Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry

Varsha Hedau[1], Derek Hoiem[2], and David Forsyth[2]

[1] Department of Electrical and Computer Engineering
[2] Department of Computer Science,
University of Illinois at Urbana Champaign
{vhedau2,dhoiem,daf}@uiuc.edu

Abstract. In this paper we show that a geometric representation of an object occurring in indoor scenes, along with rich scene structure can be used to produce a detector for that object in a single image. Using perspective cues from the global scene geometry, we first develop a 3D based object detector. This detector is competitive with an image based detector built using state-of-the-art methods; however, combining the two produces a notably improved detector, because it unifies contextual and geometric information. We then use a probabilistic model that explicitly uses constraints imposed by spatial layout – the locations of walls and floor in the image – to refine the 3D object estimates. We use an existing approach to compute spatial layout [1], and use constraints such as objects are supported by floor and can not stick through the walls. The resulting detector (a) has significantly improved accuracy when compared to the state-of-the-art 2D detectors and (b) gives a 3D interpretation of the location of the object, derived from a 2D image. We evaluate the detector on beds, for which we give extensive quantitative results derived from images of real scenes.

1 Introduction

We spend much of our lives in a box. We eat, work, and sleep in areas that are limited by orthogonal planes, populated with carefully arranged furniture. Yet, despite their importance and rich structure, such environments are highly challenging for current recognition methods, largely because the near-field objects do not conform to orthographic projective assumptions.

In this paper, we propose an approach to *think inside the box*, building tightly constrained models of appearance and interactions of objects in a way that reflects the dominant structure of the indoor scene. Our assumption is that the objects are aligned with the dominant directions of the scene. This allows us to integrate global scene orientation cues in an object appearance model, which considerably simplifies the otherwise challenging view invariant object detection. Using the perspective inside the room, we recover an approximate 3D localization of an object, which further facilitates incorporating even richer spatial interactions between objects and room's global geometry.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part VI, LNCS 6316, pp. 224–237, 2010.

Fig. 1. We "think inside the box" to detect objects, which are modeled as axis aligned cuboids (shown in yellow) with the scene. The scene is represented as a box layout from Hedau et al. [1] (shown in red). By using the surrounding scene perspective to help model appearance, we can improve detection and localize the object in 3D. Furthermore, we show that supplying more information about the spatial interactions with the scene layout produces better detection.

We build on our earlier work [1] to obtain estimates of the room layout, which is modeled by a 3D oriented 'box' and a pixel labeling of major surfaces. Our focus is to use that layout information to improve object recognition in two key ways. First, we adapt the 2D sliding window detector strategy to a 3D domain, searching for parts in frontal-rectified features and stitching them together with a sliding 3D cuboid. Second, we model the relations of the objects with respect to the room, encoding soft constraints of size, visibility, and likely position within the room. In experiments on bed recognition for indoor scenes, we demonstrate that both of these innovations yield significant improvements.

1.1 Related Work

Our work builds on a wide range of techniques from literature on object recognition, 3D scene modeling, and contextual reasoning. In object recognition, the idea of sliding window detection with statistical templates has long been a mainstay [2,3,4,5] due to its simplicity and effectiveness for many categories. Within this framework, Dalal and Triggs [6] demonstrate that spatially local histograms of gradient (HOG) are effective features for pedestrian detection. More recently, Felzenszwalb et al. [7] extend this model to allow deformable latent parts, each modeled by its own HOG-based statistical template detector. We extend these ideas to 3D. Our object models are cuboids, composed of 3D planar parts whose orientations are defined with respect to the dominant orientations of the room. Like recent work [6,7], we detect these parts using HOG-based detectors, but our gradient images are frontally rectified such that rectangles in 3D become rectangles in the rectified image. This modification makes our detector invariant to viewpoint and is necessary for the near field case of indoor scenes, where object orientation changes rapidly with its location in the image. Similar to 2D

sliding window detectors, we search for the objects by scanning a 3D cuboid at increments of ground plane position and scale.

Several recent works (e.g. [8,9]) also explore 3D-based object models, typically modeling objects as a collection of affine-transformed parts that have some spatial relation to each other or to the object center. These methods require complicated processes to align parts across views and instances and to build categorical spatial models. Because we annotate the corners of objects in training, our training and inference processes are very simple but effective. We have found our 3D-based features to be complementary to existing 2D models and show that a combination outperforms either alone.

