3D Reconstruction of Indoor Scenes using a Single RGB-D Image

Panagiotis-Alexandros Bokaris[1], Damien Muselet[2] and Alain Trémeau[2]
[1] *LIMSI-CNRS, University of Paris-Saclay, Univ. Paris-Sud, 91405 Orsay Cedex, France*
[2] *Laboratoire Hubert Curien, CNRS, UMR 5516, Université Jean Monnet, 42000 Saint-Étienne, France*

Keywords: 3D Reconstruction, Cuboid Fitting, Kinect, RGB-D, RANSAC, Bounding Box, Point Cloud, Manhattan World.

Abstract: The three-dimensional reconstruction of a scene is essential for the interpretation of an environment. In this paper, a novel and robust method for the 3D reconstruction of an indoor scene using a single RGB-D image is proposed. First, the layout of the scene is identified and then, a new approach for isolating the objects in the scene is presented. Its fundamental idea is the segmentation of the whole image in planar surfaces and the merging of the ones that belong to the same object. Finally, a cuboid is fitted to each segmented object by a new RANSAC-based technique. The method is applied to various scenes and is able to provide a meaningful interpretation of these scenes even in cases with strong clutter and occlusion. In addition, a new ground truth dataset, on which the proposed method is further tested, was created. The results imply that the present work outperforms recent state-of-the-art approaches not only in accuracy but also in robustness and time complexity.

1 INTRODUCTION

3D reconstruction is an important task in computer vision since it provides a complete representation of a scene and can be useful in numerous applications (light estimation for white balance, augment synthetic objects in a real scene, design interiors, etc). Nowadays, with an easy and cheap access to RGB-D images, as a result of the commercial success of the Kinect sensor, there is an increasing demand in new methods that will benefit from such data.

A lot of attention has been drawn to 3D reconstruction using dense RGB-D data (Izadi et al., 2011; Neumann et al., 2011; Dou et al., 2013). Such data are obtained by multiple acquisitions of the considered 3D scene under different viewpoints. The main drawback of these approaches is that they require a registration step between the different views. In order to make the 3D reconstruction of a scene feasible despite the absence of a huge amount of data, this paper focuses on reconstructing a scene using a single RGB-D image. This challenging problem has been less addressed in the literature (Neverova et al., 2013). The lack of information about the shape and position of the different objects in the scene due to the single viewpoint and occlusions makes the task significantly more difficult. Therefore, various assumptions have to be made in order to make the 3D reconstruction feasible (object nature, orientation).

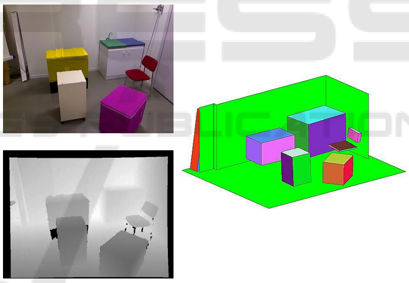

Figure 1: (left) Color and Depth input images, (right) 3D reconstruction of the scene.

In this paper, starting from a single RGB-D image, a fully automatic method for the 3D reconstruction of an indoor scene without constraining the object orientations is proposed. In the first step, the layout of the room is identified by solving the parsing problem of an indoor scene. For this purpose, the work of (Taylor and Cowley, 2012) is exploited and improved by better addressing the problem of the varying depth resolution of the Kinect sensor while fitting planes. Then, the objects of the scene are segmented by using a novel plane-merging approach and a cuboid is fitted to each of these objects. The reason behind the selection of such representation is that most of the objects in a common indoor scene, such as drawers, bookshelves,

Bokaris P., Muselet D. and TrÃĺmeau A.
3D Reconstruction of Indoor Scenes using a Single RGB-D Image.
DOI: 10.5220/0006107803940401
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 394-401
ISBN: 978-989-758-227-1

tables or beds have a cuboid shape. For the cuboid fitting step, a new “double RANSAC”-based (Fischler and Bolles, 1981) approach is proposed. The output of the algorithm is a 3D reconstruction of the observed scene, as illustrated in Fig. 1. In order to assess the quality of the reconstruction, a new dataset of captured 3D scenes is created, in which the exact positions of the objects are measured by using a telemeter. In fact, by knowing the exact 3D positions of the objects, one can objectively assess the accuracy of all the 3D reconstruction algorithms. This ground truth dataset will be publicly available for future comparisons. Finally, the proposed method is tested on this new dataset as well as on the NYU Kinect dataset (Silberman et al., 2012). The obtained results indicate that the proposed algorithm outperforms the state-of-the-art even in cases with strong occlusion and clutter.

