Human-Centric Indoor Environment Modeling from Depth Videos

Jiwen Lu[1] and Gang Wang[1,2]

[1] Advanced Digital Sciences Center, Singapore
[2] Nanyang Technological University, Singapore

Abstract. We propose an approach to model indoor environments from depth videos (the camera is stationary when recording the videos), which includes extracting the 3-D spatial layout of the rooms and modeling objects as 3-D cuboids. Different from previous work which purely relies on image appearance, we argue that indoor environment modeling should be human-centric: not only because humans are an important part of the indoor environments, but also because the interaction between humans and environments can convey much useful information about the environments. In this paper, we develop an approach to extract physical constraints from human poses and motion to better recover the spatial layout and model objects inside. We observe that the cues provided by human-environment intersection are very powerful: we don't have a lot of training data but our method can still achieve promising performance. Our approach is built on depth videos, which makes it more user friendly.

Keywords: Scene understanding, environment modeling, human-centric, depth videos.

1 Introduction

In recent years, astonishing progress has been made in modeling indoor environments [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. However, none of these work considers humans living in the room. They estimate spatial layout of rooms purely from image appearance, which turns out to be a challenging task. Indoor environments are essentially built for humans, to afford humans' daily activities. The interaction between humans and environments can help tell us the essential information of the environments. As shown in Fig. 1, from a person who is standing, we can roughly know where the floor plane is; from a person who is sitting, we can roughly know where the chair is. Based on these observations, we argue in this paper that modeling indoor environments should fully exploit the cues provided by human-environment interaction. And human-environment interaction is a long-term process because people interact with the rooms all the time. Suppose we have a sensor installed in the room, then the life-long process of human-environment interaction can provide us a huge amount of information to help accurately model the environments. Our current experimental results are based on videos which only last for several minutes. We believe the results will

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 42–51, 2012.

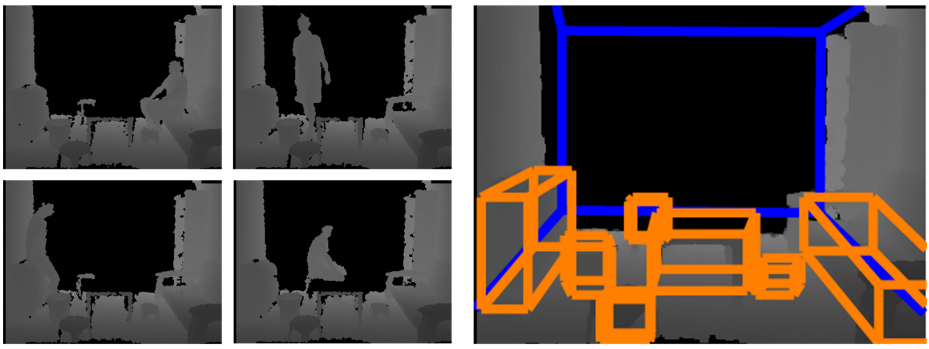

Fig. 1. Our approach extracts the spatial layout of a cluttered room and represents objects as 3-D cuboids in the environment from a depth video by exploiting information from human-environment interaction. (a) Four frames of an indoor environment depth video (the camera is stationary). (b) The modeling results.

be much more impressive if we record videos for several days or weeks, though which is beyond the scope of a "proof of concept" research paper.

Our work is based on depth videos recorded by the Microsoft Kinect sensor, different from previous work which is based on RGB images [1], [2], [3], [4], [5], [6], [7], [8]. There are two advantages of using the Kinect: 1) depth videos can better preserve the privacy information than RGB images/videos, which makes this technique easier to be accepted by the public; 2) depth images record the depth information, from which we can create a real 3-D model with more physical meanings. In this 3-D model, we can know how big an object is, how far between two objects, etc.

In this paper, we develop several methods to extract physical constraints from human poses and motion. For example, from a sitting pose, we estimate the support surface and know it must be from a sittable object. These constraints are effectively used in a statistical framework to help estimate the spatial layout of the environment and prune object hypotheses. We collect a number of depth videos with different types of human-environment interaction. The experimental results show that our method is very promising to solve this problem.

