3D Depth Perception from Single Monocular Images

Hang Xu, Kan Li, FuYu Lv, and JianMeng Pei

Beijing institute of technology, Beijing, China
{xuhang,Likan}@bit.edu.cn
lvfuyu91@sina.com, bitpeijianmeng@gmail.com

Abstract. Depth perception from single monocular images is a challenging problem in computer vision. Since the single image is lack of features of context, we only find all the cues from the local image. This paper presents a novel method for 3D depth perception from a single monocular image containing the ground to estimate the absolute depthmaps more accurately. Different from previous methods, in our method, we first generates the ground plane depth coordinate system from a single monocular image by image-forming principle, and then locates the objects in image with the coordinate system using the geometric characteristics. At last, we provide an method to estimate the accurate depthmaps. The experiments show that our method outperforms the state-of-the-art single-image depth perception methods both in relative depth perception and absolute depth perception.

1 Introduction

An increasing attention in computer vision field is paid to RGB-D image, which represents RGB images with depth information [6,7]. The depth information accommodates images more suitable for various computer vision tasks [8]. Before RGB-D images came into wide spread use, there are large numbers of images without depth information on the Internet and on some open datasets. We try to estimate the depth information from single monocular images in order to solve more computer vision tasks. But estimating the depth information from single monocular images is still a challenging task. As we know, humans appear to be extremely good at judging depth from single monocular images because of their common sense and life experience. However, it is impossible for a computer vision system to learn all these kinds of prior knowledge. Besides, for the single images, extracting the features can rely on nothing but the image itself. This paper focuses on the images containing the ground, since the ground plane will be a superior reference frame of the depth, and most scenes in images include the ground plane.

In the last decade, some methods have been proposed to estimate depths from monocular images, but almost all the methods based on Markov Random Field (MRF) are sensitive to multicolored objects in an image, and need many learning parameters, which are difficult to be learned. Some methods using geometric

X. He et al. (Eds.): MMM 2015, Part I, LNCS 8935, pp. 510–521, 2015.

characteristics only estimate the relative depth of objects or scene in an image, but can't obtain their absolute depth.

In order to resolve the problems above, we propose a novel approach based on the image-forming principle and consider the geometric characteristics. Our method need less parameters, while it can obtain the absolute depth information and reduce the effects of multicolored objects in images. Our method is based on the assumption that no objects are hung in the air and one object's depth can be determined by the points it touches the ground plane or other objects. We also assume that the optical axis of the image-forming device is always parallel to the ground plane.

2 Related Work

Depth estimating from single monocular images is the object of current attention in the literature. One of the popular methods is based on MRF proposed by Saxena's group. Saxena presented an improvement MRF algorithm to capture depths and relationships between depths at different points in images [19,20]. Schulte and his partners used an image set collected by a 3D distance scanner to train MRF and predicted depth [21]. Sun and Ng improved the MRF method and fixed the parameters to make it more suitable for 3D reconstruction [22]. Saxena *et.al.* showed that the MRF method can do well on images captured by robotic camera. They used the method to achieve robotic grasping [18]. Betra's group used a new learning model to develop the MRF method to estimate depthmaps [1]. In addition, Kratz and his partners used Expectation-maximization algorithm based on MRF to calculate the depth-energy for defogging the images, while the algorithm can also get the depth information [13]. However, MRF-based methods ignore the geometry of the object which are easily affected by the color changing of multicolored objects, resulting in a large depth error. Another disadvantage of MRF-based methods is difficult to confirm their parameters in learning phase.

Methods using the geometric characteristics of objects only can obtain the relative depth from monocular images. Hoiem and Efros made 3D reconstruction from monocular images, but they focused on generating 3D graphical images rather than calculating the absolute depth [12]. Hebert's group took a method to construct the surface layout of objects, which helps predict the depth relationship of objects [10]. Hedau *et.al.* developed an algorithm to recover the free space of indoor scenes using the relative depth information by exploiting the box like geometric structure in images [5]. However, almost all approaches using geometric characteristics cannot get the accurate depthmaps.