2 RELATED WORK

The related research to the problem examined in this paper can be separated in two different categories. The first category is the extraction of the main layout of the scene while the second one is the 3D representation of the objects in the scene.

Various approaches have been followed in computer vision for recovering the spatial layout of a scene. Many of them are based on the Manhattan World assumption (Coughlan and Yuille, 1999). Some solutions only consider color images without exploiting depth information (Mirzaei and Roumeliotis, 2011; Bazin et al., 2012; Hedau et al., 2009; Schwing and Urtasun, 2012; Zhang et al., 2014) and hence provide only coarse 3D layouts. With Kinect, depth information is available, which can be significantly beneficial in such applications. (Zhang et al., 2013) expanded the work of (Schwing and Urtasun, 2012) and used the depth information in order to reduce the layout error and estimate the clutter in the scene. (Taylor and Cowley, 2011) developed a method that parses the scene in salient surfaces using a single RGB-D image. Moreover, (Taylor and Cowley, 2012) presented a method for parsing the Manhattan structure of an indoor scene. Nonetheless, these works are based on assumptions about the content of the scene (minimum size of a wall, minimum ceiling height, etc.). Moreover, in order to address the problem of the depth accuracy in Kinect, they used the depth disparity differences, which is not the best solution as it is discussed in section 3.1.

Apart from estimating the layout of an indoor scene, a considerable amount of research has been done in estimating surfaces and objects from RGB-D images. (Richtsfeld et al., 2012) used RANSAC and NURBS (Piegl, 1991) for detecting unknown 3D objects in a single RGB-D image, requiring learning data from the user. (Cupec et al., 2011; Jiang, 2014) segment convex 3D shapes but their grouping to complete objects remains an open issue. To the best of our knowledge, (Neverova et al., 2013) was the first method that proposed a 3D reconstruction starting from a single RGB-D image under the Manhattan World assumption. However, it has the significant limitation that it only reconstructs 3D objects which are parallel or perpendicular to the three main orientations of the Manhattan World. (Lin et al., 2013) presented a holistic approach that takes into account 2D segmentation, 3D geometry and contextual relations between scenes and objects in order to detect and classify objects in a single RGB-D image. Despite the promising nature of such approach it is constrained by the assumption that the objects are parallel to the floor. In addition, the cuboid fitting to the objects is performed as the minimal bounding cube of the 3D points, which is not the optimal solution when working with Kinect data, as discussed by (Jia et al., 2013). Recently, an interesting method that introduced the “Manhattan Voxel” was developed by (Ren and Sudderth, 2016). In their work the 3D layout of the room is estimated and detected objects are represented by 3D cuboids. Being a holistic approach that prunes candidates, there is no guarantee that a cuboid will be fitted to each object in the scene. Based on a single RGB image, (Dwibedi et al., 2016) developed a deep-learning method to extract all the cuboid-shaped objects in the scene. This novel technique differs from our perspective since the intention is not to fit a cuboid to a 3D object but to extract a present cuboid shape in an image.

The two methods (Jiang and Xiao, 2013; Jia et al., 2013) are similar with our approach since their authors try to fit cuboids using RANSAC to objects of a 3D scene acquired by a single RGB-D image. (Jia et al., 2013) followed a 3D reasoning approach and investigated different constraints that have to be applied to the cuboids, such as occlusion, stability and supporting relations. However, this method is applicable only to pre-labeled images. (Jiang and Xiao, 2013) coarsely segment the RGB-D image into roughly piecewise planar patches and for each pair of such patches fit a cuboid to the two planes. As a result, a large set of cuboid candidates is created. Finally, the best subset of cuboids is selected by optimizing an objective function, subject to various constraints. Hence, they require strong constraints (such as intersections between pairs of cuboids, number of cuboids, covered area on the image plane, occlusions

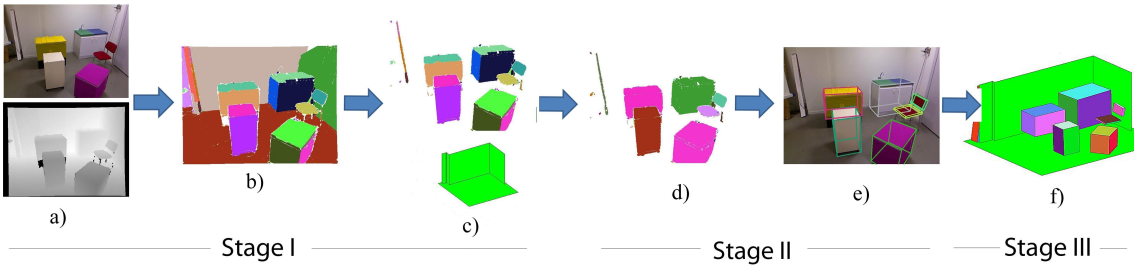

Figure 2: An overview of the proposed method.

among cuboids, etc.) during the global optimization process. This pioneer approach provides promising results in some cases but very coarse ones in others even for dramatically simple scenes (see Figs. 9 and 10 and images shown in (Jiang and Xiao, 2013)).