2 Overview of Approach

Given an indoor depth video (Fig. 2(a)), we estimate the spatial layout of the room and model objects insides as 3-D cuboids simultaneously. Our approach is illustrated in Fig. 2. Since the camera is stationary, the spatial layout of the room in each frame of the video is the same. Hence, we adopt the median filter [12] to recover the indoor scene background image (without humans inside) from the video (Fig. 2(b)). There are some missing values in the original depth images, we fill up each image by using a recursive median filter [13], [14]. We perform

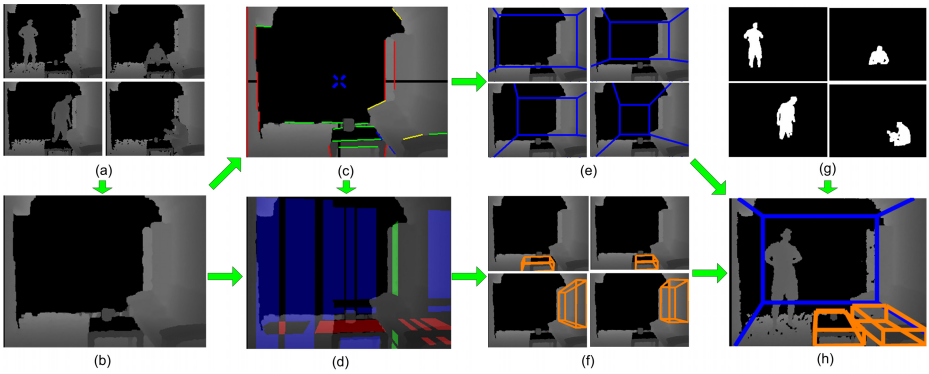

Fig. 2. Overview of our approach for indoor environment modeling. (a) Several frames in the depth video. (b) Indoor scene background image. (c) The detected long line segments and vanishing points of the scene background image. (d) Orientation map of the scene background image. (e) Room hypotheses generated from the vanishing points. (f) Object hypotheses generated from the orientation map. (g) Extracted human silhouettes for each frame of the depth video. (h) Spatial layout and objects modeled using the 3-D interactions between the room, objects, and humans.

line detection and vanishing points estimation on the scene background image (Fig. 2(c)). The vanishing points define the orientations of the major surfaces of the scene and provide constraints on its layout. Using these line segments and the vanishing points, we generate multiple room hypotheses (Fig. 2(e)) [15]. Moreover, we obtain the orientation map (Fig. 2(d)) from these line segments and vanishing points, and generate a number of object hypotheses (Fig. 2(f)). On the other hand, we subtract the scene background image from each frame of the depth video to obtain human silhouette (Fig. 2(g)). Then, we obtain human pose information in each frame and motion information over the whole video. Pose information is used to discover support surfaces, and motion information is used to discover objects. Finally, we test human-room-object compatibility and pick up the best compatible scene configuration over the whole video (Fig. 2(h)).

3 Generating Room and Object Hypotheses

A room hypothesis reflects the positions and orientations of walls, floor and ceiling. In this paper, we adopt the idea of [4] to represent the spatial layout of a room by a parametric model. Given a depth video, we first extract a scene background image by removing humans to generate room and object hypotheses. Fig. 2(b) shows the extracted indoor scene background image of the depth video. Following [4], we detect long line segments in the scene background image and find three dominant groups of lines corresponding to three vanishing points. By sampling pairs of rays from two of these vanishing points, we generate room hypotheses, and several of them are shown in Fig. 2(e).

One advantage of using depth images is that these room hypotheses can represent the 3-D geometry information, and each surface can be characterized by a plane in the 3-D coordinate system. Similar to [3], we represent objects as 3-D cuboids. We adopt the object hypothesis generation method in [3] by testing each pair of regions in the orientation map to check whether they can form convex edges. Fig. 2(f) shows four object hypotheses generated from the orientation map. Since there are a big number of hypotheses, we resort to human-environment interaction to select the best compatible scene configuration.

4 Human-Environment Interaction for Environment Modeling

We analyze human poses and motion to extract physical constraints to model the environment. While joint points of humans can be obtained by the SDK of Kinect, we find that the generated skeletons are still not stable to describe human poses for our environment modeling task because of mis-matching, scale change, and erroneous joint estimation. Hence, we adopt human silhouettes to extract human poses and motion information for indoor environment modeling.

