Inferring Spatial Layout from A Single Image via Depth-Ordered Grouping

Stella X. Yu
Computer Science Department
Boston College
Chestnut Hill, MA 02467

Hao Zhang
Computer Science Division
University of California
Berkeley, CA 94720

Jitendra Malik
Computer Science Division
University of California
Berkeley, CA 94720

Abstract

Inferring the 3D spatial layout from a single 2D image is a fundamental visual task. We formulate it as a grouping problem where edges are grouped into lines, quadrilaterals, and finally depth-ordered planes. We demonstrate that the 3D structure of planar objects in indoor scenes can be fast and accurately inferred without any learning or indexing.

1. Introduction

Understanding the 3D space depicted in an image is essential for recognition and navigation. We aim to infer depth-ordered planes that partition the 3D space without any learning or specific knowledge of the objects in the scene.

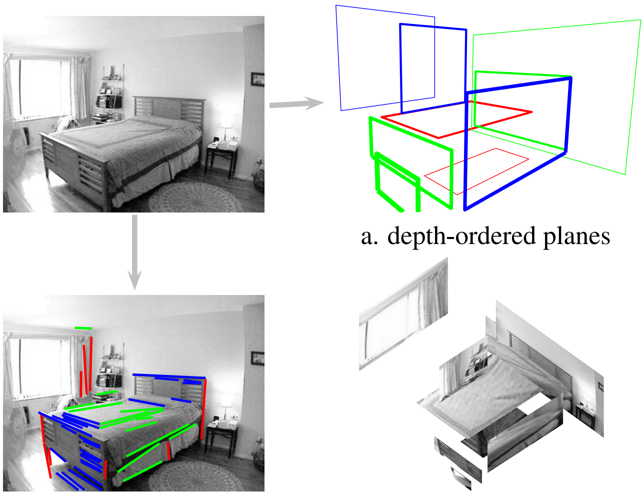


Figure 1: Spatial layout from a single image of an indoor scene. Planes are ordered in depth along dominant directions in the 3D space, thicker outlines for those closer to the viewer. Lines in the image are partitioned into occluders and those that belong to the spatial frame. We can further visualize their relative depth in 3D.

The problem of 3D reconstruction has been studied extensively in the past. It involves two relatively independent subjects: geometry and statistics. The former studies the constraints on 2D features imposed by the camera imaging process (e.g. *vanishing points, homography* [7]), while the latter studies what types of cues in the image are most correlated with the perception of the 3D real world [9] (e.g. *shape-from-X* modules, where X could be shading [19, 14], texture [8, 16], or line junctions [15, 2, 18]).

If camera calibration and manual cue inputs can be assumed, remarkable 3D reconstruction can be achieved using geometrical approaches [5, 13, 1, 22], whereas if the type of shape-from-X can be known *a priori*, then subtly curved surfaces can be precisely reconstructed using statistical approaches [17, 20]. However, these prerequisites are non-trivial to satisfy, limiting the use of these methods.

One way to overcome the limitation is to constrain possible spatial relationships. [10] presents an attribute graph grammar with 6 production rules for parsing objects, surfaces, rectangles and their spatial relations in man-made scenes. Impressive results are shown on the detection and grouping of repetitive tiles of rectangles. Other aspects of spatial layout remain to be seen on more general images.

An alternative to such a rule-based generative approach is statistical learning [11, 6, 12]. The idea is to extend stylized Shape-from-X features to a list of features whose associations with 3D attributes can be learned from annotated images. Given a new image, the list of features are evaluated, and the most likely 3D attributes are retrieved from the memory of associations. The success often depends on how similar the test image is to the training images.

We propose a grouping formulation where pixels are turned into depth-ordered planar surfaces based on proximity, curvilinearity, parallelism, orthogonality, perspectivity, and depth ordering (Fig. 2). With some heuristics to deduce proportional depth relationships from a real image [4], we can visualize relative depth by back-projecting the texture in the 2D image onto planes in the 3D space (Fig. 1).