Other main depth perception methods from monocular images also have their limitations. Lin *et.al.* provided a method based on Shape From Defocus (SFD) [14,16] to estimate depth from single defocused image [15], but it is unable to distinguish the difference between blurred textures and defocused sharp textures. Methods based on T-junction cues can only obtain the relative depth of objects, not the absolute depth [2]. In addition, Nicolas *et.al.* presented an

unsupervised method to approximate basic scene geometry properties [17]. Their method can obtain the absolute depth of objects, but it has a high requirement for imaging device. Cherian's work is very similar to ours, but it is not accurate enough especially on vertical regions [3].

In this paper, we present a novel 3D depth perception method from single monocular images. Our contributions are: 1. Derive the relationship of points between the 3D real world and the 2D image by image-forming principle and propose a method to generate the depth coordinate system on images. 2. Propose the image depth perception method to obtain the accurate depthmaps of images. 3. Present a measure of relative depth perception performance. 4. Verify the performance of our method in both relative depth accuracy and absolute depth accuracy.

3 Ground Plane Depth Coordinate System

As we know, the relationship of the 3D real world and 2D digital image can be described by a coordinate system. In our approach, we first analyze the image-forming principle and then propose a method to generate the ground plane depth coordinate system on images.

3.1 Image-Forming Principle

In Fig. 1, point O is the position of image-forming device lens. Now we use depth coordinate system $X_C Y_C Z_C$ to describe the real world space. In this coordinate system, a point $P(X, Y, Z)$ in real world is transformed to a 2D point $p(x, y)$ on the real imaging plane. In image-forming principle, device can receive light from outside and form a upside down image on the real imaging plane behind lens with a distance F, which is called focal distance. In order to help understanding, we add a virtual imaging plane in front of lens with a distance F. An erected image will be formed on the virtual imaging plane. Obviously, it is impossible to calculate the precise depth value between the two different coordinate systems, but we can get their relationship which is helpful to our work.

In the condition of Fig. 1, we can easily prove that $\triangle OO'p$ is similar to $\triangle OO''P$ and $\triangle O'pq$ is similar to $\triangle O''PQ$. Then we get the relationship in Formula (1).

$$\frac{OO'}{OO''} = \frac{O'p}{O''P}, \frac{O'p}{O''p} = \frac{pq}{PQ} = \frac{O'q}{O''Q}$$

$$\frac{F}{Z} = \frac{y}{Y} \tag{1}$$

Formula (1) shows the relationship between the point 2D coordinates and its 3D coordinates, which helps us figure out how the depth changing in the real world react on the image.

We assume that there are two points $P_0(X, Y, Z_0)$ and $P_1(X, Y, Z_1)$ in 3D real world coordinate system, and the distance between the two points on axis Z_C is

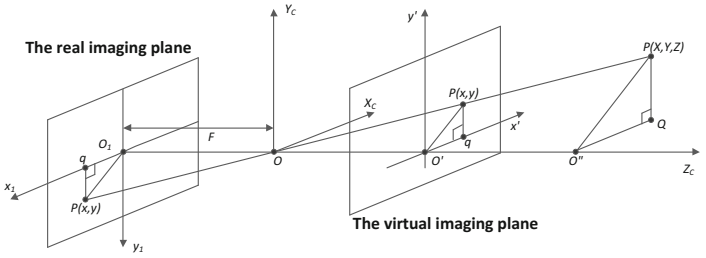

Fig. 1. Optical image-forming principle

d. In Fig. 2, point P_0 is mapped to $p_0(x_0, y_0)$ on the imaging plane, and point P_1 is mapped to $p_1(x_1, y_1)$. Based on Formula (1), we can obtain the Formula (2) and the relationship between the two points' 2D coordinates and their depth in real world as Formula (3).