4.1 Support Surface Estimation from Poses

Human poses are constrained by the environment. Humans have to be supported by stable surfaces. Hence, we can estimate support surfaces from human silhouettes. Having obtained the scene background image, we subtract it from each frame of the depth video to obtain human silhouette [12]. We apply two erosion and one dilation operations with a 3×3 template for each segmented human silhouette to remove noises and obtain one connected region [12]. Fig. 2(g) shows four segmented human silhouettes.

For each binary human silhouette image, we detect horizontal lines in the silhouette image and compute the length of each horizontal line. If the length is larger than a threshold L, we assume there is a support surface. In our experiment, L is set as 10. Given a horizontal edge line extracted from the silhouette image, suppose there are N_l pixels in the line and the 3-D coordinate of the ith point is $(X(u_i, v_i), Y(u_i, v_i), Z(u_i, v_i))$, where $1 \leq i \leq N_l$. We assume that the support surface P is parallel to the floor plane, then the plane equation of the support surface P can be written as $X = c$, where c is the parameter. We obtain c by solving the following optimization problem:

$$\min_{c} \sum_{i=1}^{N_l} \|X(u_i, v_i) - c\|_2^2 \tag{1}$$

Since there are some erroneous human silhouettes due to imperfect image segmentation, which may produce false positive surfaces, we only consider surfaces which are consistent in a sequence with at least N_T frames. In our experiments, N_T is set to 50.

Having estimated the support surfaces, we use them to generate constraints for indoor environment modeling. There are two types of surfaces that can be estimated: floor surfaces and object surfaces. For example, if the person stands or walks on the floor without occlusions, the estimated support surface should be the floor plane; if the person sits on a chair, the estimated surface should be the top surface of the chair cuboid. We differentiate different types of surfaces according to the aspect ration of the the human silhouette. Assume P_d be the floor surface and P_e be the eth object surfaces discovered from human-environment interaction. We exploit human-object and human-room interaction for environment modeling, as the constraints used in Eqs (8) and (9) in Section 5.

4.2 Object Discovery from Motion

Human motion also provides useful information for indoor environment modeling. For example, if human silhouettes are occluded by objects, we can localize these objects from the occluded silhouettes. Hence, we analyze human motion for object discovery.

We calculate the integral projections [16] of human binary silhouettes. For the tth human silhouette $I_t(u, v)$, we compute the horizontal integral projection (HIP_t) as follows:

$$HIP_t(u) = \sum_{v=1}^{wid} I_t(u, v) \tag{2}$$

Having calculated the HIP_t for the tth human silhouette $I_t(u, v)$, we detect the lowest row position u_t^{low} where $HIP_t(u_t) \geq 1$. The position u_t^{low} reflects the lowest position of the tth human silhouette $I_t(u, v)$. Now, we compute the absolute difference $D_t = |u_t - u_{t+1}|$ of u_t and u_{t+1} for two sequential frames, and determine the occlusion condition as follows:

1. If $D_t < \tau$, there is no occlusion change in the tth and $(t + 1)$th frames. In our experiments, τ is set to 10.
2. If $D_t \geq \tau$, there is an occlusion change in the tth and $(t + 1)$th frames. Specifically, if $u_t < u_{t+1}$, the person is occluded by some objects in the $(t + 1)$th frame and not occluded by the tth frame; otherwise, he/she is occluded by some objects in the tth frame and not occluded by the $(t+1)$th frame.

Having determined the image frames with occlusions, we detect an object surface in the environment. Assume the object rests on the floor, we can estimate the top surface of the object. This cue is used as a constraint in Eq. (9) in Section 5.

4.3 Human-Object and Human-Room Volumetric Reasoning

Humans must be compatible with objects and environment in terms of volume. Similar to [3], we perform human-object and human-room volumetric reasoning.

Fig. 3. An example to show how human-object and human-room volumetric reasoning can remove the incompatible object and room hypotheses. (a) The human cuboid obtained from human silhouette. (b) An incompatible object hypothesis that intersects with the human cuboid. (c) An incompatible room hypothesis which doesn't fully contain the human cuboid.

After obtaining the 3-D coordinate for each pixel of a human silhouette, we model the human as a 3-D cuboid. Ideally, the volumetric intersection between any object and the human cuboid should be empty. And the human cuboid should be fully contained in the free space defined by the walls of the room. With these two constraints, we can remove objects and room hypotheses which are incompatible with humans, as shown in Fig. 3.