In this paper, in order to improve the quality of the reconstruction, we followed a different approach and propose an accurate segmentation step using novel constraints. The objective is to isolate the objects from each other before fitting the cuboids due to the fact that the cuboid fitting step can be significantly more efficient and accurate when working with each object independently.

3 METHOD OVERVIEW

The method proposed in this paper can be separated in three different stages. The first stage is to define the layout of the scene. This implies to extract the floor, all the walls and their intersections. For this purpose, the input RGB-D image is segmented by fitting 3D planes to the point cloud. The second stage is to segment all the objects in the scene and to fit a cuboid to each one separately. Finally, in stage 3 the results of the two previous stages are combined in order to visualize the 3D model of the room. An overview of this method can be seen in Fig. 2

3.1 Parsing the Indoor Scene

In order to parse the indoor scene and extract the complete layout of the scene, an approach based on the research of (Taylor and Cowley, 2012) is used. According to this work, the image is separated in planar regions by fitting planes to the point cloud using RANSAC, as can be seen in Fig 2b. Then the floor and the walls are detected by analyzing their surfaces, angles with vertical and angles between them. This method provides the layout of the room in less than 6 seconds. The final result of the layout of the scene, visualized in the 3D Manhattan World, can be seen in the bottom of Fig. 2c.

While working with depth values provided by the Kinect sensor, it is well known that the depth accuracy is not the same for the whole range of depth (Andersen et al., 2012), i.e. the depth information is more accurate for points that are close to the sensor than for points that are farther. This has to be taken into account in order to define a threshold according to which the points will be considered as inliers in a RANSAC method. Points with a distance to a plane inside the range of Kinect error should be treated as inliers of that plane. In order to address this problem, (Taylor and Cowley, 2012) proposed to fit planes in the disparity (inverse of depth) image instead of working directly with depth. This solution improves the accuracy but we claim that the best solution would be to use a threshold for the computation of the residual errors in RANSAC that increases according to the distance from the sensor. This varying threshold is computed once by fitting a second degree polynomial function to the depth values provided by (Andersen et al., 2012). The difference between the varying threshold proposed by (Taylor and Cowley, 2012) using disparity and the one proposed here can be seen in Fig. 3. As observed in the graph, our threshold follows significantly better the experimental data of (Andersen et al., 2012) compared to the threshold of (Taylor and Cowley, 2012).

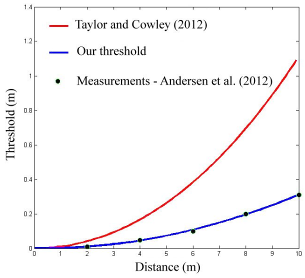

Figure 3: Comparison of the varying threshold set in (Taylor and Cowley, 2012) and the one proposed in this paper.

The impact of the proposed threshold on the room layout reconstruction can be seen in the two characteristic examples in Fig. 4. As it can be easily noticed, with the new threshold the corners of the walls are better defined and complete walls are now detected. This adaptive threshold is further used in the cuboid fitting step and significant improvements are obtained for various objects, as it is discussed in section 3.3.

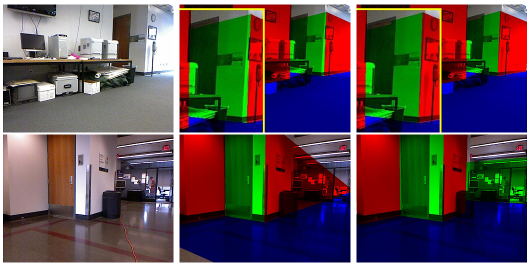

Figure 4: Impact of the proposed threshold in the room layout reconstruction. (left column): Input image (middle column): Threshold in (Taylor and Cowley, 2012). (right column): Threshold proposed here.