2. Method details

Our method has four grouping stages (Fig. 2): **1.** Edges are extracted and grouped into lines. **2.** Lines are clustered based on local parallelism, curvilinearity, orthogonality, and convergence to common vanishing points. **3.** Quadrilater-

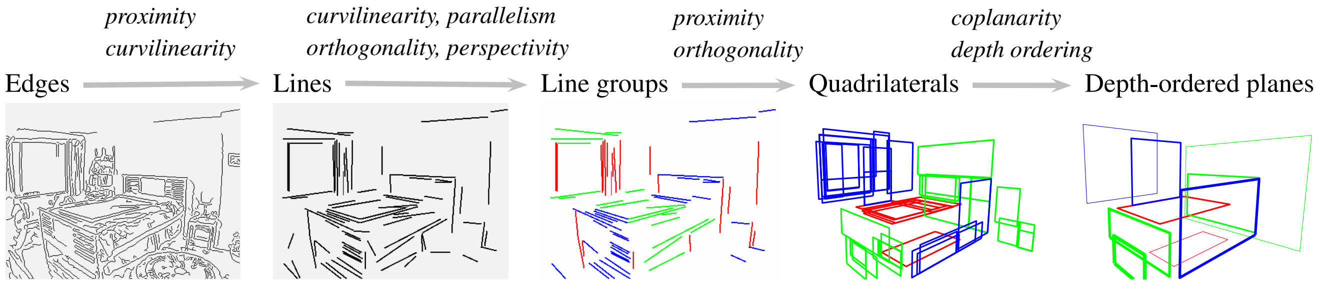


Figure 2: Our method groups edges into lines, line clusters, quadrilaterals, and depth-ordered planes based on cues in the single image.

als are formed by adjacent lines of different directions. **4.** Depth-ordered planes are obtained by grouping quadrilaterals based on coplanarity and relative depth.

2.1. From Edges to Line Segments

We link edge pixels into line segments. Shown in Fig. 3, the distribution of line orientations often reveals the spatial frame of an indoor scene. If we correct the orientations with vanishing points and rule out shorter lines, the distribution becomes peakier, indicating stronger correlation.

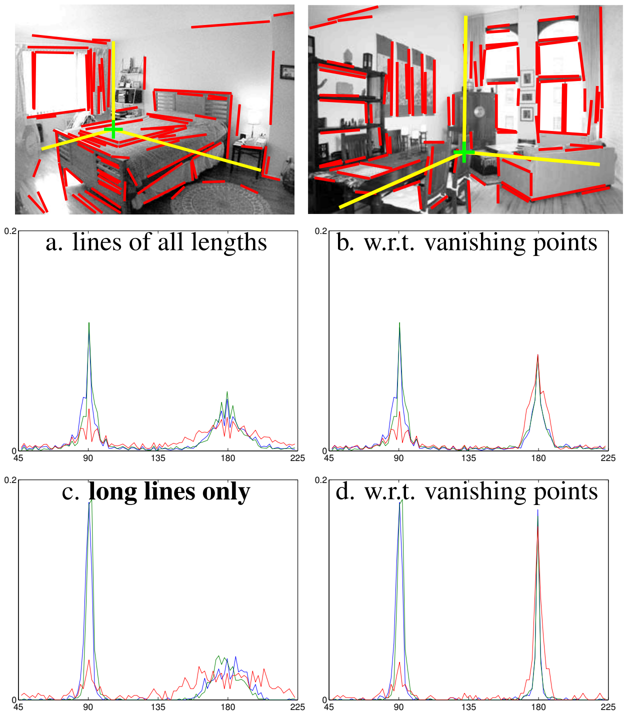


Figure 3: Line orientations are indicative of dominant 3D directions. **Row 1**: Two room images marked with lines and the spatial frame. The frame divides the image into floor, left and right walls. Lines within each region often assume the two dominant directions that define the region. **Rows 2-3**: Each plot has three empirical probability distributions of orientations, one for each region (e.g. red for the floor), collected over 150 such room images. In b and d, the line angle is defined with respect to the corresponding vanishing point. In c and d, lines over a certain length are selected.

Such a statistical correlation has been utilized in [3]. However, their solution to extrinsic spatial frames depends on the orthogonality of three dominant directions, which we do not assume. Whereas they use raw edges at each pixel location, we use long lines only. We thus ignore non-informative short edges (which often result from texture and small objects), discount dependent edge evidence, and increase the estimation reliability of frame directions.