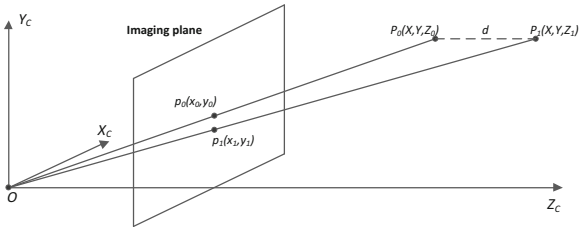

Fig. 2. The relationship between 3D and 2D coordinates

$$\frac{F}{Z_0} = \frac{y_0}{Y}, \frac{F}{Z_1} = \frac{y_1}{Y} \tag{2}$$

$$\frac{y_1}{y_0} = \frac{Z_0}{Z_1} \tag{3}$$

3.2 Depth Coordinate System Generating Algorithm

In this subsection, we will propose a method to generate the depth coordinate system. In the Formula (3), the relationship between the two points' 2D coordinates and their depth in real world is provided. From this relationship, we can get an inference that all points on the ground plane in the image have the similar relationship. So we can use the relationship to generate the depth coordinate system of the ground plane in the image.

We suppose there is a set P, which contains n points in 3D real world. These points always have the same coordinates on axis X_C and Y_C, and every two neighboring points $P_i(X, Y, Z_i)$ and $P_{i+1}(X, Y, Z_{i+1})$ have the same distance d (actually, these points belong to the same line, and the line is on the ground

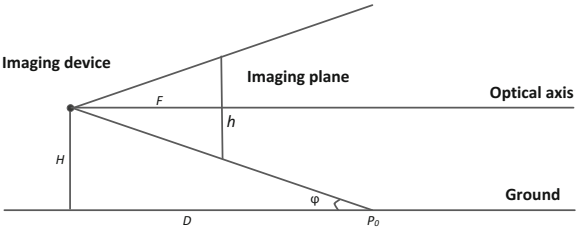

Fig. 3. The features of the device

plane and parallel to the optical axis of the device). We can represent the depth coordinate system on the image using a set C.

$$C = \{y_i | P_i \in P, i = 0, 1, \cdots, n\} \tag{4}$$

From the previous conclusion in Formula (3) and Formula (4), it seems not difficult to propose the method of generating the depth coordinate system. But another problem is here that how to determine the depth of the first point $P_0 (P_0 \in P)$.

In Fig. 3, the vertical distance from the imaging device to the ground is H. The height of the imaging plane is h, and the distance from imaging device to the imaging plane is focal distance F. The angle between visual range and optic axis is represented as φ. We assume that the depth of the bottom most row of the image on the ground plane from the imaging device is D. So we can obtain the expression of D as Formula (5).

$$D = \frac{H}{\tan \varphi} = \frac{2FH}{h} \tag{5}$$

From Formula (5), D is only affected by the focal distance F, the altitude H of the device and the altitude h of the imaging plane. When processing robotic vision images or images with the device parameters, we can easily get these statuses to calculate the value of D. We can also use some experienced values to determine the value of D if there is no information about device status. The value of $D \in [5, 10](meters)$ can be used in most images on the Internet or in the open datasets. Then we provide a recurrence formula in general case as follows.

$$y_{n+1} = (\frac{D + nd}{D + (n + 1)d}) y_n \tag{6}$$

Then we give the 3D coordinate generating algorithm as follows.

Algorithm 1. Depth coordinate system generating algorithm.

Input:	The feature parameters of device h, F; The state parameters of device H; The unit depth d and the point set P.
Output:	The coordinate set C.
Step 1:	Calculate the horizontal distance D by Formula (5).
Step 2:	Set $C = \emptyset$, $i = 0$, $y_0 = -\frac{ROW}{2}$, put y_0 into set C. (ROW is the total rows of the processed image.)
Step 3:	Use y_i to calculate the coordinate of the next point $P_{i+1} \in P$ with further unit depth, y_{i+1}, by Formula (6), then put y_{i+1} into set C
Step 4:	If we have the next point $P_{i+2} \in P$, set $i = i + 1$ and go step 3; otherwise, go step 5.
Step 5:	Return the coordinate set C.