Our work also adds to numerous efforts in image-based 3D scene estimation [10,11,12,13,14] and contextual reasoning [15]. Many previous approaches such as [16,17,18] use context in form of rough geometric constraints such as relative location and depth estimates, to improve object recognition in 2D. Our goal is to recover full 3D spatial extent of an object coherent with the surroundings, which requires stricter and richer constraints. The planar parts of our object cuboids are oriented 3D rectangles, which were shown to be useful for structure modeling in [19,20]. We build on our indoor spatial layout method [1], which estimates the principal orthogonal vanishing points, a 3D box layout, and a pixel labeling of the room surfaces. We use this method to obtain the orientation and our initial estimate of room layout.

Our probabilistic contextual reasoning most closely resembles works by Hoiem et al. [21] and Leibe et al. [22]. Like Hoiem et al. [21], we softly enforce size consistency through probabilistic inference. However, our 3D object models allow us to avoid making assumptions of roughly level cameras and orthographic projection, which is crucial for estimating object size in the near-field. Leibe et al. [22] detect and track objects from a moving video while constraining the objects to lie within the recovered street corridor. Because our scene estimates are recovered from a single image, we marginalize over the layouts while softly constraining that objects should lie within the room. Additionally, we model the spatial position of objects with respect to the walls.

1.2 Overview of Our Approach

Fig. 2 shows the overview of our approach. We start by detecting vanishing points corresponding to 3 orthogonal directions of the world using the vanishing point method from [1]. This gives us the orientation of object cuboids. By assuming that objects rest on floor we generate object candidates at several scales and translations by sliding a cuboid on floor planes at several different heights below camera. Fig. 2(c) shows some sample candidates obtained by our approach. Object cuboid is represented in terms of its planar sides or 'faces'. We detect objects by searching for their axis aligned faces using rectified gradient features as shown in Fig. 2(d). Fig. 2(e) shows several detected horizontal and vertical cuboid faces parallel to the room orientation. These responses are combined together to score the object cuboids which are further refined probabilistically by using the object and scene layout interaction. Fig. 2(f) shows the detected object.