5 Evaluating Environment Configurations

Given a depth video, we generate a set of room hypotheses $\{R_1, R_2, \cdots, R_N\}$, a set of 3-D object cuboid hypotheses $\{O_1, O_2, \cdots, O_K\}$, $R_n = \{F_1^n, F_2^n, \cdots, F_5^n\}$ defines the five faces [4], corresponding to the ceiling, left wall, middle wall, right wall, and floor of the nth room hypothesis, respectively. We also obtain the human silhouette for each frame. Our objective now is to find the best environment configuration, under which the spatial layout, objects, and humans are most compatible with each other. We find the best room hypothesis and object configurations which are most compatible with humans in the environment based on the following objective function:

$$\min_{R_n, Y} \sum_{i=1}^{5} \sum_{j=1}^{N_i} d(Q_{ij}^n, F_i^n) + \alpha \sum_{i=1}^{5} S(F_i^n, M) - \beta \sum_{k=1}^{K} Y_{O_k} \tag{3}$$

$$s.t.\ \ C_1(O_k, O_l) \leq \delta_1, \forall k, l, and\ k \neq l \tag{4}$$

$$C_2(O_k, F_i^n) \leq \delta_2, \forall k, i \tag{5}$$

$$C_3(H_s, O_k) \leq \delta_3, \forall s, k \tag{6}$$

$$C_4(H_s, F_i^n) \leq \delta_4, \forall s, i \tag{7}$$

$$C_5(P_d, F_5^n) \leq \delta_5, \forall d \tag{8}$$

$$Y_{O_k} = 1, if\ C_6(P_e, O_k^T) \leq \delta_6, \forall e, k \tag{9}$$

where N_i denotes the number of pixels in the ith plane and it is zero if the ith face is missing in the the spatial layout of the room, $d(Q_{ij}^n, F_i^n)$ is the distance between the jth pixel and the ith plane of the nth room hypothesis, $S(F_i^n, M)$ denotes the inconsistency between the ith plane of the nth room hypothesis and the orientation map, α and β are two parameters to balance the scales of these two terms, and they were empirically set to 1000 and 1000 in our experiments. $C_1(O_k, O_l)$ denotes the volumetric intersection between the kth and the lth objects, $C_1(O_k, O_l)$ must be small (ideally, 0) if they are both present in the scene configuration. $C_2(O_k, F_i^n)$ denotes the volume of O_k (or a part of O_k) which is not contained by the plane F_i^n. Again, $C_2(O_k, F_i^n)$ must be small as an object should be contained by the room. We keep all the object hypotheses which don't violate volumetric constraints by maximizing $\beta \sum_{k=1}^{K} Y_{O_k}$. $C_3(H_s, O_k)$ denotes the volumetric intersection between the human cuboid estimated from the sth frame and the kth object hypothesis, $C_3(H_s, O_k)$ must be small if O_k is present. $C_4(H_s, F_i^n)$ denotes the volume of the human cuboid (or a part of the cuboid) which is not contained by the plane F_i^n. Similarly, $C_4(H_s, F_i^n)$ must be small for a valid room hypothesis R_n. $C_5(P_d, F_5^n)$ denotes the distance between the floor discovered from humans and the floor plane of the nth room hypothesis. This distance must be small for a good room hypothesis. $C_6(P_e, O_k^T)$ denotes the distance between the eth object surface discovered from humans and the top surface of the kth object hypothesis, and O_k should be present in the scene configuration if $C_6(P_e, O_k^T)$ is small enough. δ_1, δ_2, δ_3, δ_4, δ_5 and δ_6 are six parameters and empirically set to 300, 1000, 300, 1000, 5, and 2, respectively.

We first obtain the 3-D coordinate for each pixel and obtain the plane equations of F_i^n. Then, the distance $d(Q_{ij}^n, F_{ti})$ can be computed. On the other hand, we penalize inconsistency between the orientation map and room surfaces. A good room hypothesis's floor and ceiling should be parallel to the YOZ plane of the orientation map, left and right walls should be parallel to the XOZ plane of the orientation map, and the middle wall should be parallel to the XOY plane of the orientation map, respectively. We find our approach is insensitive to these parameters because there is usually a big difference between instances which obey these constraints and instances which violate these constraints. These parameters are not tuned on our test videos.