2.2. From Line Segments to Line Clusters

To classify lines according to their alignment with the spatial frame, one could estimate vanishing points using RANSAC [1, 22]. However, such a model-based classification has two problems: it gets easily confused in the presence of clutter and does not respect spatial coherence. Illustrated in Fig. 4a, lines that are parallel in the 3D space converge to a vanishing point in the image, but the converse is not true: lines that converge to a single point in the image do not necessarily correspond to parallel lines in 3D.

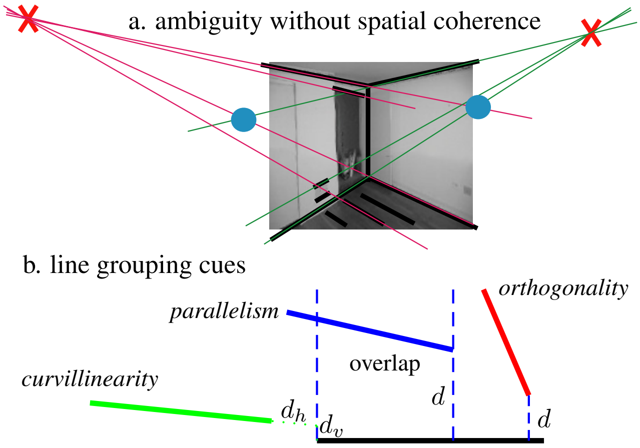


Figure 4: **a:** Ambiguity of vanishing points without spatial coherence. The wrong set of points (×) might explain all the black lines better than the correct ones (●) in terms of line convergence. **b:** Enforcing spatial coherence with grouping cues between lines. The black horizontal line is curvilinear with the left line, parallel with the middle line, and orthogonal with the right vertical line.

We address the problem in two steps. The first step is to cluster lines based on curvilinearity A_-, parallelism $A_\|$

and orthogonality $R_\perp$. Shown in Fig. 4b, extended lines have high A_-; side-by-side lines have high $A_\|$; and nearly orthogonal lines have high $R_\perp$.

$$A_- = \exp\left(-\frac{d_h^2}{2\sigma_{c1}^2} - \frac{d_v^2}{2\sigma_{c2}^2} - \frac{1-\cos^2\theta}{2\sigma_{c3}^2}\right)$$

$$A_\| = \exp\left(-\frac{d^2}{2\sigma_{p1}^2} - \frac{(1-\text{overlapping ratio})^2}{2\sigma_{p2}^2} - \frac{1-\cos^2\theta}{2\sigma_{p3}^2}\right)$$

$$R_\perp = \exp\left(-\frac{d^2}{2\sigma_{o1}^2} - \frac{\cos^2\theta}{2\sigma_{o2}^2}\right),$$

$$d = \begin{cases} 0, & \text{if intersect} \\ \min(\text{endpoints to line distance}), & \text{if no overlap} \\ \min(\text{overlap endpoints to line distance}), & \text{if overlap} \end{cases}$$

where θ is angular difference between two lines. We use $\sigma_{c1} = 3$, $\sigma_{c2} = 0.1$, $\sigma_{c3} = 0.7$, $\sigma_{p1} = 3$, $\sigma_{p2} = 0.1$, $\sigma_{p3} = 0.1$, $\sigma_{o1} = 2$, $\sigma_{o2} = 0.3$. All these grouping cues diminish over distance, and the distance is normalized with respect to the length of the reference line. A longer line thus has a larger neighbourhood to evaluate. To account for perspectivity, we also increase the angular tolerance for parallelism as the distance between lines increases.

These local grouping cues are effective at enforcing spatial coherence which is missing in Fig. 4a, i.e., the ceiling and floor lines on the same wall overlap enough to be considered parallel despite large angular difference, whereas those on different walls have too large a vertical gap to be collinear despite their smaller angular difference.

While A_- and $A_\|$ measure how likely two lines belong in the same group, $R_\perp$ measures how likely two lines belong to different groups. We treat them as pairwise attraction and repulsion respectively, and use the graph cuts method in [24, 25] to find 3 clusters (Fig. 5a).