By Algorithm 1, we obtain the coordinate set C, which contains all 2D coordinates of 3D points on ground plane with different depth. We can get different accurate-scale coordinates by adjusting the unit depth d. In Fig. 4, we set the unit depth $d = 2(m)$ of the coordinate, and set the middle one as $d = 1(m)$, the right one as $d = 0.5(m)$. The performance shows that the smaller the d is, the more accurate the coordinate is. But setting the value of d too small is unnecessary and inefficient. Based on a large number of experiments, we find that $d = 1(m)$ can obtain a good performance and keep a high efficiency.

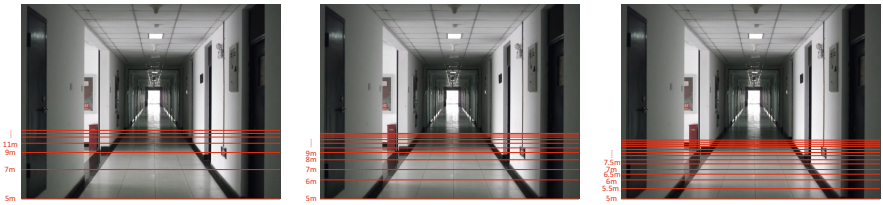

Fig. 4. Coordinates with different unit depth

4 Image Depth Perception Algorithm

In order to determine the depth of every point in an image, we should figure out the different areas in an image firstly. We use a texture segmentation algorithm based on [9,4,11], which is using the geometric characteristics. Then we propose a depth perception algorithm to obtain the depthmap of the image based on segmentation preprocessing result.

4.1 Texture Segmentation

Since the depth coordinate system is used for mapping the ground plane to different depth, the images should be preprocessed to obtain the ground area

firstly. Based on the geometric characteristics, the objects in images are categorized precisely [9], where the sky, the ground and vertical in images are divided from each other. In Fig. 5, image A is the original image, image B is the result of categorized. In image B, the blue part is the sky area, the green one is the the ground area, and the red one is vertical area. The arrows "↑", "←" or "→" on the objects show the depth changing direction. The marks "◯" or "×" on the objects represent that those objects are solid or hollow respectively.

After the sky, ground and vertical areas have been separated, we divide the vertical area more precisely [4,11] to get the location of the objects (see Fig. 5, image B and C).

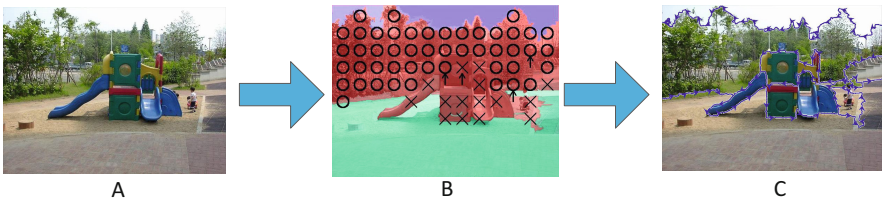

Fig. 5. Texture segmentation process. A: original image. B: the image is categorized into sky, vertical and ground. C: the vertical is segmented more precisely.

4.2 Image Depth Perception

In this subsection, we propose an image depth perception method, which includes the ground plane depth perception and the vertical area depth perception. Obviously, we can set the depth of sky infinite.

After the preprocessing, we obtain a image in which every objects and areas are separated from each other. Any parts of the segmentation result can be described as a quad of $R_i(C_i, M_i, B_i, P_i)$. This quad provides the attributes and location information of a segmentation region. In this quad, C_i and M_i both represent the category of the region. C_i shows the region belonging to sky, ground plane or vertical, and M_i shows which kind of mark is on the region. B_i is the bottom line number of the region. P_i is the 2D coordinate of the top-left point of the region.

We suppose a set R that includes all the regions of the segmented image. The set R can be described as $R = \{R_i | i = 0, 1, \cdots, n\}$, where n is the total number of regions in a image.

