Layout Estimation of Highly Cluttered Indoor Scenes Using Geometric and Semantic Cues

Yu-Wei Chao[1], Wongun Choi[1], Caroline Pantofaru[2], and Silvio Savarese[1]

[1] Department of Electrical Engineering and Computer Science,
University of Michigan, Ann Arbor, MI 48109, USA
{ywchao,wgchoi,silvio}@umich.edu
[2] Willow Garage, Inc., Menlo Park, CA 94025, USA
pantofaru@willowgarage.com

Abstract. Recovering the spatial layout of cluttered indoor scenes is a challenging problem. Current methods generate layout hypotheses from vanishing point estimates produced using 2D image features. This method fails in highly cluttered scenes in which most of the image features come from clutter instead of the room's geometric structure. In this paper, we propose to use human detections as cues to more accurately estimate the vanishing points. Our method is built on top of the fact that people are often the focus of indoor scenes, and that the scene and the people within the scene should have consistent geometric configurations in 3D space. We contribute a new data set of highly cluttered indoor scenes containing people, on which we provide baselines and evaluate our method. This evaluation shows that our approach improves 3D interpretation of scenes.

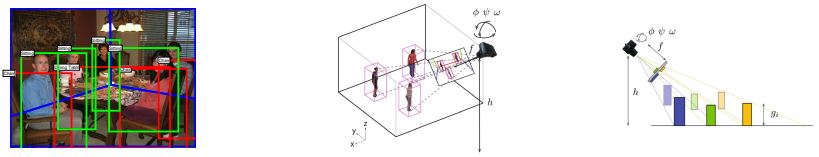
Keywords: scene understanding, vanishing point estimation, layout.

1 Introduction

Enabling machines to understand visual scenes has been a focus of computer vision research. Recently there has been significant work focused on solving for the spatial layouts of indoor scenes [8,10,12,13,15,16]. Given an image of a room, as shown in Fig. 1a, the goal is to automatically identify the extent of the floor, walls, and ceiling as labeled by blue lines. These methods adopt a common procedure for estimating indoor scene layout: 1) detect long straight lines and estimate the vanishing points (VP) corresponding to three orthogonal surface directions, 2) generate candidate layouts, and 3) select the best layout.

Step 1) is very sensitive to clutter in the scene. This step typically relies on associating line segment features (such as the boundaries between walls) to the three VPs [14]. However in cluttered scenes these structural boundaries are often occluded, and the observed lines are instead generated by the clutter of people, chairs, tables, and other objects. So clutter can lead to a poor set of vanishing points, which leads to a poor set of candidate hypotheses, from which even the best layout choice is still wrong. The success of estimating scene geometry hinges on the accurate estimation of the three VPs. There have been previous attempts

A. Petrosino (Ed.): ICIAP 2013, Part II, LNCS 8157, pp. 489–499, 2013.



(a) (b) (c)

Fig. 1. In cluttered rooms (a), room features (blue lines) and objects (red boxes) like dining tables and chairs are severely occluded and difficult to detect. In these cases, people are often easier to detect. We use human detections (green boxes) to estimate the three orthogonal vanishing points of the scene, and then solve for the room layout. Our vanishing point estimation approach (detail in Sec 4) is illustrated in (b) and (c).

to incorporate such clutter into the scene geometry understanding process, such as [8,10,12,16], however they have incorporated clutter reasoning only at the last step of candidate selection. In order to obtain the best possible geometric understanding, we must identify such non-geometric clutter earlier.

The relationship between scene geometry and objects in the scene is a rich source of contextual information which previous work has attempted to exploit. Bao *et al.* [1] uses the 3D locations of detected objects to help estimate the geometric properties of the scene by assuming objects are supported by a common plane. Lee *et al.* [12] explicitly models the relationship between the objects presented in the scene and the scene layout. Unfortunately, in highly cluttered indoor scenes, robust object detection is difficult due to severe occlusions and large intraclass variation (see Fig.3). For instance, the dining table in the middle of Fig. 1a (red boxes) is heavily occluded by the people in the front, while the chairs behind the dining table are occluded by the dining table and the people sitting on them. Furthermore, the two chairs in the front have different shapes. Detecting generic objects is extremely challenging in highly cluttered scenes.