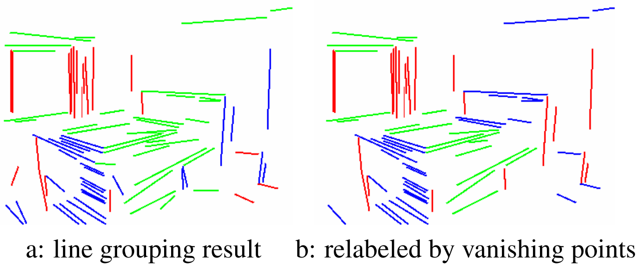

a: line grouping result b: relabeled by vanishing points

Figure 5: Line grouping facilitates the estimation of vanishing points, which in turn lead to a globally constrained classification. Red, green and blue colors mark vertical lines, left and right dominant lines respectively throughout the paper.

We estimate one vanishing point for each of the three line clusters [22]. The point-fitting process converges much faster among the lines in the same cluster. The set of vanishing points in turn provide a more accurate reclassification of all the lines (Fig. 5b).

2.3. From Lines to Quadrilaterals

Lines that are next to each other but point in different directions are likely to come from the same planar surface in the 3D space (Fig. 6). Since we know the corresponding vanishing point for each line, we can represent the plane by a quadrilateral that tightly covers the two defining lines. This also corrects estimation errors in the line orientations.

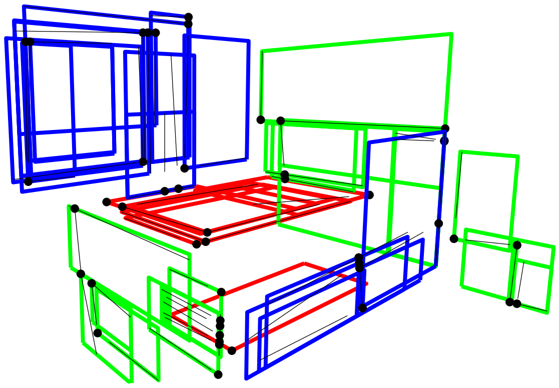


Figure 6: Quadrilaterals are determined by adjacent lines of different orientations and their vanishing points. The black dots indicate the intersections of such line pairs. A quadrilateral assumes the color that is complementary to the colors of the lines: e.g., red and green lines in Fig. 5 define blue quadrilaterals in Fig. 6.

2.4. From Quadrilaterals to Ordered Planes

Quadrilaterals are grouped into planes based on coplanarity and relative depth orders.

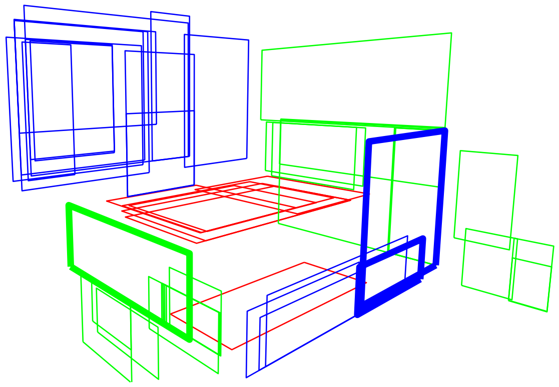
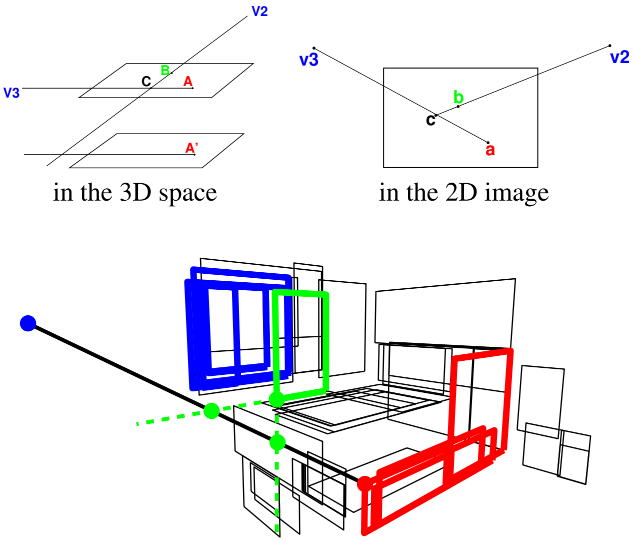


Figure 7: Coplanarity between quadrilaterals is measured by their degree of overlap. Only quadrilaterals of the same orientation is considered. Highlighted in thick outlines, two blue planes overlap well and thus have large coplanarity, whereas the green one overlaps little with others and becomes a lone quadrilateral.

