer the 3D layout of building facades, where each plane is represented by a continuous orientation vector and a distribution of depth values. 2) In such a model, multiple cues, such as semantic segmentation, surface layout, and vanishing lines, are utilized to detect and decompose the building region into distinctive planes. 3) The quality of an individual candidate plane is determined by its compatibility with both 2D evidence from image features and 3D evidence such as camera and building height. 4) We model different types of 3D geometric relationships among candidate planes, and apply a CRF to both determine their validity and infer their optimal depths. 5) We do not assume ground is horizontal or buildings are vertical with respect to the camera, nor do we take the Manhattan world assumption.

More specifically, we first use a novel sampling algorithm to segment out a set of candidate facade planes in the image, together with their 3D orientations. We then infer the validity and optimal depth of each candidate facade plane based on its individual compatibility with 2D and 3D cues, as well as its mutual compatibility with adjacent facade planes. Individual compatibility checks if the image features from a candidate facade plane agree with its semantic label and 3D orientation. It also checks if the depth of a candidate facade plane is geometrically plausible. Mutual compatibility checks 3D geometric constraints between a pair of facade planes, including convex corner constraint, occlusion constraint, and alignment constraint. Such a reasoning process can be efficiently achieved using a CRF.

Details of the sampling algorithm and the optimization method involving various 2D and 3D constraints are described in the paper. Our experiments show that our approach yields a more informative interpretation of building facades while maintaining a competitive performance on several quantitative measures. Compared with the block-based approach [1] that yields a coarse and qualitative reconstruction, our method returns a numeric parameterization of a set of planes that compose facades. Compared with super-pixel based approaches that return a depth map [5], our method enables reasoning over planes and blocks and provides a higher-level understanding of the scene.

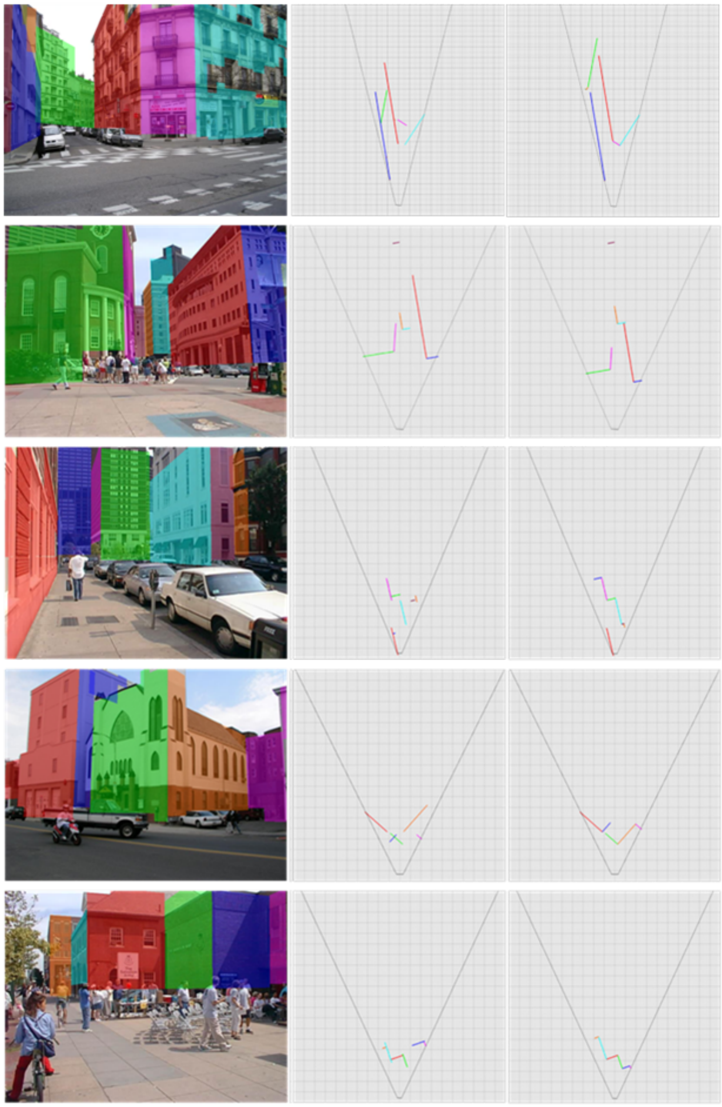

Figure 1: From a single image, our algorithm detects facade planes of different 3D orientations, and infers their optimal depth in 3D space. Left: Detected distinctive facade planes. Middle/Right: Top-down view of recovered facade planes before/after considering inter-planar geometric constraints.

[1] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: image understanding using qualitative geometry and mechanics. *ECCV*, 2010.

[2] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007.

[3] D. Hoiem, A. A. Efros, and M. Hebert. Closing the loop on scene interpretation. *CVPR*, 2008.

[4] K. Karsch, C. Liu, and S. B. Kang. Depthtransfer: Depth extraction from video using non-parametric sampling. *PAMI*, 2014.

[5] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. *CVPR*, 2010.

[6] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. *NIPS*, 2005.

[7] A. Saxena, M. Sun, and A. Y. Ng. Make3d: learning 3-d scene structure from a single still image. *PAMI*, 2009.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage. Jiyan Pan was at the Robotics Institute, Carnegie Mellon University when the work was performed.