Manhattan Junction Catalogue for Spatial Reasoning of Indoor Scenes

Srikumar Ramalingam[1] Jaishanker K. Pillai[2*] Arpit Jain[2] Yuichi Taguchi[1]
[1]Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA
[2]Dept. of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA
{ramalingam,taguchi}@merl.com, {jsp,ajain}@umiacs.umd.edu

Abstract

Junctions are strong cues for understanding the geometry of a scene. In this paper, we consider the problem of detecting junctions and using them for recovering the spatial layout of an indoor scene. Junction detection has always been challenging due to missing and spurious lines. We work in a constrained Manhattan world setting where the junctions are formed by only line segments along the three principal orthogonal directions. Junctions can be classified into several categories based on the number and orientations of the incident line segments. We provide a simple and efficient voting scheme to detect and classify these junctions in real images. Indoor scenes are typically modeled as cuboids and we formulate the problem of the layout estimation as an inference problem in a conditional random field. Our formulation allows the incorporation of junction features and the training is done using structured prediction. We outperform other single view geometry estimation methods on standard datasets.

1. Introduction and Related Work

In Figure 1, two or more line segments intersect at different points and we refer to these intersections as junctions. Based on the patterns formed by the incident line segments, we can classify them into different categories such as **L**, **Y**, **W**, **T** and **X** junctions [10, 11]. The types and locations of these junctions provide several geometrical cues about the scene. In this paper, we detect these junctions automatically from a given image and use them to improve the spatial understanding of indoor scenes.

As Sugihara [10] pointed out, human beings invented a noble class of pictures called *line drawings* to represent 3D shapes of objects. The problem of interpreting line drawings was considered as a means to reach the final goal of single view 3D reconstruction. While it is almost straightforward to detect junctions in a given line drawing, detecting junctions in real images is hard and ambiguous. Rather than reconstructing a 3D scene using geometric primitives, Hoeim et al. [3] represented a 3D scene as a popup model normally used to build stages for children's book. Saxena et al. [8] took a different approach to infer the absolute depth directly using both image features and weak assumptions based on coplanarity and connectivity. For modeling indoor scenes, Hedau et al. [2] used a cuboid model to approximate the geometry of the room. Under this model, the pixels in a given image are classified into left wall, middle wall, right wall, floor and ceiling. The importance of corners in images have been emphasized in [6, 5, 1]. Our primary contribution in this work is to provide an efficient voting-based method to detect junctions from real images. Using a CRF model, we show that junction features can be useful to improve the performance of indoor scene understanding algorithms.

Figure 1. *A living room with several junctions of types* **L**, **T**, **Y**, **X** *and* **W**. *We present a novel method to detect these junctions and use them for recovering the spatial layout of a scene.*

2. Junction Detection

In our work, we detect the junctions using a simple two-stage algorithm: (1) we vote for 6 accumulator arrays using line segments along the vanishing points for every pixel p, and (2) we detect different types of junctions by applying a product operation to the contents of the 6 accumulator arrays. We refer to the 6 accumulators as $\mathsf{V}_{\vec{x}}$,

*now at A9, a subsidiary of Amazon.

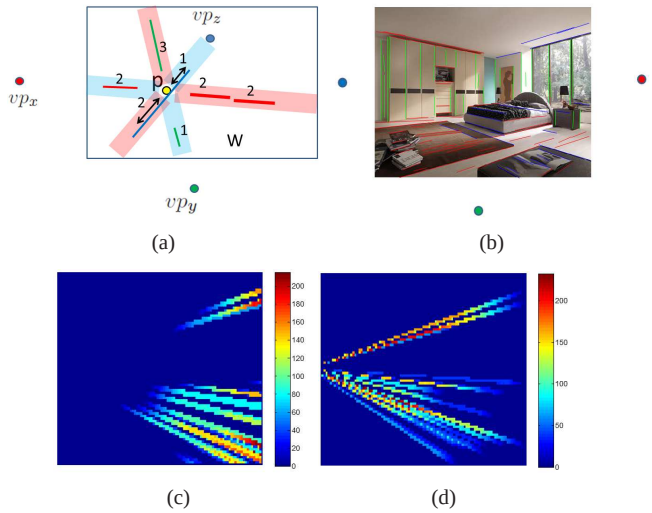

Figure 2. *The idea behind our voting method. We build 6 accumulator arrays, each of which stores votes from lines along one of the three principal orthogonal directions. (a) In this example, the contents of the accumulators corresponding to the point p are given by $V_{\overrightarrow{x}}(p) = 2$, $V_{\overleftarrow{x}}(p) = 4$, $V_{\overrightarrow{y}}(p) = 1$, $V_{\overleftarrow{y}}(p) = 3$, $V_{\overrightarrow{z}}(p) = 1$ and $V_{\overleftarrow{z}}(p) = 2$. (b) An indoor scene along with the detected line segments and vanishing points. (c, d) The contents of the accumulators $\mathsf{V}_{\overrightarrow{z}}$ and $\mathsf{V}_{\overleftarrow{z}}$, respectively, for the image shown in (b).*