The coplanarity $A_\square$ between two quadrilaterals increases with the extent of their overlap: More overlap in the image, more likely they lie on the same planar surface in the

3D space. Since the quadrilaterals share common vanishing points, the extent of overlap can be measured more straightforwardly in a rectified plane. That is, we compute a homography that maps quadrilaterals into rectangles, with the aspect ratio determined by the lengths of two defining lines. The area of intersection between two rectangles is trivial to compute. We take the value normalized by the minimum of the two rectangular areas as the degree of overlap (Fig. 7).

Fig. 8 illustrates our relative depth test. Two points can be ordered along a direction if two conditions hold: 1) Their pixels must be collinear with the vanishing point that corresponds to the direction; 2) As the vanishing point indicates an infinite distance away from the viewer, the one that is closer to the vanishing point must lie behind. If two points do not line up in that direction in the 3D space, but they lie in a plane that is parallel to the direction, we can align them by projecting one point perpendicularly to intersect that direction. This intersection can then be used to infer relative depth. This procedure has a counterpart in the 2D image by connecting pixels with vanishing points, and the sign of displacement often indicates relative depth even when the two points are not in a plane that is parallel to the direction.



in the 3D space

in the 2D image

Figure 8: **Row 1:** Relative depth between two points along a particular direction. If two points A and B lie in the plane that is parallel to the direction V_3, we can order them by projecting B to the line AV_3. This procedure has a counterpart in the image. Since c lies closer to the vanishing point v_2, C and thus B must lie behind A. For A' and B, such intersection C does not exist. Nevertheless, the corresponding operation in the image still produces an intersection c, which is often indicative of the relative ordering of A' and B. **Row 2:** Relative depth between quadrilaterals along their normal direction. The red quadrilateral is considered in front of the green, because both green intersection points lie closer to the blue vanishing point. Here color indicates depth: red is in front of green, green is in front of blue.

The relative depth R_d between two quadrilaterals is determined by the relative depth of their end points. A quadrilateral is considered in front, only if its end point is in front along both defining directions.

The coplanarity $A_\square$ provides cues to merge two quadrilaterals into one plane, whereas the relative depth R_d provides cues to segregate them in depth. Treating them as pairwise attraction and directional repulsion in a graph cuts framework [23], we integrate both types of cues simultaneously to reach a global depth ordering of quadrilaterals.

We obtain a set of depth-ordered planes by computing the unions of quadrilaterals at the same depth. The union is trivial to compute in a homography-rectified coordinate system. We can also classify all the lines into either frame lines or occluders. Lines making up those quadrilaterals that do not lie farthest back must be occluders.

3. Results and Discussions

We implement our algorithm in MATLAB. The same set of parameters are used for all the results shown in Fig. 9 and Fig. 10. The most time consuming operation is edge detection, for which we use the MATLAB Canny edge detector (about 5 seconds for an image of size 400×400 on a linux machine with 2GHz CPU). The rest grouping operations deal with significantly fewer and larger geometrical entities, which altogether can be done in less than 0.5 seconds. There is no learning or indexing involved.

Our results show that it is possible to use a grouping mechanism to recover 3D spatial layout information from a single image. A number of interesting issues remain.

1. Grouping lines into quadrilaterals. Our current algorithm hypothesizes a quadrilateral area with two adjacent lines of different directions. This simple recipe is local and thus prone to false positives. The middle blue plane in Fig. 1a is such a case. A red vertical line is taken from the window curtain, while a green line is taken from the edge of the bed. This plane does not exist in the 3D space. Such over-generalizations can also be seen in Fig. 9 (row 2, column 1). Shading, texture, and cross validation from nearby planes could help eliminate these false plane hypotheses.

2. Sensitivity of relative depth cues. Our relative depth test is remarkably accurate at small depth differences. In Fig. 9 (row 1, column 2): the front side of the night stand is correctly evaluated to be in front of the left wall, not because there are green lines between the legs of the stand, but because both vertical lines corresponding to the two legs are detected and the spatial relationship between their endpoints provides a powerful cue for perspectivity. On the other hand, this sensitivity also makes a wall fixture such as the heater in Fig. 9 (row 1, column 3) to be in front, which might not be desired.