3.2 Segmenting the Objects in the Scene

As an output of the previous step, the input image is segmented in planar regions (Fig. 2b). Moreover, it is already known which of these planar regions correspond to the walls and to the floor in the scene (bottom of Fig. 2c). By excluding them from the image, only planar regions that belong to different objects in the image are left, as can be seen in the top of Fig. 2c. In order to segment the objects in the scene, the planar regions that belong to the same object have to be merged. For this purpose, the edges of the planar surfaces are extracted using a Canny edge detector and the common edge between neighboring surfaces is calculated. Then, we propose to merge two neighbor surfaces by analyzing i)the depth continuity across surface boundaries, ii)the angle between the surface normals and iii)the size of each surface.

For the first criterion, we consider that two neighboring planar surfaces that belong to the same object have similar depth values in their common edge and different ones when they belong to different objects. The threshold in the mean depth difference is set to 60 mm in all of our experiments. The second criterion is necessary in order to prevent patches that do not belong to the same object to be merged. In fact, since this study is focused on cuboids, the planar surfaces that should be merged need to be either parallel or perpendicular to each other. The final criterion forces neighboring planar surfaces to be merged if both of their sizes are relatively small (less than 500 points). The aim is to regroup all small planar regions that constitute an object that does not have a cuboid shape (sphere, cylinder, etc.). This point is illustrated in Fig. 5, where one cylinder is extracted. The proposed algorithm checks each planar region with respect to its neighboring regions (5 pixels area) in order to decide whether they have to be merged or not. This step is crucial for preparing the data before fitting cuboids in the next step.

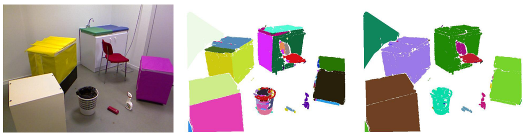

Figure 5: An example of merging objects that are not cuboids.(left) original input image. (middle):Before merging. (right):After merging.

3.3 Fitting a Cuboid to Each Object

The aim of this section is to fit an oriented cuboid to each object. As discussed by (Jia et al., 2013), the optimal cuboid is the one with the minimum volume and the maximum points on its surface. Since the image has been already segmented, i.e. each object is isolated from the scene, the strong global constraints used by (Jiang and Xiao, 2013) can be relaxed and more attention to each cuboid can be drawn. Therefore, we propose the following double-RANSAC process. Two perpendicular planar surfaces are sufficient to define a cuboid. Hence, in order to improve the robustness of the method, we propose to consider only the two biggest planar surfaces of each object. In fact, in a single viewpoint of a 3D scene only two surfaces of an object are often visible. Thus, first, for each segmented object, the planar surface with the maximum number of inliers is extracted by fitting a plane to the corresponding point cloud using RANSAC (with our adaptive threshold described in section 3.1). The orientation of this plane provides the first axis of the cuboid. We consider that the second plane is perpendicular to the first one but this information is not sufficient to define the second plane. Furthermore, in case of noise or when the object is thin (few points in the other planes) or far from the acquisition sensor, the 3D orientation of the second plane might be poorly estimated. Hence, we propose a robust solution which projects all the remaining points of the point cloud on the first plane and then fits a line using another RANSAC step to the projected points. The orientation of this line provides the orientation of the second plane. This is visualized in Fig. 6. In the experiments section, it is shown that this double RANSAC process provides very good results while fitting cuboids to small, thin or far objects.

Furthermore, as a second improvement of the

RANSAC algorithm, we propose to analyze its quality criterion. In fact, RANSAC fits several cuboids to each object (10 cuboids in our implementation) and selects the one that optimizes a given quality criterion. Thus, the chosen quality criterion has a big impact on the results. As it was discussed before, in RGB-D data a well estimated cuboid should have a maximum of points on its surface. Given one cuboid returned by one RANSAC iteration, we denote $area_{f1}$ and $area_{f2}$ the areas of its two faces and $area_{c1}$ and $area_{c2}$ the areas defined by the convex hull of the inlier points projected on these two faces, respectively. In order to evaluate the quality of the fitted cuboid, Jiang and Xiao proposed the measure defined as $min(\frac{area_{c1}}{area_{f1}}, \frac{area_{c2}}{area_{f2}})$ which is equal to the maximum value of 1 when the fitting is perfect. This measure assimilates the quality of a cuboid to the quality of the worst plane among the two, without taking into account the quality of the best fitting plane. Nevertheless, the quality of the best fitting plane could help in deciding between two cuboids characterized by the same ratio. Furthermore, the relative sizes of the two planes are completely ignored in this criterion. Indeed, in case of a cuboid composed by a very big plane and a very small one, this measure does not provide any information about which one is well fitted to the data, although this information is crucial to assess the quality of the cuboid fitting. Consequently, we propose to use a similar criterion which does not suffer from these drawbacks: $ratio = \frac{area_{c1}+area_{c2}}{area_{f1}+area_{f2}}$. Likewise, for an ideal fitting this measure is equal to 1. In order to illustrate the improvement due to the proposed adaptive threshold (of section 3.1) and the proposed ratio in the cuboid fitting step, 3 typical examples are shown in in Fig. 7. There, it can be seen that the proposed method (right column) increases significantly the performance for far and thin objects.