In this paper we follow the intuition that in these types of indoor scenes people can be more robustly detected, as shown in Fig. 1a (green boxes). When people are present in indoor photographs they are typically the focus of the image, and so are less occluded than tables and chairs. [8] also explores a similar concept, but their method is benefitted from the functional regions obtained from accumulating observations of human actions over time. Inspired by the previous work with objects, we adopt the common supporting plane assumption for humans in indoor scenes, and exploit human detection and 3D geometric information to better estimate vanishing points. We show that from those estimated vanishing points we can generate a more robust understanding of scene geometry in highly cluttered environments.

2 Related Work

Scene understanding has attracted interest in the computer vision community of late. Compared to outdoor scenes, indoor environments have richer structure,

allowing the use of stronger priors. Under the Manhattan world assumption, every surface belonging to the scene structure is aligned to one of the three orthogonal directions, which can be represented by three vanishing points (VP) on the image. Lee *et al.* [13] uses the detected wall boundaries on the image to estimate VPs and solve the scene structure accordingly. However, those boundaries are not likely to be observed in practice. Methods have been proposed to estimate layout by modeling the clutter [10,12,16]. Hedau *et al.* [10] identifies the cluttered regions by training a classifier with manually labeled images. Wang *et al.* [16] models the clutter with a latent variable, and applies priors on the appearance to learn the layout model. Lee *et al.* [12] assumes strong geometric features on the cluttered objects, and learns the spatial relationship between objects and layouts. All these methods assume a set of candidate layouts, which is typically generated from detected vanishing points using straight line features (which are more likely to come from the cluttered foreground). Therefore, the generated layout candidates will be inaccurate and limits the performance of final result.

The presence of objects can provide geometric constraints on the scene. Bao *et al.* [1] uses the result of object detection to jointly infer the presence of objects and their support plane. However, object detection is less robust in the face of occlusion and view-point changes, and the results decline with increased clutter. Many human detection techniques have been proposed recently [2,3,7]. So we take the advantage of the fact that people can be detected more robustly than objects in indoor scenes because their discriminative visual features (such as head-and-should silhouette) are less occluded than other objects. Inspired by [1], we parameterize the scene by the ground plane and camera parameters. Instead of using only the camera pitch angle, however, we also model the yaw and roll angle to recover three orthogonal VPs. Note that Fouhey *et al.* [8] also uses people as a cue for layout estimation. The strength of their method relies on estimating the functional regions (e.g. walkable, sittable, reachable) within the image by accumulating observations of human actions over time. However, their method is still based on the VPs and layouts generated by [10].

3 Estimating a Room Layout

We follow [10] and represent an indoor space by a 3D box. In each scene, the camera can observe at most five interior faces of the box model: floor, ceiling, left, center, and right walls. Given the Manhattan world assumption, each pair of the faces are either parallel or perpendicular in 3D. The projection of each face on the image is a polygon, as shown in Fig. 1a. The goal of layout estimation is to identify the boundaries between two faces in the image, (the polygon edges), and recover the 3D box structure of the indoor space.

Our approach follows the general procedure of [8,10,12,13,15,16] to generate the layout of the room. First, we estimate the three orthogonal vanishing points of the scene to obtain the orientation of the 3D box. Different from [10], which estimate the vanishing points solely from image line segments, our method exploits the 3D geometric relationship between people and the room box to jointly

estimate the vanishing points, camera height, and 3D locations of the people (detailed in Sec. 4). Once the VPs are estimated, we follow [10] to generate layout hypotheses by translating and scaling the faces of the box, and finally find the candidate layout which is most compatible with the image observation.

4 Vanishing Point Estimation from Human Detection and 3D Geometric Information

We propose a novel framework for estimating three orthogonal vanishing points using human detections and their 3D geometric relationships with the scene. The intuition behind our method is that people in the scene should have a consistent geometric configuration with the scene layout. Specifically, given that all of the people are the same height, they should share a common supporting ground plane. This intuition is expressed as an energy maximization framework, described in Sec. 4.1. Each component of our model is addressed in Sec. 4.2, and finally the optimization procedure is described in Sec. 4.3.

4.1 The Model

Given an image I, our goal is to jointly estimate the set of 3D human locations H and the scene geometry S. We parameterize the scene geometry by $S = \{f, \omega, \psi, \phi, h\}$, where f is the camera focal length, ω, ψ, ϕ are the roll, yaw, pitch angle (in the order of rotation performed) of the camera, and h is the camera height. The coordinates of three orthogonal vanishing points can be uniquely determined by $\{f, \omega, \psi, \phi\}$, and vice versa [9].