3. Plane and line ownerships. Our test for occluders

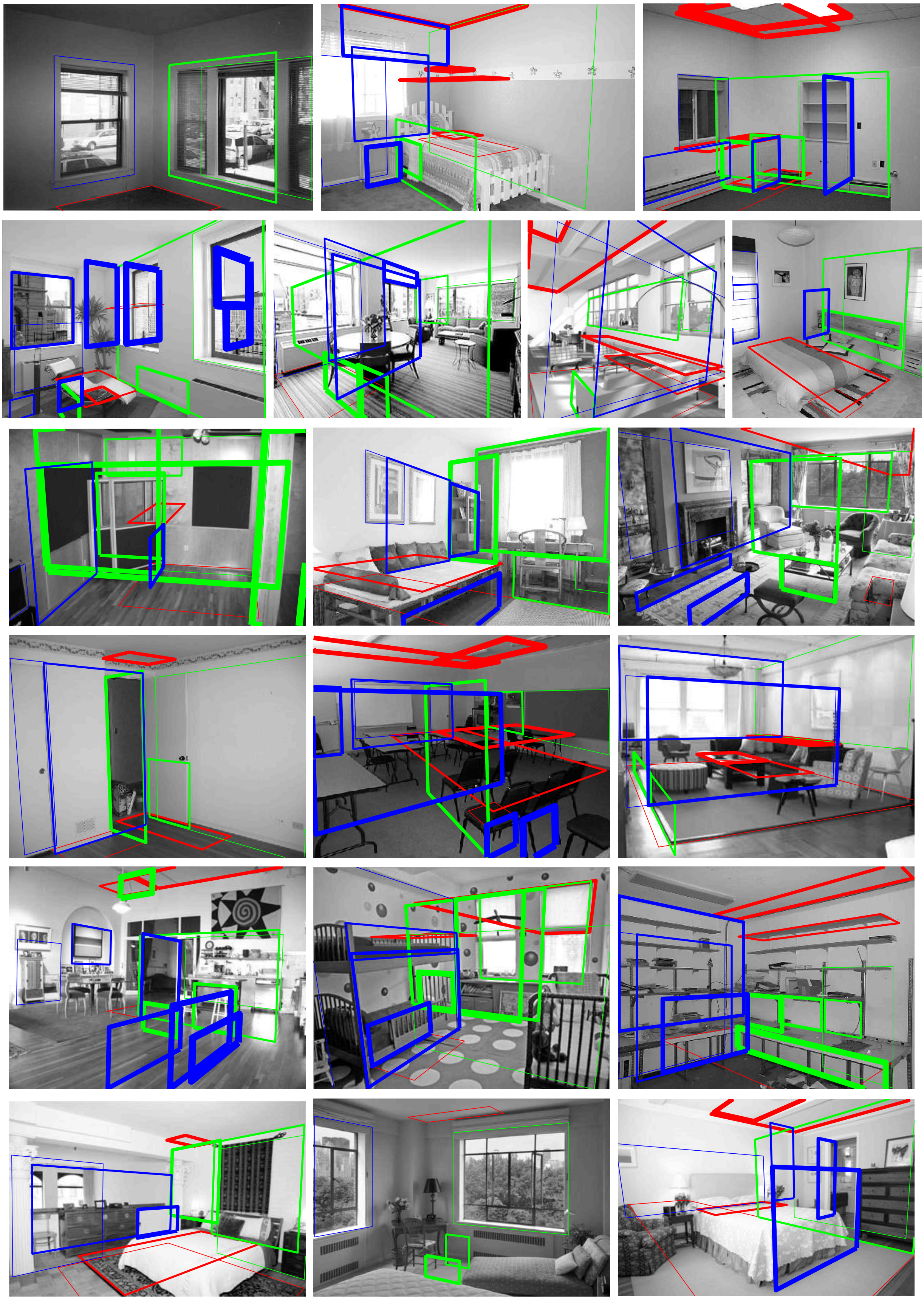


Figure 9: Depth-ordered planes for a set of indoor images. Planes with thicker lines are in front.