By the assumption that no objects are hung in the air and one object's depth can be determined by the points it touches the ground plane, we can easily estimate the depth information of most regions. But there exist some regions covered by other regions or not touching the ground plane directly, so it is difficult to estimate their depth information by their bottom line. Since the work of [12] shows that neighboring regions can provide important cues to infer the depth information of one region. Inspired by [12], we propose a method to

estimate the region's depth information by its neighboring regions based on their touching length. Every neighbors having been determined their depth should make a contribution to deciding the depth of this region. We use the following equation to determine the depth of regions which are not touching the ground directly:

$$dep = \sum_{j=1}^{m} r_j dep_j \tag{7}$$

In the Formula (7), m is total number of neighboring regions having been determined the depth, r_j is the rate of two regions' touching length. The value of r_j is given by $r_j = \frac{TL_j}{\sum_{j=1}^{m} TL_j}$. TL_j in the equation means the touch length of two regions. Meanwhile, it must satisfy that $\sum_{j=1}^{m} r_j = 1$.

Now we give the image depth perception algorithm as follows.

Algorithm 2. Image depth perception Algorithm.

Input:	The coordinate set C; The region set R. ($R = \{R_i(C_i, M_i, B_i, P_i) \mid i = 1, 2, \cdots, n\}$)
Output:	The depthmap of the image.
Step 1:	Set a set $RF = \emptyset$; Initialize the depthmap empty.
Step 2:	Traverse R, find all $R_i \in R$, where $C_i = SKY$. Set depth of each R_i the max value, remove R_i from R and put it into RF.
Step 3:	Traverse R, find all $R_i \in R$, where $C_i = GROUND$. Set depth of each R_i the corresponding value based on the coordinate set C, remove R_i from R and put it into RF.
Step 4:	Scan the other elements in R and find $R_i \in R$ which touches the ground plane and has minimum B_i. If it succeeds, go step 5; otherwise, go step 6.
Step 5:	According to the M_i and the touching points of the R_i, set depth of the R_i, remove R_i from R and put it into RF. Go step 4.
Step 6:	Scan the other elements in R and find $R_i \in R$ which has minimum B_i. If it succeeds, go step 7; otherwise, go step 8.
Step 7:	Calculate the depth of the R_i base on Formula (7), set depth of the R_i, remove R_i from R and put it into RF. Go step 6.
Step 8:	Return the depthmap.

5 Experiments

5.1 Dataset

In this paper, we use an image set of 350 images/depthmaps provided by Stanford (`http://make3d.cs.cornell.edu/data.html`) to finish the experiments. The data is collected by a 3D distance scanner, and comprised of a large set of

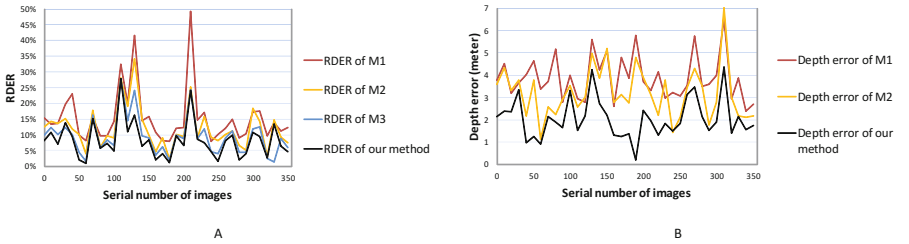

Fig. 6. Experimental results. A: The $RDER$ of different methods. B: The depth error of different methods.

original images and their corresponding ground-truth depthmaps. Each image is of size 1704×2272 and each depth map is of size 55×305.

In the experiments on depth perception, we provide some other researchers' method as contrast. The MRF method [1] is a state-of-the-art method in this filed, which is called M1. Cherian's work [3] is very similar to ours, which is called M2. Hoiem's method [12] has a good performance, which is called M3. But it cannot obtain the depthmaps, and we only use it for relative depth experiment.