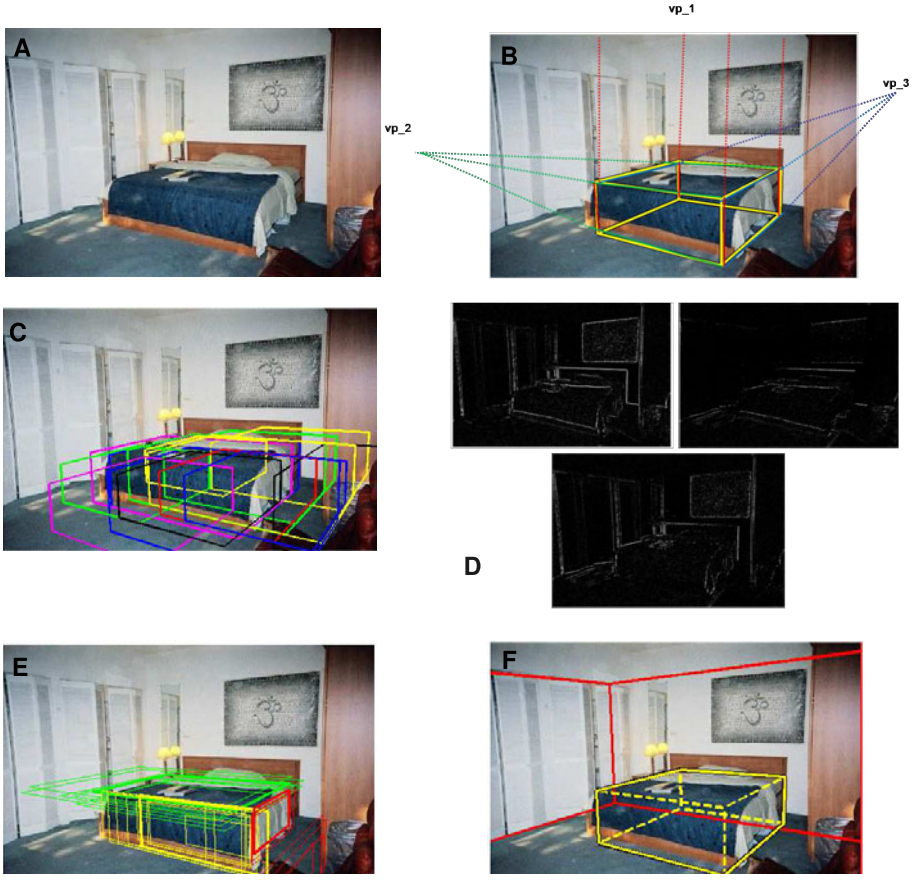

Fig. 2. Our algorithm takes original image and estimates the vanishing points of the three orthogonal directions of the world using method of [1]. This fixes the orientation of objects. We then generate many object candidates by sliding a cuboid in 3D. A sample of candidates are shown with different colors in (c). We detect cuboids by searching for their axis aligned 'faces' using rectified gradient features(shown in (d)). Some face samples with high response in each of the three orientation are shown in (e) with red, green and yellow. Bed cuboid detected by this procedure is further refined with the constraints provided by box layout of scene (shown in red) using a simple probabilistic model. The highest scoring bed cuboid is shown in yellow.

2 Detecting Objects

Classical sliding window approaches provide 2D localization of objects and cannot be easily used to predict 3D location and extent. Treating the objects as 2D planar cardboards has a disadvantage of knowing very little about their spatial interaction with the background. Indoor scenes are highly structured, and this information can be used to obtain a reasonable 3D localization of an object,

which is our target in this paper. By assuming that the faces of object cuboids are parallel to the walls, the orientation of objects can be obtained from the rooms orientation given by vanishing points. Following this we build a searching strategy by sliding a cuboid in 3D to obtain object hypotheses and looking for consistent projected gradients in the image to score these hypotheses effectively.

We model objects as cuboid shaped boxes resting on floor in 3D. Most of the objects in indoor settings can be approximated by cuboid-like structure, e.g., beds and other furniture. Also, in a typical setting, the objects are parallel to room walls. Towards detecting objects in indoor scenes, we thus first estimate vanishing points corresponding to the three orthogonal directions of the room, which also serve as the vanishing points for the objects. These vanishing points fix the object orientation with respect to the camera. To estimate its translation, we generate object hypotheses constrained according to the vanishing points. For each of these hypotheses, we extract specialized features using perspective cues, and score them using a function learned from training images. In this paper we evaluate our method on beds, but, in principle our modeling procedure can also be extended to other similar objects such as chairs, cupboards, and tables.

2.1 Generating Object Hypotheses

To estimate the orientation of object cuboids in the image, we first estimate the vanishing points corresponding to the three orthogonal directions of the room. We use the method of Hedau et al. [1] for vanishing point estimation, which builds upon the method of [23]. The method detects long straight lines in the image. The lines corresponding to principal orthogonal directions in 3D should intersect at the corresponding vanishing points. The method thus accumulates votes for intersection points of these lines based on angular distance between the lines and the points. A triplet of points that gathers maximum votes and satisfies the orthogonality constraints gives three vanishing points. To speed up the search for this triplet, the space of points is quantized with variable bin size increasing as one goes away from the center of the image. Using the vanishing points one can estimate the rotation of the camera with respect to the room (and objects), as well as camera intrinsic parameters [23]. We next describe how we generate object hypotheses using the information of vanishing points.

Given the vanishing points corresponding to the three orthogonal directions of the room, $\{\overline{vp}_i\}_{i=1}^3$, assuming a camera with zero skew and square pixels, one can estimate the camera intrinsic matrix K and its rotation with respect to the room, R. Let us consider the coordinate system centered at the camera optical center whose axes are along the room directions: x-axis along the room width (left to right), y-axis along room height (bottom to top), and z-axis along room depth (towards the camera). For a point $\overline{X}$ in this coordinate system, its homogeneous coordinate projection in image plane $\overline{x}$ can be computed using the following projection relation.

$$c\overline{x} = KR\overline{X} \tag{1}$$

We make the following assumptions about the objects:

1. The object planes are parallel to the walls. It is possible to search angles that are not parallel to any wall, but we keep to the three principal directions in this work.
2. The object base touches the floor. This is true for most furniture in a room. Given a reference base corner point $\overline{X}$ of the object in 3D, the other k corner points $\{\overline{X}_i\}_{i=1}^{k}$ can be computed for given dimensions of the object cuboid.