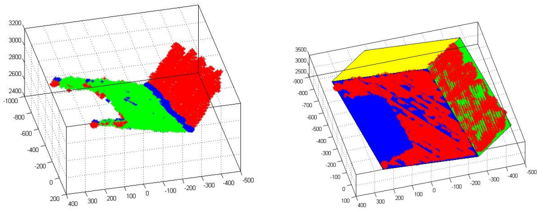

Figure 6: Illustration of our cuboid fitting step. (left): The inliers of the first fitted 3D plane are marked in green. The remaining points and their projection on the plane is marked in red and blue, respectively. A 3D line is fitted to these points. (right): The fitted cuboid.

In the final step of the method, the fitted cuboids are projected in the Manhattan World of the scene, in order to obtain the 3D model of the scene, as illustrated in Fig. 2f. Additionally, the cuboids are projected on the input RGB image in order to demonstrate how well the fitting procedure performs (see Fig. 2e).

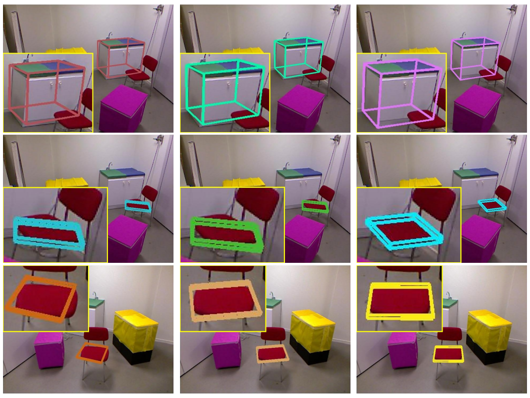

Figure 7: Impact of the selected threshold and ratio on the cuboid fitting. (left): Fixed global threshold and ratio proposed here. (middle): Varying threshold proposed here and ratio proposed in (Jiang and Xiao, 2013) (right): Threshold and ratio proposed here.

4 NEW GROUND TRUTH DATASET

For an objective evaluation, a new dataset with measured ground truth 3D positions was built. This dataset is composed by 4 different scenes and each scene is captured under 3 different viewpoints and 4 different illuminations. Thus, each scene consists of 12 images. For all these 4 scenes, the 3D positions of the vertices of the objects were measured using a telemeter. These coordinates constitute the ground truth. As the reference point was considered the intersection point of the three planes of the Manhattan World. It should be noted that the measurement of vertices positions in a 3D space with a telemeter is not perfectly accurate and the experimental measurements show that the precision of these ground truth data is approximately $\pm 3.85mm$. Some of the dataset images can be seen in the figures of the next section.

5 EXPERIMENTS

5.1 Qualitative Evaluation

As a first demonstration of the proposed method some reconstruction results are shown in Fig. 8. It can be seen that it performs well even in very demanding scenes with strong clutter. Moreover, it is able to handle small and thin objects with convex surfaces. Subsequently, our method is compared with the recent method proposed by (Jiang and Xiao, 2013) since their method not only performs cuboid fitting to RGB-D data but also outperforms various other approaches. A first visual comparison can be performed on both our dataset and the well-known NYUv2 Kinect Dataset (Silberman et al., 2012) in Figs. 9 and 10, respectively. It should be noted that all the thresholds in this paper were tuned to the provided numbers for both ours and the NYUv2 dataset. This point highlights the generality of our method that was tested in a wide variety of scenes. (Jiang and Xiao, 2013) have further improved their code and its last release (January 2014) was used for our comparisons. A random subset of 40 images that contain information about the layout of the room was selected from the NYUv2 Kinect dataset. The results imply that our method provides significantly better reconstructions than this state-of-the-art approach. Furthermore, in various scenes in Fig. 9, it can be observed that the global cuboid fitting method of (Jiang and Xiao, 2013) can result in cuboids that do not correspond to any object in the scene. The reason for this is the large set of candidate cuboids that they produce for each two planar surfaces in the image. The strong constraints that they apply afterwards, in order to eliminate the cuboids which do not correspond to an object, do not always guarantee an optimal solution. Another drawback of this approach is that the aforementioned constraints might eliminate a candidate cuboid that does belong to a salient object. In the next section, the improvement of our approach is quantified by an exhaustive test on our ground truth dataset.