5.2 Relative Depth Perception

Since many applications of computer vision only need relative depth of images to finish their tasks, the error rate of relative depth is also a valuable indicator of the depth perception. We firstly make the experiment on relative depth perception.

We propose the Relative Depth Error Rate ($RDER$), a measure of the relative depth perception performance. On the dataset, we already have the ground-truth depthmaps. Based on the ground-truth, we sort every pixel in the ascending order, and treat the original image's sequencing result as the standard depth order. By using depth perception method, we obtain the image's depth information and sort its pixels to get a new depth order. Then the inverted sequence number between the new depth order and the standard depth order shows the relative depth errors in the depth estimating result. As we know, every order has a maximal inverted sequence number, which shows the maximal value of relative depth errors in the image.

We define the $RDER$ as $RDER = \frac{ISN}{MISN}$, where the numerator is inverted sequence number (ISN) of the estimating result and the denominator is the maximal inverted sequence number ($MISN$) of the standard order.

Fig. 6 A shows that the $RDER$ of all images on the dataset, which is obtained by our method, M1, M2 and M3. The figure demonstrates that our method outperforms M1 and M2 on the relative depth perception, and it is about the same between our method and M3. Since our method and M3 take the object's geometric characteristics into consideration, which make sure the relative depth order between objects rarely change.

As the result, the average $RDER$ of our method is 8.173%, and the average $RDER$ of M1, M2 and M3 are 15.359%, 11.872% and 9.367% respectively.

5.3 Absolute Depth Perception

In this subsection we evaluate the absolute depth perception performance of our approach on the image dataset. We use our method, M1 and M2 to estimate the depthmap of each image. Then we compute the average difference of the ground-truth depthmap with the result depthmaps, and the difference shows the absolute depth estimating error. Apparently, the smaller depth error is, the more accurate the depth estimating method is.

Fig. 6 B shows that the depth error on all the images obtained by our method, M1 and M2. We observe in the figure that our method shows better performance on absolute depth perception than other methods. In our method, the geometric characteristics based texture segmentation make sure the atomicity of every objects, and avoid the influence of the color changing in an object perfectly. The results reflect this effort clearly.

The experiment results show that the average depth error of our method is 2.1359 meters, while the average depth error of M1 and M2 is 3.7843 meters and 3.1970. Our method has succeeded in improving the accuracy of the depth information perception.

5.4 Experimental Results

Fig. 7 shows a number of test results of experiments. The column A of each row is the original image, the column B is the relative depth obtained by M2, the column C is the absolute depth estimated by M1, and the column D is the absolute depth result of our method. Comparing with M2's approach, we can see that our approach outperforms in many detail areas, such as top of the tree and the forest behind the building. M2's approach is good at obtain the relative depth, but it cannot estimate the accurate depthmap. Comparing with the M1, our approach outperforms both in relative depth perception and absolute depth perception obviously. Since the M1 is sensitive with color changing, many depth estimating errors happen as shown in the figure.

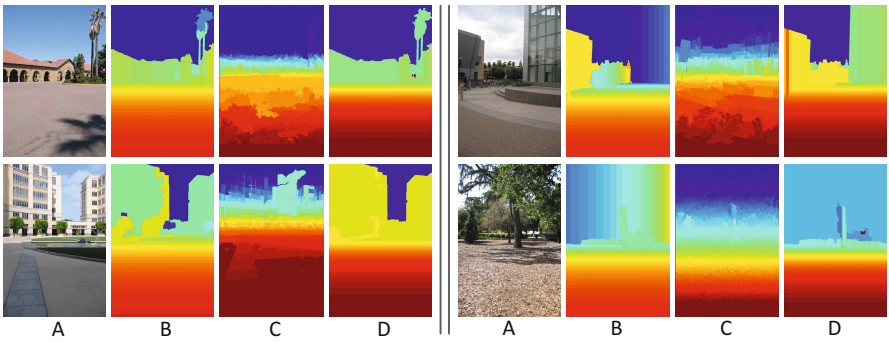

Fig. 7. A: original image. B: relative depth obtained by M2. C: absolute depth obtained by M1. D: absolute depth obtained by our method.