To generate object hypotheses, we first fix the camera height h_c to an arbitrary value. Any object base corner point lying on the floor $\overline{X}$, should thus satisfy $\overline{X}^T n + h_c = 0$, where $n = (0, 1, 0)$ is the normal to the floor plane. We use this constraint to fix the reference base corner point of the object; the other corners are computed using object dimensions. The projections of these corners in the image can be computed using equation (1). We generate these object hypotheses for different discrete values of camera heights and object dimensions. For our experiments in Sec. 4 we vary camera height from 2.5 ft. to 8.5 ft. with 1 foot increments and use the aspect ratios of $2.5 \times 6 \times 5$ and $2.5 \times 7 \times 6$ ft. for beds in 3D. Note that the extent of floor plane for a camera height is bounded by the horizon line, the vanishing line joining the horizontal vanishing points. We use this constraint to limit the number of generated hypotheses. We typically get $4K$ to $30K$ object hypotheses per image.

2.2 Scoring Object Hypotheses

Part based approaches to modeling objects have shown good results. We model an object cuboid $\overline{c}$ as a collection of its faces, i.e., $\overline{c} = \{f_i\}_{i=1}^{F}$, where F is the number of faces. Given the configuration of the object, some faces are occluded, hence we attach a face visibility variable $\overline{v} = \{v_i\}_{i=1}^{F}$. Since the perspective distortion of each face is known, we extract features from each face after correcting for this distortion, denoted by $G = \{\overline{g}_i\}_{i=1}^{F}$ (see features described next). We independently score each face using a linear function on respective features, $s(f_i) = \overline{w}_i^t \overline{g}_i$, where $\overline{w}_i$ is weight vector learned from linear SVM. To deal with variations in object dimensions and for better localization, we allow the individual faces to deform slightly. For this, we modify the score of each face by the best scoring face in the neighboring object hypotheses $f_j \in \mathcal{N}(f_i)$. The final score of an object hypothesis $\overline{c}$ is thus given by

$$scr(\overline{c}) = \frac{\sum_i v_i \max_{f_j \in \mathcal{N}(f_i)} s(f_j)}{\sum_i v_i} \tag{2}$$

Features. Standard histogram of oriented gradients (HOG) features implicitly assume that the 2D image projection is the natural coordinate frame in which to view objects. Our method supposes that a 3D coordinate frame oriented with the room is better. We bin the gradients of image with respect to the direction of each pair of vanishing points. We use 6 orientation bins in our experiments. Fig. 2 (D) shows gradients binned in directions of each pair of vanishing points. Efficient

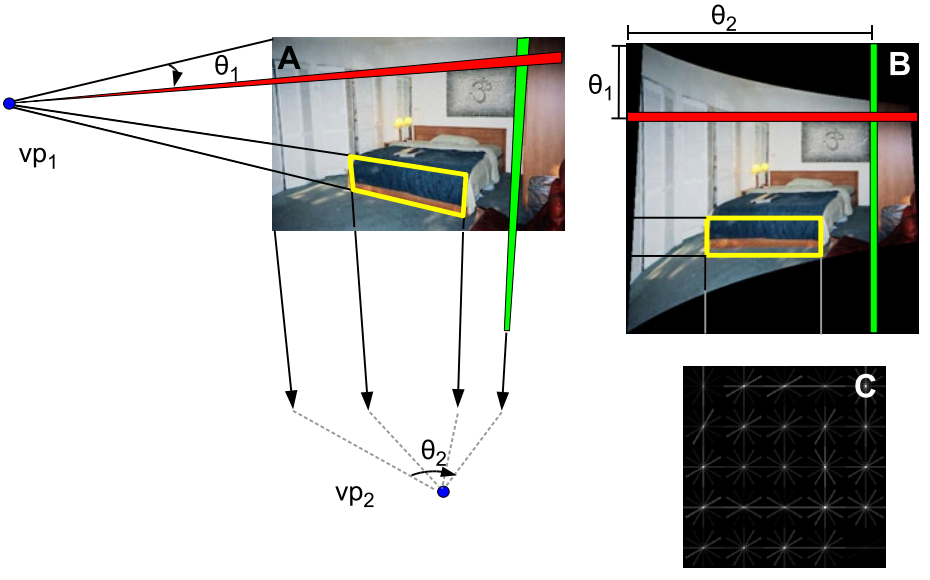

Fig. 3. Computing rectified HOG. We construct the orientation histograms with respect to the surrounding scene orientation (vanishing points directions). Fig. 2(D) shows gradients binned in direction of each pair of vanishing points. We further rectify these gradient images by using a transformation induced by indexing pixels in original image by the angles they subtend at the vanishing points. Note that gradient images are rectified and the original images are shown here only for simplicity. Such a rectification allows efficient computation of HOG in rectangular regions via integral images.