Suppose we obtain N candidate human detections, then we can denote $H = \{B, P, T\}$. $B = \{b_i | i = 1, \ldots, N\}$ represents human detection bounding boxes, with $b_i = \{x, y, width, height\}$. Each person can take one of K poses, so $P = \{p_i | i = 1, \ldots, N\}$ represents people's poses, with $p_i \in \{1, \ldots, K\}$. Finally, each detection hypothesis may or may not be correct, so $T = \{t_i | i = 1, \ldots, N\}$ models the correctness of each detection hypothesis with a binary flag.

Given H and S, the 3D locations of the people can be uniquely determined by back-projecting the bottom of the bounding boxes onto the 3D ground plane, as shown in Fig. 1b. We formulate the estimation of H and S as an energy maximization framework, with energy:

$$E(S, H, I) = \alpha\Psi(S, H) + \beta\Psi(I, H) + \gamma\Psi(I, S). \quad (1)$$

$\Psi(S, H)$ is the compatibility between the scene hypothesis and the human locations, which is the difference between the observed 3D human heights and the expected heights of different human poses. $\Psi(I, H)$ is the compatibility between the observed image and human locations as measured by the human detector score. $\Psi(I, S)$ is the compatibility between the observed image and the scene hypothesis, measured by how well the image line segments fit the hypothesized vanishing points. α, β, and γ are the model weight parameters.

4.2 Model Components

Below we explain each component of the model. Note that the human positions are assumed to be independent in the scene.

Scene-Human Compatibility $\Psi(S, H)$: This potential measures the likelihood of the human location $H = \{B, P, T\}$ given the scene hypothesis S. Assuming that the locations of different people are independent, we have,

$$\Psi(S, H) = \frac{1}{N} \sum_{i=1}^{N} \Psi(S, H_i) \tag{2}$$

We model each human by a pose-dependent cuboid in 3D space. Given S, we first back-project the bottom of the ith person's bounding box onto the ground plane to get the 3D location where the ith cuboid is supported by the ground plane. Assuming the cuboids and the ground plane have the same normal, we can get the top of the ith cuboid by back-projecting the top of ith detection bounding box, as illustrated in Fig. 1c. The 3D height of the ith person detection g_i is the corresponding cuboid height. We apply a prior on the 3D height $\mathcal{N}(\mu_k, \sigma_k)$ for the human pose class k. The potential $\Psi(S, H_i)$ is formulated as

$$\Psi(S, H_i) = \begin{cases} \ln \mathcal{N}(g_i - \mu_{p_i}, \sigma_{p_i}) & \text{if } t_i = 1 \\ \ln(1 - \mathcal{N}(g_i - \mu_{p_i}, \sigma_{p_i})) & \text{if } t_i = 0 \end{cases} \tag{3}$$

Image-Human Compatibility $\Psi(I, H)$: The compatibility between person locations H and image I is defined by the detection confidence as,

$$\Psi(I, H) = \frac{1}{N} \sum_{i=1}^{N} \Psi(I, H_i) \tag{4}$$

where $\Psi(I, H_i)$ is a function of the detection score s_i of b_i. In practice, we take $\Psi(I, H_i) = \ln g(s_i)$, where $g(\cdot)$ is the sigmoid function.

Image-Scene Compatibility $\Psi(I, S)$: This potential measures the compatibility between the observed image line segments and the vanishing points computed from the scene hypothesis S. Following [10], we first detect long straight lines $\{l_n | n = 1, \ldots, L\}$ in I. Then we take $\{f, \omega, \psi, \phi\}$ from the scene hypothesis S and compute the three orthogonal vanishing points v_1, v_2, v_3. As in [10], the lines vote for each vanishing point using an exponential voting scheme. Line l_n votes for vanishing point v_m with a score of

$$V(v_m, l_n) = |l_n| \cdot \exp -(\frac{\alpha_{mn}}{\sigma_V}) \tag{5}$$

, where α_{mn} is the angle between l_n and the line connecting v_m and the midpoint of l_n, σ_V controls the peakedness of the voting score, and $|l_n|$ is the length of l_n. The potential $\Psi(I, S)$ aggregates these votes:

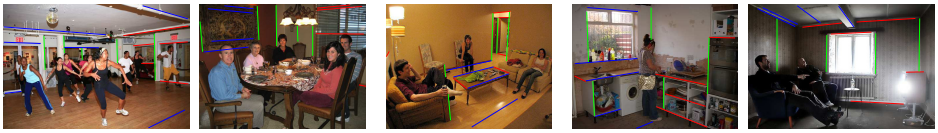
$$\Psi(I, S) = \sum_{m=1}^{3} \sum_{n=1}^{L} V(v_m, l_n) \tag{6}$$

4.3 Solving the Optimization Problem

Given the image I and the human detection B, P, we want to solve the scene information $S = \{f, \omega, \psi, \phi, h\}$ and the presence of the person T. This can be obtained by maximizing the energy in Eq. 1:

$$\{\hat{S}, \hat{T}\} = \max_{S,T} E(S, H, I) = \max_{S,T} \alpha\Psi(S, H) + \beta\Psi(I, H) + \gamma\Psi(I, S) \qquad (7)$$

Since we explicitly model the camera and scene parameters, we can sample a discrete set of parameters values and search for the best combination. We fix a set of uniformly distributed samples for each ϕ and ψ, and normally distributed samples for each f, ω, and h.



	(a) dancing	(b) having dinner	(c) talking	(d) washing dishes	(e) watching TV
FL:	601.24	607.48	418.67	591.26	820.16
Roll:	2.10	0.42	-2.05	0.14	0.31
Yaw:	-44.37	36.85	-40.11	-42.47	26.55
Pitch:	0.26	7.44	17.06	6.20	2.48

Fig. 2. Our collected *Indoor-Human-Activity* dataset is composed of five activity classes: dancing (a), having dinner (b), talking (c), washing dishes (d), and watching TV (e). The top row shows example images with annotated line segments which are used to compute the ground truth of three orthogonal vanishing points. The bottom row shows the camera focal length and angles computed from the vanishing points.

5 Experiments and Results

We aim to evaluate our algorithm on highly-cluttered indoor images which include people. None of the existing datasets were appropriate for this task, so we contribute a new dataset called the *Indoor-Human-Activity* dataset. The dataset contains 911 images of five human activity classes: dancing (187), having dinner (183), talking (193), washing dishes (183), and watching TV (165). Different activity classes contain different levels of clutter, as seen in Fig. 2. For each image, we have annotated the line segments associated to the three principle directions, from which we have computed the ground-truth vanishing points. In addition, we provide annotations of scene layout and human detections, as well as four object classes (sofa, chair, table, and dining table) for future use.

We first evaluate several state-of-the-art object and human detectors on our dataset. Object detectors are trained using DPM [6] on the *furniture* dataset proposed in [5]. For the human detector, we the off-the-shelf DPM detector [6] and the poselet detector [2,3]. Fig. 3 shows precision-recall curves. The human detectors perform better overall than object detectors in every activity class. Among the human detectors, the poselet detector performs best. Therefore we use the poselet detector to provide candidate human bounding boxes.

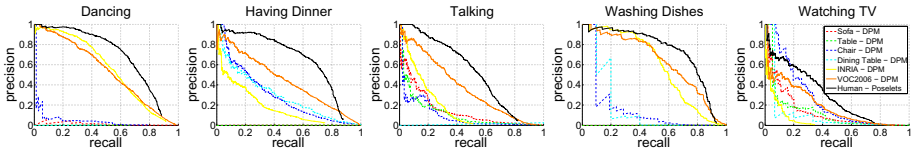


Fig. 3. Precision-recall curves for the DPM object and human detectors [6], and the poselet human detector [2]. People are detected better than objects in our dataset.

Table 1. VP estimation error by Hedau [10], our model without people, with poselet detection and with ground-truth bounding boxes (F: focal length, R: roll, Y: yaw, P: pitch). Our method outperforms the baselines in almost all parameters.

	Dancing				Having Dinner				Talking				Washing Dishes				Watching TV			
	F	R	Y	P	F	R	Y	P	F	R	Y	P	F	R	Y	P	F	R	Y	P
Hedau [10]	346	1.63	9.72	5.84	336	1.96	9.36	6.47	219	1.84	8.60	4.27	179	1.10	4.61	3.66	331	1.57	**8.89**	4.80
W/O HMN	242	1.48	8.36	4.58	206	1.38	9.43	4.94	160	1.39	9.07	4.43	**145**	0.99	4.06	3.20	209	1.30	11.30	5.06
PSLT	**221**	**1.39**	**8.31**	**4.44**	**187**	**1.23**	**8.46**	**3.92**	**130**	**1.21**	**7.87**	**3.10**	147	**0.96**	**3.58**	**2.87**	**197**	**1.28**	10.39	**3.80**
GTBB	226	1.35	8.09	4.13	180	1.17	8.25	3.90	120	1.17	7.34	2.83	131	0.93	3.79	2.80	185	1.30	9.52	3.75