6 Conclusions

In this paper, we firstly analyze the image-forming principle and propose a method to generating the depth coordinate system in image. Then we use the geometric characteristics to avoid the influence of color changing in an object and propose a novel image depth perception method to obtain the accurate depthmaps of images. In experiment, a measure is provided to judge the relative depth perception performance. The experiments show that our method outperforms the-state-of-art methods both in relative depth perception and absolute depth perception.

In the future, we will consider the priori information of objects to help us compute the depth of the images especially in vertical area. Considering the angel changing of device is another field we can accommodate in the future. In addition, we will also use more diverse datasets to testify our method's robustness.

Acknowledgments. The Research was supported in part by Natural Science Foundation of China(No.60903071), Specialized Research Fund for the Doctoral Program of Higher Education of China, and Training Program of the Major Project of BIT.

References

1. Batra, D., Saxena, A.: Learning the right model: Efficient max-margin learning in laplacian crfs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2136–2143 (2012)
2. Calderero, F., Caselles, V.: Recovering relative depth from low-level features without explicit t-junction detection and interpretation. International Journal of Computer Vision 104(1), 38–68 (2013)
3. Cherian, A., Morellas, V., Papanikolopoulos, N.: Accurate 3d ground plane estimation from a single image. In: 2009 IEEE International Conference on Robotics and Automation (ICRA), pp. 2243–2249. IEEE (2009)
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
5. Hedau, V., Hoiem, D., Forsyth, D.: Recovering free space of indoor scenes from a single image. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2807–2814. IEEE (2012)
6. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. The International Journal of Robotics Research 31(5), 647–663 (2012)
7. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In: Experimental Robotics, pp. 477–491. Springer (2014)
8. Hoiem, D., Adviser-Efros, A.A., Adviser-Hebert, M.: Seeing the world behind the image: spatial layout for three-dimensional scene understanding. Carnegie Mellon University (2007)
9. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: 2005 10th IEEE International Conference on Computer Vision, pp. 654–661. IEEE (2005)

10. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision 75(1), 151–172 (2007)
11. Hoiem, D., Efros, A.A., Hebert, M.: Closing the loop in scene interpretation. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE (2008)
12. Hoiem, D., Efros, A.A., Hebert, M.: Recovering occlusion boundaries from an image. International Journal of Computer Vision 91(3), 328–346 (2011)
13. Kratz, L., Nishino, K.: Factorizing scene albedo and depth from a single foggy image. In: 2009 12th IEEE International Conference on Computer Vision, pp. 1701–1708. IEEE (2009)
14. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM Transactions on Graphics (TOG) 26(3), 70 (2007)
15. Lin, J., Ji, X., Xu, W., Dai, Q.: Absolute depth estimation from a single defocused image. IEEE Transactions on Image Processing: a publication of the IEEE Signal Processing Society 22(11), 4545 (2013)
16. Namboodiri, V.P., Chaudhuri, S.: Recovery of relative depth from a single observation using an uncalibrated (real-aperture) camera. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–6. IEEE (2008)
17. Nicolas, H.: Depth analysis for surveillance videos in the h. 264 compressed domain. In: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 146–149. IEEE (2012)
18. Saxena, A.: Monocular depth perception and robotic grasping of novel objects. Ph.D. thesis, Citeseer (2009)
19. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Neural Information Processing Systems Conference (NIPS), vol. 18, pp. 1–8 (2005)
20. Saxena, A., Chung, S.H., Ng, A.Y.: 3-d depth reconstruction from a single still image. International Journal of Computer Vision 76(1), 53–69 (2008)
21. Saxena, A., Schulte, J., Ng, A.Y.: Depth estimation using monocular and stereo cues. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 2197–2203. Morgan Kaufmann Publishers Inc. (2007)
22. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Depth perception from a single still image. In: AAAI Conference on Artificial Intelligence (AAAI), pp. 1571–1576 (2008)