computation of HOG features is possible in rectangular windows however the projection of oriented rectangles in 3D are not rectangular in images. For this reason we compute histogram of gradients for a face in rectified coordinates corresponding to its vanishing points where it is frontal. Each face is divided into 5×5 cells and local normalization is done as described in [7]. Fig. 3(A,B) illustrates this rectification for a face with vanishing points vp_1, vp_2 and the corresponding HOG features are shown in Fig. 3(C). For simplicity we show the rectification on the original image however in principle gradient images, Fig. 2(D) are rectified. For computing face score HOG features computed with respect to the vanishing points of that face are used. Apart from HOG features we also use the line based features which is count of number of line pixels consistent with the orientation and average object label confidence, obtained from surface label estimates of [1] for each face.

Integrating the scores of a 2D Detector. There has been noticeable progress in 2D recognition approaches. We show how our cuboid detector can easily benefit from the state of art 2D methods [7]. Towards this we add the score of detector [7] in the bounding box of the cuboid to the cuboids score. Sec. 4 shows that we obtain a improved detector by incorporating the information from 2D detector.

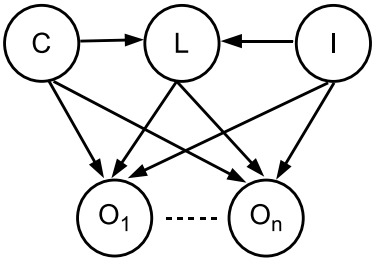

Fig. 4. Joint model of objects and the spatial layout. Objects are independant given layout and camera, leading to simple inference. Several spatial constraints such as objects can not penetrate walls and tend to occur at certain distances from wall can be encoded via this model (read text for more details). We show that incorporating such constraints leads to significant improvement in detecting objects.

3 Modeling the Interaction between Objects and Spatial Layout

Objects live in the scene and thus have to follow certain constraints due to the structure of the scene. These constraints can be used to help improve object detection. Towards this we propose to explicitly model the spatial interactions of objects with scene in a simple probabilistic framework. For scene structure we use our previous work, [1]. This work describes spatial layout of scene in terms of (a) a box layout that defines extent of walls, floor and ceiling, and (b) surface layout that gives pixel labeling of different surfaces as walls, floor, ceiling and objects. We have obtained the the spatial layout estimates on our images using the trained models from [1]. The choice of this spatial layout representation is intuitive for reasoning about spatial interaction between objects and the scene.

The box layout provides extent of walls and floor. The objects inside the box can not cross the boundaries of the walls. Also some objects tend to appear in certain spatial configurations with respect to the box. For instance, beds inside the rooms tend to be close to the walls. Thus knowing the extent of walls and floor provide important information about the placement of objects. Similarly an estimate of location of different objects inside the image can be used to refine the extents of wall floor boundaries.

Towards joint reasoning of objects and layout we propose a simple generative model, shown in Fig. 4. Here, $\{O_i\}_{i=1}^{N}$, $O_i \in \{0, 1\}$ are the object variables, O_i is whether a particular object is present or not, N is the number of objects, L is the box layout of the scene, C is the camera height and I is the image evidence. We consider all the detections left after doing a soft non-max suppression on the output of our cuboid detector (Sec. 2). The non-max suppression step greedily selects the highest the scoring detections while rejecting the ones that overlap more than a certain threshold with the existing selected detections. We use the thresold of 0.85 in our experiments. In this paper we have used beds as objects,

however our framework is general enough to be applicable to other objects as well. The joint distribution over of objects, layout and camera can be written as

$$P(O_1, \ldots, O_N, L, C|I) = P(C)P(L|C, I)\prod_{i=1}^{N} P(O_i|L, C, I) \tag{3}$$