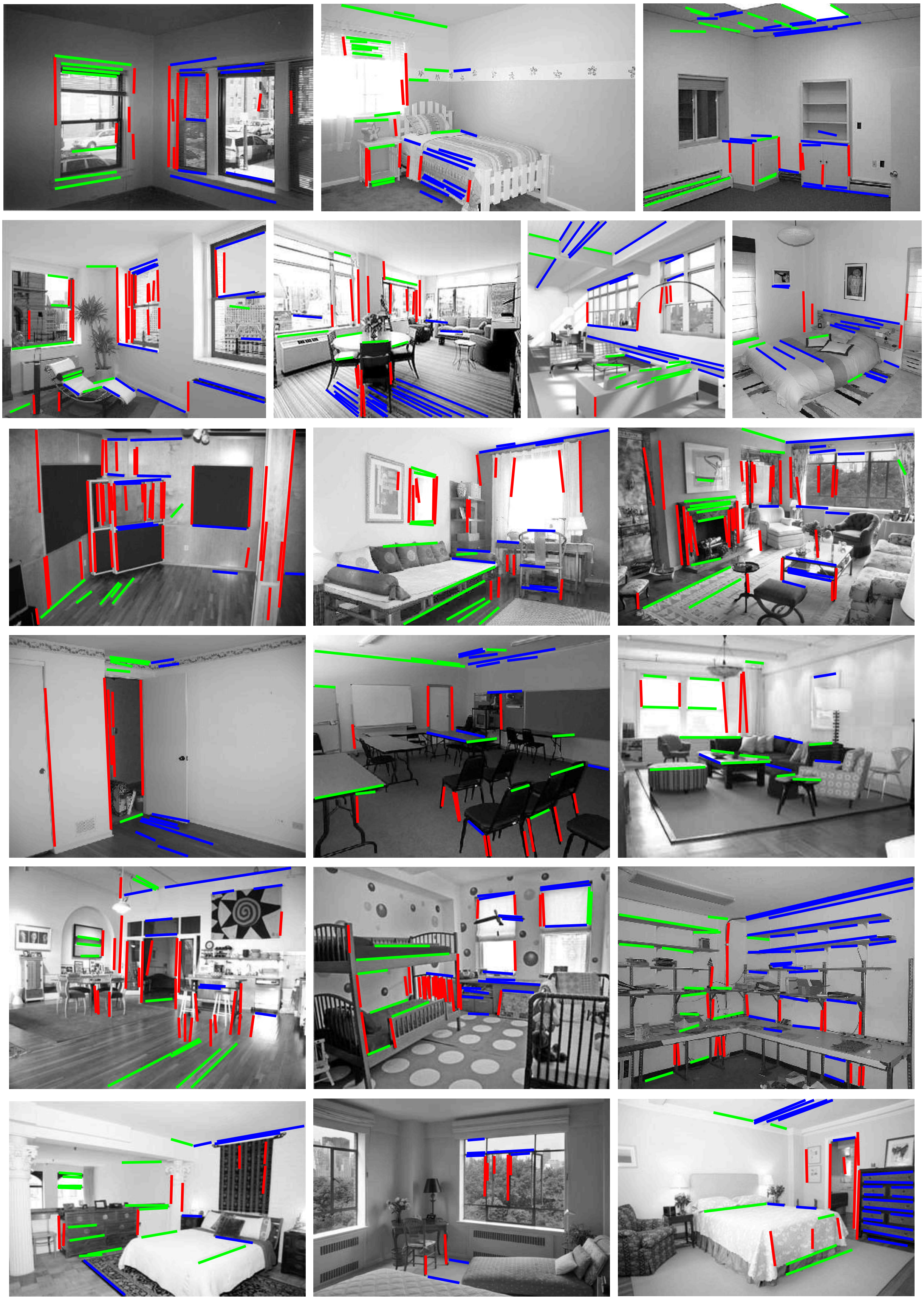


Figure 10: Occluder lines for the same set of images in Fig. 9.

is straightforward: lines forming quadrilaterals that do not lie farthest back are occluders. However, this leaves some unpaired lines unaccounted for, even if they are contained in some quadrilateral. Many results in Fig. 10 can be improved if line-plane grouping is jointly considered.