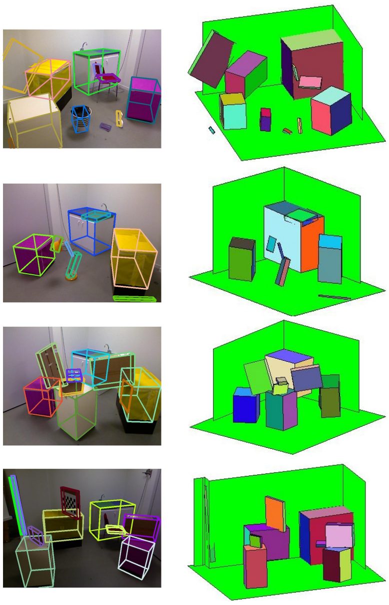

Figure 8: Various results of the proposed method on different real indoor scenes.

Table 1: Mean value (μ) and standard deviation (σ) of the Euclidean distances in mm between the ground truth and the reconstructed vertices over the 12 images of each scene and mean running time (t) in seconds of each algorithm.

	Our method			(Jiang and Xiao, 2013)		
	μ	σ	t^*	μ	σ	t^*
Scene 1	52.4	8.8	8.8	60.9	19.6	25.3
Scene 2	60.4	20.9	12.3	132.7	65.9	26.1
Scene 3	69.7	20.2	14.2	115.7	48.3	27.2
Scene 4	74.9	35.3	12.2	145.3	95.4	26.8

* *Running on a Dell Inspiron 3537, i7 1.8 GHz, 8 GB RAM*

5.2 Quantitative Evaluation

In order to test how accurate is the output of the proposed method and how robust it is against different viewpoints and illuminations, the following procedure was used. The 3D positions of the reconstructed vertices are compared to their ground truth positions by measuring their Euclidean distance. The mean value (μ) and the standard deviation (σ) of these Euclidean distances as well as the mean running time of the algorithm over the 12 images of each scene are presented in Table 1. The results using the code of (Jiang and Xiao, 2013) are included in the table for comparison. It should be noted that since this method does not provide the layout of the room, their estimated cuboids are rotated to the Manhattan World obtained by our method for each image.

During the experiments, it was noticed that the results of (Jiang and Xiao, 2013) were very unstable and various times their method could not provide a cuboid for each object in the scene. Moreover, since the RANSAC algorithm is non-deterministic, neither are both our approach and the one of (Jiang and Xiao, 2013). In order to quantify this instability, each algorithm was run 10 times on the exact same image (randomly chosen) of each scene. The mean (μ) and standard deviation (σ) of the Euclidean distance between the ground truth and the reconstructed 3D positions were measured. The results are presented in Table 2. It should be noted that the resulting 3D positions of both algorithms are estimated according to the origin of the estimated layout of the room. Thus, the poor resolution of the Kinect sensor is perturbing the estimation of both the layout and the 3D positions of the objects and the errors are cumulating. However, the values of the mean and standard deviation for our method are relatively low with respect to the depth resolution of Kinect sensor at that distance, which is approximately 50 mm at 4 meters (Andersen