Here, $P(C)$ is the prior on camera height, assumed to be a Gaussian with mean $\mu = 5.5$ ft. (about eye level) and standard deviation $\sigma = 3$ ft. $P(L|C, I)$ is the layout likelihood conditioned on the camera which is estimated using layout scores obtained from the layout detector of [1] and features such as box layout height and depth given the camera height. $P(O_i|L, C, I)$ is the object likelihood conditioned on the layout and camera modeled as a logistic function given by,

$$P(O_i|L, C, I) = 1/(1 + \exp(-w^T \phi(O_i, L, C))) \tag{4}$$

where $\phi(O_i, L, C)$ is the feature set, consisting of (1) Scores from our object detector (described in Sec. 2); (2) Inferred object height given the camera height and horizon; and (3) Object-layout interaction features (described next). Objects are assumed to be independent given layout and the camera after non-max suppression, which leads to simple inference. We compute object marginals over a discrete set of sample values for camera heights and box layout. In our experiments we marginalize over top 100 layouts returned by the method of [1].

3.1 Interaction Features

We propose the object and layout interaction features which model 3D spatial constraints. This is possible due the 3D localization of objects provided by our object detector and the 3D extent of walls, floor obtained from [1]. As interaction features we use (a) overlap between object's footprint and the floor as an indicator of the extent of object sticking outside the floor i.e. into the walls; (b) distance between object and the walls which is computed as distance between the object and the nearest wall boundary, capturing the tendency of objects to occur at fairly consistent positions with respect to the layout.

Each of the above conditional likelihood is trained using logistic regression. The outputs of logistic regression are well calibrated probabilites. Inference is exact and straightforward on the above model.

Table 1. Average Precision (AP) for beds. Our 3D Cuboid detector is comparable with the state-of-art 2D object template detection method of Felzenszwalb et al. [7]. The combination of two detectors results in improvement in performance over each. The precise 3D extent of object provided by our cuboid detector facilitates incorporation of richer scene context which improves object detection significantly further.

Method	1.Cuboid detector	2. Felzenszwalb et al.	1+2	1+2+scene layout
Average Precision	0.513	0.542	0.596	0.628

4 Experiments

We evaluate our object detector and the joint layout model for beds on a dataset of 310 images of indoor scenes collected from Flickr and LabelMe [24]. We have labeled ground truth corners of beds in these images. We split this set randomly into 180 training and 130 test images.

Cuboid face detectors are trained using only bed images. We train one common detector for all the vertical faces of a cuboid and one for horizontal face. Faces that have overlap less than 50% with the ground truth are used as negative samples for the detector. We train the face detector using a linear SVM. For a face we allow some deformation by choosing its score as the maximum amongst all the faces having more than 75% overlap with it. Fig. 7 shows the precision-recall curves for our bed detector. Precision is the number of correct detections, and recall is the number of objects that are retrieved. We compare

Fig. 5. Examples of high scoring detection selected from the first 100 top ranked detections of our algorithm on the test set images. First four rows show true positives, ground truth positive that are detected as positive and last row shows examples of false positives negatives that are detected as positives. Many false positives such as the the dining table and the sofa are due to high response of our detector to the strong oriented gradients in these areas.