It is intractable to obtain a closed-form solution to the constrained optimization problem in Eqs. (3)-(9). To address this, we resort to using a fast greedy search method to obtain an approximate solution. Specifically, we first select the top C room hypotheses $R_{top} = \{R_1, R_2, \cdots, R_C\}$ with the smallest objective function values in the first two terms of Eq. (3). In our experiments, the parameter C is empirically set as 10. Then, we remove the false object hypotheses according to the constraints in Eqs. (4) and (5), and keep the positive object hypotheses according to the constraint in Eq. (9), respectively. For the remaining room and object hypotheses, we compute the number of frames which satisfy the constraints in Eqs. (6-(8) over the whole video and select the one which has the largest number of frames satisfying these constraints as the final room hypothesis. For a depth video with 300 frames, our algorithm can find the best

Table 1. Statistics of our dataset. NF: number of frames; RH: number of room hypotheses; OH: number of object hypotheses; NO: number of objects interacted with humans in the room. Note that only the objects with human-object interactions in the environments were considered in our experiments.

Dataset	NF	RH	OH	NO	Dataset	NF	RH	OH	NO	Dataset	NF	RH	OH	NO
Video 1	348	324	10	2	Video 2	423	324	10	2	Video 3	294	360	12	2
Video 4	331	360	66	2	Video 5	345	144	17	1	Video 6	344	144	17	1
Video 7	306	196	5	1	Video 8	264	324	32	1	Video 9	305	324	32	2
Video 10	344	252	19	2	Video 11	370	324	48	2	Video 12	511	324	16	2
Video 13	371	324	4	2	Video 14	277	252	28	1	Video 15	297	252	24	1

scene configuration in less than 3 minutes with unoptimized matlab code using an Intel Core 2.80GHz CPU.

6 Experimental Results

6.1 Dataset

We collect a new dataset for our experiments. We cannot collect a large-scale dataset because it is very hard to get permissions from many room owners. In our study, we test on 15 depth videos with resolution of 320×240. These videos are collected in 5 different rooms, and the camera is stationary. In each room, we capture 2-6 video clips by changing the viewpoint of the depth camera. In each video, there is one person moving (e.g., standing, sitting, walking) freely in the room. We have two different persons for these 15 videos. Table 1 tabulates some statistics of our dataset including the number of frames for each video, the number of room hypotheses, the number of object hypotheses generated in Section 2, and the number of objects. Some example frames are shown in Fig. 2(a).

6.2 Results

To show the advantage of using human-environment interaction, we consider two baselines: 1) Baseline 1: Environment modeling without human and object information, and 2) Baseline 2: Environment modeling using objects but without human information.

Qualitative Evaluation: Fig. 4 illustrates some qualitative results of Baseline 1, Baseline 2 and our approach of three rooms, which clearly shows the benefits of human-environment interaction: our approach can better extract the spatial layout of the rooms and model objects than the other two baselines.

Quantitative Evaluation: We also quantitatively evaluate the performance of our approach in estimating the spatial layout. For each depth video, we manually labeled the ground truth of the five planes (i.e., three walls, ceiling plane and

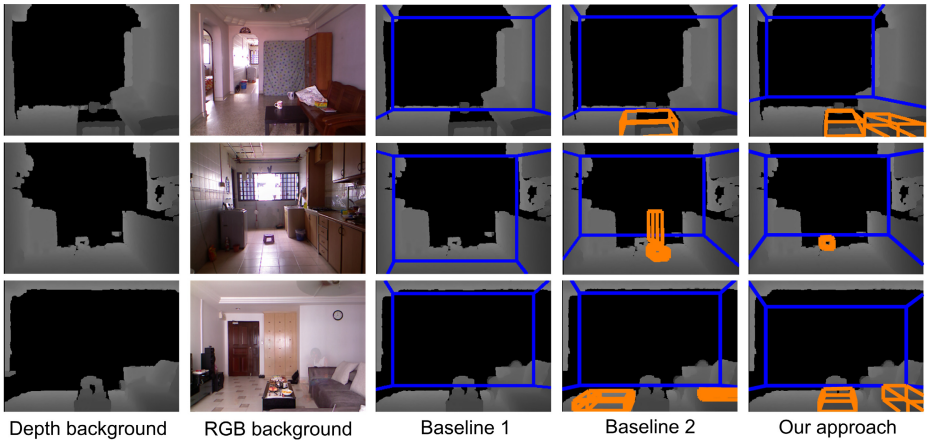

Fig. 4. Qualitative results for our indoor environment modeling (best viewed in color). From top to down are the 1st, 5th, and 10th videos in our dataset, respectively. Our approach can better find the floors of the rooms (rows 1-2) and model objects in the environments (rows 1-3) than the other two baselines.