In our implementation, we model humans with two pose classes ($K = 2$): standing and sitting. The prior on 3D heights was set to $(\mu_{stand}, \sigma_{stand}) = (1.68, 0.2)$ and $(\mu_{sit}, \sigma_{sit}) = (1.32, 0.1)$ meters. A SVM classifier is used to classify a person's pose [4]. The classifier is trained using 50 images from each activity class, and the rest are used for evaluating the vanishing points and layout estimation. We consider the predicted human bounding boxes with more than 50% overlap with ground-truth bounding boxes to be our training data. As pose features, we use the weighted poselet activation vector and the ratio between full body and torso heights. A 5-fold cross validation achieved 83% accuracy.

We first evaluate the accuracy of vanishing point estimation (Sec 5.1). In Sec 5.2, we demonstrate that better estimated vanishing points can generate better candidate layout hypotheses, and then we analyze the layout estimation error by different input vanishing points.

5.1 Vanishing Point Estimation

Our goal is to estimate the vanishing points, however comparing vanishing point positions directly is not a good measure of accuracy. This is because the absolute error in vanishing point position increases in sensitivity to inaccurate camera parameters with increased distance from the camera center. A better-normalized comparison is between the camera parameter errors, which we use to evaluate our approach. Given three orthogonal vanishing points, we can uniquely determine the roll, yaw, pitch angles, and the focal length of the camera. Note that we can not evaluate the estimated camera height because the ground-truth can not be obtained from a single image.

We compare the VP estimation results of Hedau *et al.* [10] and three versions of our method: 1) without using human detections (W/O HMN), using only

Table 2. Pixel error of estimated layouts. Our estimated VPs shows comparable results to Hedau's [10].

	Best Candidate			Estimation		
	Hedau [10]	Ours	GT VP	Hedau [10]	Ours	GT VP
Dancing	5.51 %	**4.75** %	3.65 %	**19.74** %	20.24 %	18.36 %
Having Dinner	5.19 %	**5.06** %	3.53 %	24.00 %	**23.92** %	21.87 %
Talking	5.12 %	**4.83** %	3.61 %	23.84 %	**20.58** %	19.89 %
Washing Dishes	**3.58** %	3.80 %	3.51 %	**26.30** %	27.63 %	25.48 %
Watching TV	**4.94** %	5.87 %	3.60 %	**19.14** %	22.74 %	18.28 %

Table 3. Intersection/union of observable 3D space between estimated and ground-truth layouts. Our estimated VPs outperforms Hedau's [10] in all activity classes due to better 3D reasoning.

	Best Candidate			Estimation		
	Hedau [10]	Ours	GT VP	Hedau [10]	Ours	GT VP
Dancing	43.99 %	**50.95** %	83.32 %	17.60 %	**24.25** %	46.95 %
Having Dinner	51.15 %	**61.19** %	90.51 %	24.75 %	**35.82** %	52.08 %
Talking	60.91 %	**65.03** %	90.59 %	34.26 %	**40.24** %	53.32 %
Washing Dishes	68.94 %	**70.08** %	90.62 %	32.78 %	**33.90** %	46.76 %
Watching TV	51.01 %	**57.88** %	89.70 %	27.84 %	**33.08** %	55.83 %

$\Psi(I, S)$, 2) using poselet detection (PSLT), and 3) using ground-truth human bounding boxes (GTBB) to remove the detection error and provide a lower bound on the error. Table 1 contains the average errors for each activity class.

First we observe that our partial method obtains comparable or better results than [10] in most parameters. This is because [10] generates the VP hypotheses by the intersection of lines in 2D, while we parameterize the VPs by 3D camera parameters. We can prune out some unlikely hypotheses by putting priors on the parameter search space. Using poselet human detection, our full method outperforms the baselines in almost all activity classes. Since the roll angles are generally very small, the back-projected human height depends mostly on the pitch angle of the camera. And indeed, our approach improves the pitch angle most, as can be seen in the Table 1 columns labeled 'P'. The amount of error also reflects the level of clutter in different activity classes; images from "washing dishes" contain less clutter than "having dinner", and so the error rates show less improvement when human detections are used. Qualitative examples of VP estimation are shown in the first five rows of Fig. 4.