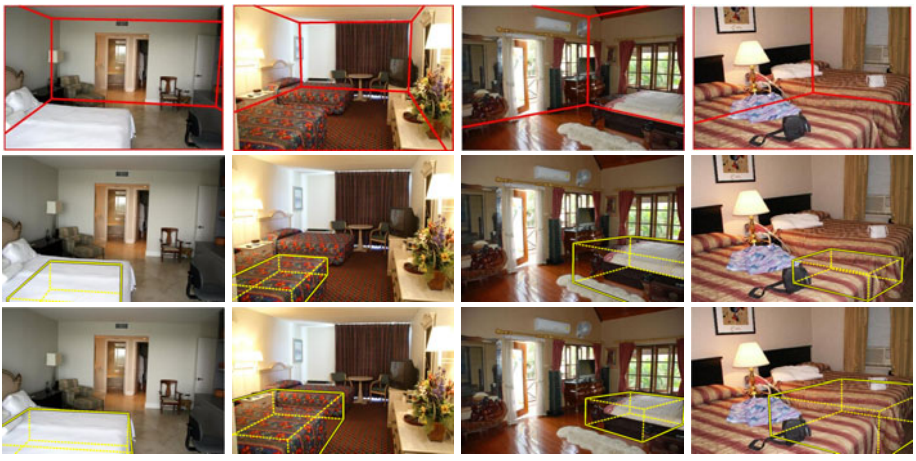

Fig. 6. Examples of improved object detections by joint modeling of objects and the scene layout. First row shows the best ranked box layout of scene obtained from Hedau et al. [1]. Second row shows the highest scoring beds by our cuboid detector in each image. Third row shows the highest scoring best detection obtained from our joint model. Note that first row shows only the best box layout for an image ranked by the box layout detector of [1], the bed detection is however obtained using marginal estimates of objects over discrete sample set of multiple high scoring box layouts. Notice how the joint model captures the tendency of beds occurring close to wall and the camera height prior prunes out the the detections with wrong scale estimates

our detector with the state of art 2D detector of Felzenszwalb et al. [7], which we train on our dataset. We use evaluation criteria similar to VOC Challenge. Precision recall curves are obtained for bounding box of the detected cuboids in order to compare with our baseline [7], which outputs bounding boxes. Average precision(AP) is computed over the entire test set. Our cuboid detector for beds has AP of 0.513 vs. 0.542 for the well-engineered 2D detector of [7].

To evaluate the additional information that is captured by our cuboid detector as compared to the 2D detector of [7] we combine the detection scores of this detector with our cuboid detector scores. For this we simply add to our score, the score of this detector in the bounding box of the cuboid. We obtain an improvement of 0.05 AP over [7] and 0.08 AP on our original cuboid detector (see Table 1). This suggests that 2D and cuboid detectors each have information to contribute to the other. We also show precision-recall curves in Fig. 7(b) computed using overlaps of the projected convex hull of the cuboids. Note that this is a stricter localization criterion as it also requires the object faces to overlap. We get similar improvement in performance by adding the score of 2D detector to our cuboid detector.

In Fig. 5 we show selected high ranked true positives (first four rows) and false positives (last row) of our cuboid detector. The cuboid detector often accurately localizes the bed in the images. It confuses other objects that have strong oriented

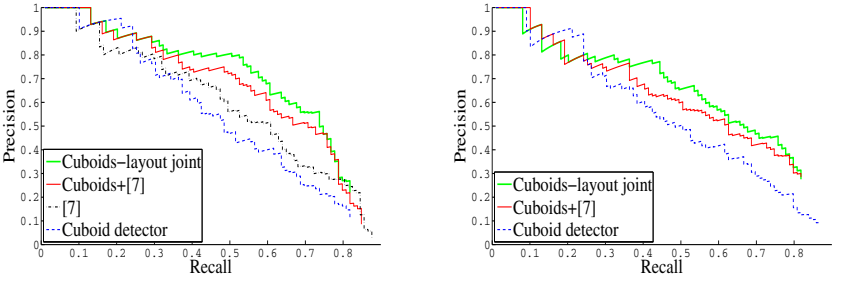

Fig. 7. Precision-Recall curves for bed cuboid detector trained on our dataset of indoor images, (computed as bounding box overlap in **(a)**). We compare our method (blue curve) with the state of art 2D object template detection method of Felzenszwalb et al. [7] (black curve). Better performance is achieved by combining the scores of 2D detector with our cuboids suggesting some amount of complementary information provided by each. Cuboid detector is further improved by using the interactions with scene layout via a joint model (green curve). In **(b)**, we show the precision-recall curves computed using overlap of convex hull of cuboids. Here we achieve results similar to **(a)**. Note that this is a stricter criterion for evaluating localization. We can not compute this measure for [7] since its output is a bounding box.

gradients on them as beds (5th row, 3rd and 4th column). As seen Fig. 5, the detector is robust to cropping (3rd row, 1st col.), occlusion (4th row, 2nd col.), and clutter (4th row, 4th col.).

Finally we evaluate the performance of our joint object layout model. We achieve an AP of 0.628 from marginal estimates of objects obtained from our joint model. Fig. 6 shows several examples of improved object detections obtained by joint reasoning of the box layout, camera and the object cuboid. Notice how the interaction features of object and box layout helps to push the beds closer to the walls. The camera height prior helps in pruning out the detects with unlikely dimensions in 3D.