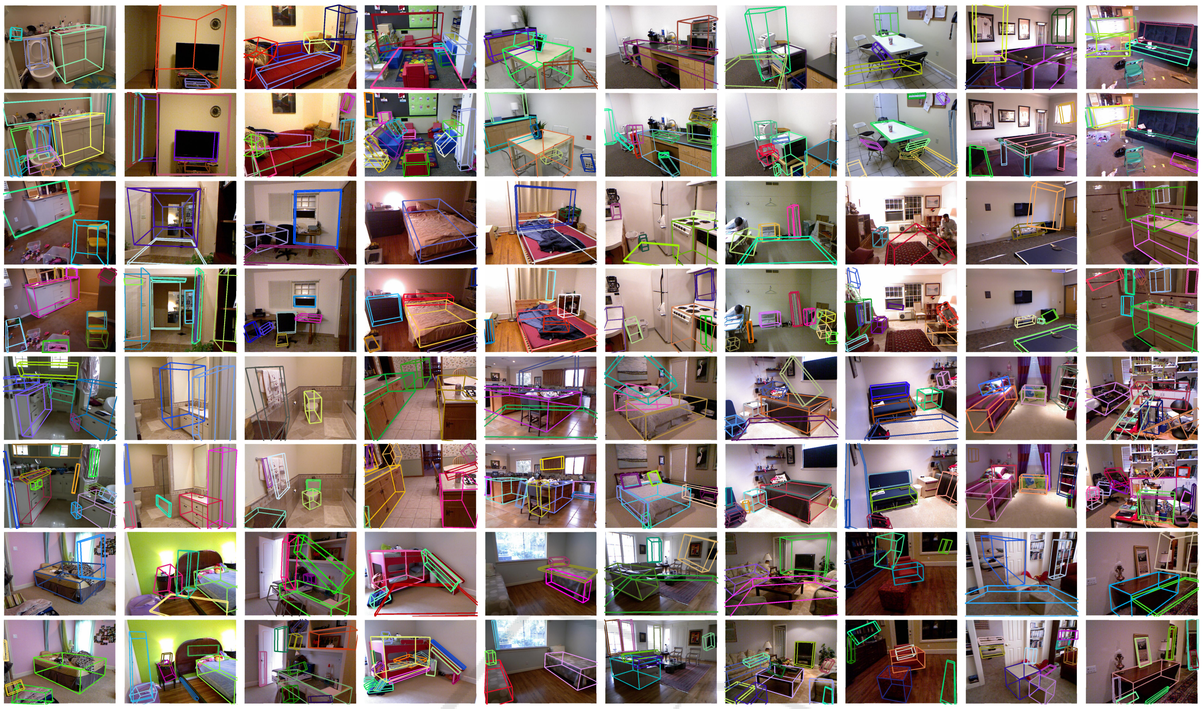

Figure 9: Comparison of the results obtained by (Jiang and Xiao, 2013) (odd rows) and the method proposed in this paper (even rows) for the NYUv2 Kinect dataset.

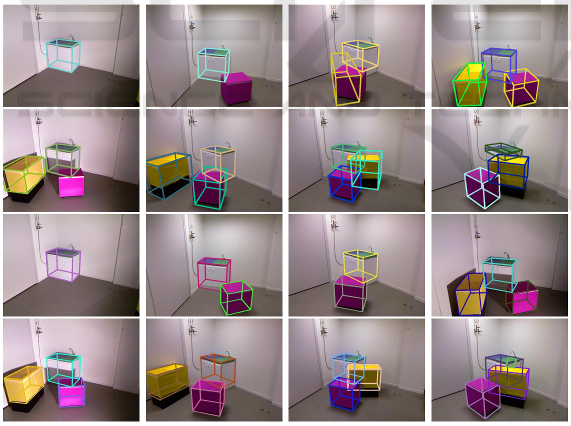

Figure 10: Random results of (Jiang and Xiao, 2013) (top 2 rows) and the corresponding ones of our method (bottom 2 rows) for the ground truth dataset.

et al., 2012). Furthermore, the standard deviations of Table 2 are considerably low and state a maximum deviation of the result less than 4.5 mm.

Finally, as can be seen in Table 1, the computational cost of our method is dramatically lower than the one of (Jiang and Xiao, 2013). It should be noted that in this running time our method estimates the complete 3D scene reconstruction of the scene. It requires around 9 seconds for a simple scene and less than 20 seconds for a demanding scene with strong clutter and occlusion on a Dell Inspiron 3537, i7 1.8 Ghz, 8 GB RAM. It is worth mentioning that no optimization was done in the implementation. Thus, the aforementioned running times could be considerably lower.

Table 2: Mean value (μ) and standard deviation (σ) of the Euclidean distances between the ground truth and the reconstructed vertices over 10 iterations of the algorithm on the same image.

	Our method		(Jiang and Xiao, 2013)	
	μ (mm)	σ (mm)	μ (mm)	σ (mm)
Scene 1	50.2	2.7	54.3	10.5
Scene 2	57.0	2.9	104.9	37.8
Scene 3	81.9	2.7	93.2	20.5
Scene 4	72.0	4.5	195.4	35.7

6 CONCLUSIONS

In this paper, a new method that provides accurate 3D reconstruction of an indoor scene using a single RGB-D image is proposed. First, the layout of the scene is extracted by exploiting and improving the method of (Taylor and Cowley, 2012). The latter is achieved by better addressing the problem of the non-linear relationship between depth resolution and distance from the sensor. For the 3D reconstruction of the scene, we propose to fit cuboids to the objects composing the scene since this shape is well adapted to most of the indoor objects. Unlike the state-of-the-art method (Jiang and Xiao, 2013) that runs a global optimization process over sets of cuboids with strong constraints, we propose to automatically segment the image, as a preliminary step, in order to focus on the local cuboid fitting on each extracted object. It is shown that our method is robust to viewpoint and object orientation variations. It is able to provide meaningful interpretations even in scenes with strong clutter and occlusion. More importantly, it outperforms the state-of-the-art approach not only in accuracy but also in robustness and time complexity. Finally, a ground truth dataset for which the exact 3D positions of the objects have been measured is provided. This dataset can be used for future comparisons.