5.2 Room Layout Estimation

We compare the estimated layouts obtained by Hedau's VPs [10], our estimated VPs, and the ground-truth VPs. In most literature [8,10,12,13,15,16], layout estimation are evaluated based on the 2D pixel error, i.e. the percentage of pixels that is labeled different from the ground truth. However, as suggested in [11], good 2D estimation does not usually indicate good 3D estimation. To provide 3D evaluation, we propose a new metric for evaluating layouts: intersection-union of observable 3D space.

Fig. 4. Qualitative results for vanishing points (first five rows) and layout estimation (last five rows). Row 1: image with annotated line segments and the ground-truth VPs. Row 2: detected line segments and the VPs computed by Hedau *et al.*'s method [10]. Line segments are colored with the associated vanishing points (green: vertical, red: further horizontal, blue: close horizontal, cyan: not associated). Row 3: input human detection to our method. Rows 4 & 5: output of our method. The line association has been improved using our method. Rows 6-8: the generated layouts using ground-truth VPs, Hedau's estimated VP, and our estimated VPs (green: ground-truth, yellow: best candidate, red: estimated layout), along with the corresponding pixel errors. Rows 9 & 10: the observable 3D space of best candidate and estimated layouts (red: ground-truth, blue: [10], green: ours, cyan: GT VP).

First, we evaluate the layout estimation by the commonly-used pixel error. In Table 2, we report both the error of best candidate layout (the oracle result) and the estimated layout. Layout candidates are generated by sampling 20 rays from each VP [10]. Observe that by improving vanishing point estimation, the best candidate layout can achieved lower pixel error. However, for estimated layouts, we obtain comparable results using Hedau [10] and our VPs, and slightly better results using ground-truth VPs. 2D metrics can not fully capture the difference in 3D space. As in the watching TV example in Fig. 4, bad VP estimation gives the same or even better estimated layout in terms of pixel error.

To show that our method can achieve better 3D estimation, we propose a new 3D metric: intersection/union of observable 3D space between the estimation and ground-truth. Assuming a fix camera height, the observable 3D space is obtained by back-projecting the observable 2D layout extent into the 3D space. This is determined by the camera focal length, angles, and the 2D layout, as shown in the last two rows of Fig. 4. Similar to pixel error, we report the result for both the best candidate and estimation in Table 3. Our method outperforms [10] with respect to 3D reasoning about the scene.

6 Conclusion

Understanding the geometric structure of a room is an important stepping stone on the way to understanding the semantic content of an indoor image. In this paper, we have provided a method for improving the computation of geometric room structure from a single image by using human detections in the scene. Since humans are often the focus of the scene, they are more frequently detected than other objects, and so provide robust information which complements previously used line segments as features. We have contributed a new Indoor-Human-Activity dataset and provided experiments that show that our method improves upon previous scene geometry understanding by increasing the accuracy of line segment associations, vanishing points, and in turn 3D structural plane boundaries, camera height and camera focal length. We look forward to applying this method to future work on indoor activity understanding.

References

1. Bao, S.Y., Sun, M., Savarese, S.: Toward coherent object detection and scene layout understanding. In: CVPR (2010)
2. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 168–181. Springer, Heidelberg (2010)
3. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV (2009)
4. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (2011)
5. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Understanding indoor scenes using 3d geometric phrases. In: CVPR (2013)

6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Discriminatively trained deformable part models., http://people.cs.uchicago.edu/pff/latent-release4/
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI (2010)
8. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single-view geometry. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 732–745. Springer, Heidelberg (2012)
9. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press (2004) ISBN: 0521540518
10. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV (2009)
11. Hedau, V., Hoiem, D., Forsyth, D.: Recovering free space of indoor scenes from a single image. In: CVPR (2012)
12. Lee, D.C., Gupta, A., Hebert, M., Kanade, T.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: NIPS (2010)
13. Lee, D.C., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: CVPR (2009)
14. Rother, C.: A new approach for vanishing point detection in architectural environments. IVC (2002)
15. Schwing, A.G., Hazan, T., Pollefeys, M., Urtasun, R.: Efficient structured prediction for 3d indoor scene understanding. In: CVPR (2012)
16. Wang, H., Gould, S., Koller, D.: Discriminative learning with latent variables for cluttered indoor scene understandingy. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 497–510. Springer, Heidelberg (2010)