5 Conclusion

We have developed a detector to locate objects of a specific geometry in an indoor scene, while using object geometry, scene geometry, and their mutual arrangement. Using just a single image, the detector computes object localization in 3D that includes its location, orientation and extent, which is a lot more information when compared to 2D object detectors. The 2D localization performance of the detector is comparable to the state-of-the-art. When we combine our detector with a state-of-the-art 2D detector, there is a significant boost in performance, which indicates that the geometric constraints are highly informative.Furthermore, the visual results indicate that the detector can localize the object nicely, upto the level of its individual parts.

Such a 3D object detector can be used for generating a complete 3D layout of an image, which can in-turn aid graphics applications such as free space estimation, 3D walkthroughs, and image editing. In this paper, have demonstrated the concept of a sliding cuboid detector for a single object category, i.e., beds. However, in principle, the algorithm and the techniques discussed in this paper can also be extended to other objects such chair, sofa, table, dresser etc. Each of these objects can be modeled as a cuboid or as a cuboid with attached back rest, for instance chair and sofa. Likewise, our contextual framework could be extended to include other objects and people, with the goal of producing a complete, coherent parse of an image.

Acknowledgements. This work was supported in part by the National Science Foundation under IIS -0916014 and in part by the Office of Naval Research under N00014-01-1-0890 as part of the MURI program.

References

1. Hedau, V., Hoiem, D., Forsyth, D.A.: Recovering the spatial layout of cluttered rooms. In: Proc. ICCV (2009)
2. Sung, K.K., Poggio, T.: Example based learning for view-based human face detection. Technical report, Cambridge, MA, USA (1994)
3. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. In: CVPR, p. 203. IEEE Comp. Society, Los Alamitos (1996)
4. Schneiderman, H., Kanade, T.: A statistical model for 3-d object detection applied to faces and cars. In: CVPR (2000)
5. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV 57 (2004)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
7. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI 99 (2009)
8. Hoiem, D., Rother, C., Winn, J.: 3d layoutcrf for multi-view object class recognition and segmentation. In: CVPR (2007)
9. Su, H., Sun, M., Fei-Fei, L., Savarese, S.: Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In: ICCV, Kyoto, Japan (2009)
10. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3-d scene structure from a single still image. In: PAMI (2008)
11. Lee, D., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: Proc. CVPR (2009)
12. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. IJCV 75 (2007)
13. Barinova, O., Konushin, V., Yakubenko, A., Lee, K., Lim, H., Konushin, A.: Fast automatic single-view 3-d reconstruction of urban scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 100–113. Springer, Heidelberg (2008)
14. Gupta, A., Efros, A.A., Hebert, M.: Blocks world revisited: Image understanding using qualitative geometry and mechanics. In: ECCV (2010)
15. Hoiem, D., Efros, A.A., Hebert, M.: Closing the loop on scene interpretation. In: Proc. CVPR (2008)

16. Gould, S., Fulton, R., Koller, D.: Decomposing a scene into geometric and semantically consistent regions. In: ICCV (2009)
17. Sudderth, E., Torralba, A., Freeman, W.T., Wilsky, A.: Depth from familiar objects: A hierarchical model for 3D scenes. In: Proc. CVPR (2006)
18. Heitz, G., Gould, S., Saxena, A., Koller, D.: Cascaded classification models: Combining models for holistic scene understanding. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) NIPS, pp. 641–648. MIT Press, Cambridge (2008)
19. Yu, S., Zhang, H., Malik, J.: Inferring spatial layout from a single image via depth-ordered grouping. In: CVPR Workshop (2008)
20. Tu, Z., Chen, X., Yuille, A.L., Zhu, S.C.: Image parsing: Unifying segmentation, detection, and recognition. IJCV 63, 113–140 (2005)
21. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. In: CVPR (2006)
22. Leibe, B., Schindler, K., Cornelis, N., van Gool, L.: Coupled object detection and tracking from static cameras and moving vehicles. PAMI 30, 1683–1698 (2008)
23. Rother, C.: A new approach to vanishing point detection in architectural environments. IVC 20 (2002)
24. Russell, B., Torralba, A., Murphy, K., Freeman, W.: Labelme: A database and web-based tool for image annotation. IJCV 77 (2008)