REFERENCES

Andersen, M. R., Jensen, T., Lisouski, P., Mortensen, A., Hansen, M., Gregersen, T., and Ahrendt, P. (2012). Kinect depth sensor evaluation for computer vision applications. Technical Report ECE-TR-06, Aarhus University.

Bazin, J. C., Seo, Y., Demonceaux, C., Vasseur, P., Ikeuchi, K., Kweon, I., and Pollefeys, M. (2012). Globally optimal line clustering and vanishing point estimation in manhattan world. In *CVPR*, pages 638–645.

Coughlan, J. M. and Yuille, A. L. (1999). Manhattan world: Compass direction from a single image by bayesian inference. In *ICCV*, pages 941–947.

Cupec, R., Nyarko, E. K., and Filko, D. (2011). Fast 2.5d mesh segmentation to approximately convex surfaces. In *ECMR*, pages 49–54.

Dou, M., Guan, L., Frahm, J.-M., and Fuchs, H. (2013). Exploring high-level plane primitives for indoor 3d reconstruction with a hand-held rgb-d camera. In *ACCV 2012 Workshops*, volume 7729 of *Lecture Notes in Computer Science*, pages 94–108. Springer Berlin Heidelberg.

Dwibedi, D., Malisiewicz, T., Badrinarayanan, V., and Rabinovich, A. (2016). Deep cuboid detection: Beyond 2d bounding boxes.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

Hedau, V., Hoiem, D., and Forsyth, D. (2009). Recovering the spatial layout of cluttered rooms. In *ICCV*, pages 1849–1856.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. (2011). Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *UIST*, pages 559–568.

Jia, Z., Gallagher, A., Saxena, A., and Chen, T. (2013). 3d-based reasoning with blocks, support, and stability. In *CVPR*.

Jiang, H. (2014). *Finding Approximate Convex Shapes in RGBD Images*, pages 582–596. Springer International Publishing, Cham.

Jiang, H. and Xiao, J. (2013). A linear approach to matching cuboids in rgbd images. In *CVPR*.

Lin, D., Fidler, S., and Urtasun, R. (2013). Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, pages 1417–1424.

Mirzaei, F. and Roumeliotis, S. (2011). Optimal estimation of vanishing points in a manhattan world. In *ICCV*, pages 2454–2461.

Neumann, D., Lugauer, F., Bauer, S., Wasza, J., and Hornegger, J. (2011). Real-time rgb-d mapping and 3-d modeling on the gpu using the random ball cover data structure. In *ICCV Workshops*, pages 1161–1167.

Neverova, N., Muselet, D., and Trémeau, A. (2013). 2 1/2 d scene reconstruction of indoor scenes from single rgb-d images. In *CCIW*, pages 281–295.

Piegl, L. (1991). On nurbs: a survey. *IEEE Computer Graphics and Applications*, 11(1):55–71.

Ren, Z. and Sudderth, E. B. (2016). Three-dimensional object detection and layout prediction using clouds of oriented gradients. IEEE CVPR.

Richtsfeld, A., Mörwald, T., Prankl, J., Balzer, J., Zillich, M., and Vincze, M. (2012). Towards scene understanding - object segmentation using rgbd-images. In *CVWW*.

Schwing, A. and Urtasun, R. (2012). Efficient exact inference for 3d indoor scene understanding. In *ECCV*, volume 7577 of *Lecture Notes in Computer Science*, pages 299–313. Springer Berlin Heidelberg.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer.

Taylor, C. and Cowley, A. (2011). Fast scene analysis using image and range data. In *ICRA*, pages 3562–3567.

Taylor, C. and Cowley, A. (2012). Parsing indoor scenes using rgb-d imagery. In *RSS*.

Zhang, J., Kan, C., Schwing, A. G., and Urtasun, R. (2013). Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In *ICCV*, pages 1273–1280.

Zhang, Y., Song, S., Tan, P., and Xiao, J. (2014). *PanoContext: A Whole-Room 3D Context Model for Panoramic Scene Understanding*, pages 668–686. Springer International Publishing, Cham.