$\mathsf{V}_{\overleftarrow{x}}$, $\mathsf{V}_{\overrightarrow{y}}$, $\mathsf{V}_{\overleftarrow{y}}$, $\mathsf{V}_{\overrightarrow{z}}$ and $\mathsf{V}_{\overleftarrow{z}}$. Let us denote the votes at a specific point p in an accumulator V_j as $V_j(p)$, where $j \in \{\overrightarrow{x}, \overleftarrow{x}, \overrightarrow{y}, \overleftarrow{y}, \overrightarrow{z}, \overleftarrow{z}\}$. For each point p, every line segment that is collinear with the line joining p and vp_i ($i \in \{x, y, z\}$) votes for either $V_{\overrightarrow{i}}(p)$ or $V_{\overleftarrow{i}}(p)$ depending on its location with respect to p and vp_i: If the line segment lies in the region between p and vp_i, it votes for $V_{\overrightarrow{i}}(p)$; if the line segment lies outside of the region between p and vp_i, and not adjacent to vp_i, then it votes for $V_{\overleftarrow{i}}(p)$. The vote is weighted by the length of the line segment. The subscript $\overrightarrow{i}$ refers to the line segments towards the vanishing point vp_i, while the subscript $\overleftarrow{i}$ refers to the line segments away from the vanishing point vp_i. This idea of voting is illustrated with an example in Figure 2(a). Using the 6 accumulators, we can detect junctions using simple product operations. At every point p, the corresponding 6 accumulator cells $V_j(p)$ tell us the presence of lines that are incident with this point. To detect a junction, we have to ensure that there are line segments in specific directions, and we also have to ensure that there are no line segments in the rest of the directions.

3. Inference

Our inference algorithm follows the approach of Hedau et al. [2] for learning the scoring function that is used in evaluating several possible layouts. The exact details of our CRF model and the inference algorithm can be found in [7]. Using both geometric context (GC) and orientation maps (OM), Lee et al. [4] showed a misclassification error of 18.6%. Using junction features in addition to GC and OM, we obtained an error of 13.34%. Using a higher sampling to generate layouts, Schwing et al. [9] obtained an error of 13.59%. In future, we plan to use our junction features in their framework.

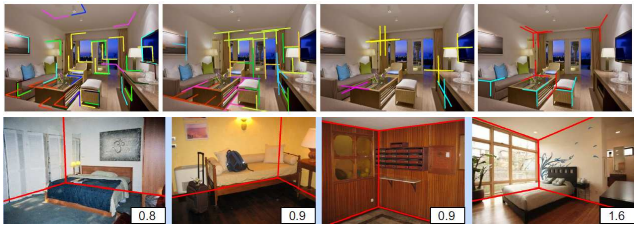

Figure 3. *On the top row, we show prominent junctions such as* **L**, **T**, **X** *and* **Y** *ordered from left to right. On the bottom we show a few top scoring layouts along with the pixel-misclassification error shown on bottom right.*

Acknowledgments: We thank Jay Thornton, Shotaro Miwa, Makito Seki, Jonathan Yedidia, Matthew Brand, Philip H.S. Torr, Karteek Alahari and Peter Varley for useful discussions on single view 3D reconstruction and line drawings.

References

[1] A. Flint, D. Murray, and I. Reid. Manhatten scene understanding using monocular, stereo, and 3D features. In *ICCV*, 2011.

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

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

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

[5] D. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, 2009.

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

[7] S. Ramalingam, J. Pillai, A. Jain, and Y. Taguchi. Manhattan junction catalogue for spatial understanding of indoor scenes. In *CVPR*, 2013.

[8] A. Saxena, S. H. Chung, and A. Y. Ng. 3-D depth reconstruction from a single still image. *IJCV*, 2008.

[9] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction for 3D indoor scene understanding. In *CVPR*, 2012.

[10] K. Sugihara. *Machine Interpretation of Line Drawings*. MIT Press, 1986.

[11] D. Waltz. Understanding line drawings of scenes with shadows. In *The Psychology of Computer Vision*, 1975.