4. Constraints on depth and extent provided by planes of different orientations. Our depth-ordering of quadrilaterals only concerns the quadrilaterals that point in the same direction. In Fig. 9 (row 2, column 3), there are no upright blue quadrilaterals detected at the far end, making the depth 1 plane, which delineates the spatial frame, far front than it should be. If we take adjacent red and green planes into account, it is obvious that a blue plane must exist farther in the back. Therefore, how planes of different orientations line up in depth need to be considered jointly in order to derive the spatial frame of a room. For example, determining the extent of the floor relies on the constraints provided by the clutters sitting upright on the floor.

4. Summary

We compute a set of depth-ordered planes from a single image with a learning-free grouping method. We show that: **1)** Despite the role of shading and texture gradients at revealing depth in certain scenarios, edges are more ubiquitous and sometimes sufficient for global depth analysis; **2)** Gestalt laws of grouping [21] can be used to reliably evaluate perspectivity; **3)** Compared to traditional shape-from-X methods, our depth from intersection point test is more sensitive to subtle depth differences and more universal in a single image. The relative depth cues allow us to deal with occlusion and weaker perspectivity, the two major difficulties in inferring spatial layout from a single indoor image.

Acknowledgments

We thank Antonio Torralba for some initial discussions and making the vanishing point figure in Fig. 4a. We also thank anonymous reviewers for constructive comments. This research was funded by DOD/ONR-MURI #FD N00014-01-1-0890 grant and NSF CAREER IIS-0644204.

References

[1] M. E. Antone and S. Teller. Automatic recovery of relative camera rotations for urban scenes. In *CVPR*, 2000.

[2] H. G. Barrow and J. M. Tenenbaum. Interpreting line drawings as three-dimensional surfaces. *AI*, 1981.

[3] J. M. Coughlan and A. L. Yuille. Manhattan world: orientation and outlier detection by Bayesian inference. *Neural Computation*, 15:1063–88, 2003.

[4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. In *ICCV*, 1999.

[5] P. Debevec, C. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image- based approach. *SIGGRAPH*, pages 11–20, 1996.

[6] H. L. Erick Delage and A. Y. Ng. A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image. In *CVPR*, 2006.

[7] O. Faugeras. *Three-dimensional computer vision: A geometric viewpoint*. MIT Press, 1993.

[8] J. Garding. Direct estimation of shape from texture. *IEEE PAMI*, 1992.

[9] J. J. Gibson. *Perception of the visual world*. Houghton Mifflin, Boston, 1950.

[10] F. Han and S. C. Zhu. Bottom-up / top-down image parsing by attribute graph grammar. In *ICCV*, 2005.

[11] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[12] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1), 2007.

[13] Y. Horry, K. Anjyo, and K. Arai. Tour into the picture: using a spidery mesh user interface to make animation from a single image. *SIGGRAPH*, pages 225–32, 1997.

[14] K. Ikeuchi and B. Horn. Numerical shape from shading and occluding boundaries. *AI*, 17:141–84, 1981.

[15] T. Kanade. A theory of origami world. *Artificial Intelligence*, 13:279–311, 1980.

[16] T. Leung and J. Malik. Detecting, localizing and grouping repeated scene elements from an image. In *ECCV*, 1996.

[17] A. Lobay and D. Forsyth. Shape from texture without boundaries. In *ECCV*, 2002.

[18] J. Malik. Interpreting line drawings of curved objects. *IJCV*, 1:73–103, 1987.

[19] V. S. Ramachandran. Perceiving shape from shading. *Scientific American*, 1988.

[20] C. Rasmussen. Texture-based vanishing point voting for road shape estimation. In *BMVC*, 2004.

[21] S. Sarkar and K. Boyer. Perceptual organization in computer vision: a review and a proposal for a classificatory structure. *IEEE SMC*, 23:382–99, 1993.

[22] T. Werner and A. Zisserman. New techniques for automated architecture reconstruction from photographs. In *ECCV*, volume 2, pages 541–55. Springer-Verlag, 2002.

[23] S. X. Yu and J. Shi. Segmentation with pairwise attraction and repulsion. In *ICCV*, Vancouver, Canada, 9-12 July 2001.

[24] S. X. Yu and J. Shi. Understanding popout through repulsion. In *CVPR*, Kauai Marriott, Hawaii, USA, 9-15 Dec 2001.

[25] S. X. Yu and J. Shi. Multiclass spectral clustering. In *ICCV*, Nice, France, 11-17 Oct 2003.