Table 2. Spatial layout estimation error (%) of different methods

Dataset	Video 1	Video 2	Video 3	Video 4	Video 5	Video 6	Video 7	Video 8
Baseline 1	37.14	37.14	29.93	36.06	39.39	39.01	38.18	34.27
Baseline 2	37.14	37.14	29.93	34.24	24.33	28.42	38.18	24.26
Our method	18.96	22.11	16.91	19.36	24.33	27.34	15.26	22.12
Dataset	Video 9	Video 10	Video 11	Video 12	Video 13	Video 14	Video 15	Average
Baseline 1	36.32	31.21	37.30	22.42	31.38	29.66	29.57	33.93
Baseline 2	25.65	31.21	28.46	22.42	24.56	29.66	29.57	29.68
Our method	22.12	21.44	22.64	19.15	22.32	21.61	24.36	21.34

floor plane) in the scene background image. We use the pixel-based measure introduced in [4] which counts the percentage of pixels on the room surfaces that disagree with the ground truth. Table 2 records the quantitative results on all these depth videos when different methods were applied. Our approach achieves an average error rate of 21.34%, while the performance number of Baseline 1 and Baseline 2 are 33.94% and 29.68%, respectively. The methods in [4] and [3] cannot be directly applied to our data because it is very hard for us to collect enough data to learn a structural model for scene configuration evaluations.

7 Conclusion and Future Work

In this paper, we have demonstrated the promise of exploiting human-environment interaction to model indoor environments. As a proof of concept

research paper, our experiments are performed on short video clips. However, we believe much better results will be observed if we have videos for days or weeks, as more human-environment interaction information will be provided. In the near future, we are interested in applying this technique in realistic scenarios, such as hospital wards, nuring homes, etc.

Acknowledgment. This study is supported by the research grant for the Human Sixth Sense Program at the Advanced Digital Sciences Center from the Agency for Science, Technology and Research (A*STAR) of Singapore.

References

1. Hoiem, D., Efros, A., Hebert, M.: Recovering surface layout from an image. IJCV 75, 151–172 (2007)
2. Lee, D.C., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: CVPR, pp. 2136–2143 (2009)
3. Lee, D.C., Gupta, A., Hebert, M., Kanade, T.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: NIPS, pp. 1288–1296 (2010)
4. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV, pp. 1849–1856 (2009)
5. Wang, H., Gould, S., Koller, D.: Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 497–510. Springer, Heidelberg (2010)
6. Li, L., Socher, R., Fei-Fei, L.: Towards total scene understanding: classification, annotation and segmentation in an automatic framework. In: CVPR, pp. 2036–2043 (2009)
7. Gupta, A., Satkin, S., Efros, A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR, pp. 1961–1968 (2011)
8. Hedau, V., Hoiem, D., Forsyth, D.: Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 224–237. Springer, Heidelberg (2010)
9. Tsai, G., Xu, C., Liu, J., Kuipers, B.: Real-time indoor scene understanding using bayesian filtering with motion cues. In: ICCV (2011)
10. Delage, E., Lee, H., Ng, A.: A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In: CVPR, pp. 2418–2428 (2006)
11. Yu, S.X., Zhang, H., Malik, J.: Inferring spatial layout from a single image via depth-ordered grouping. In: CVPRW, pp. 1–7 (2008)
12. Lu, J., Zhang, E.: Gait recognition for human identification based on ica and fuzzy svm through multiple views fusion. Pattern Recognition Letters 28, 2401–2411 (2007)
13. Herbst, E., Ren, X., Fox, D.: Rgb-d object discovery via multi-scene analysis. In: IROS, pp. 4850–4856 (2011)
14. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: ICRA, pp. 1817–1824 (2011)
15. Rother, C.: A new approach to vanishing point detection in architectural environments. Image and Vision Computing 20, 647–655 (2002)
16. Zhou, Z., Geng, X.: Projection functions for eye detection. Pattern Recognition 37, 1049–1056 (2004)