# Plane Extraction For Indoor Place Recognition

Ciro Potena, Alberto Pretto, Domenico D. Bloisi and Daniele Nardi

Department of Computer, Control and Management Engineering
Sapienza University of Rome
via Ariosto 25, 00185, Rome, Italy

**Abstract.** In this paper, we present an image based plane extraction method well suited for real-time operations. Our approach exploits the assumption that the surrounding scene is mainly composed by planes disposed in known directions. Planes are detected from a single image exploiting a voting scheme that takes into account the vanishing lines. Then, candidate planes are validated and merged using a region growing based approach to detect in real-time planes inside an unknown indoor environment. Using the related plane homographies is possible to remove the perspective distortion, enabling standard place recognition algorithms to work in an invariant point of view setup. Quantitative Experiments performed with real world images show the effectiveness of our approach compared with a very popular method.

## 1 Introduction

Place recognition is the problem of identifying from images a place already seen before, or a place represented by a set of images included in a given database. The place recognition problem is often referred to as the "loop closure detection" problem, and it has been addressed in a number of works (e.g., [4, 9, 12]).

However, most of these methods assume that same places are seen at multiple times with approximately the same position and orientation. This means that, if the same place is seen more times form different points of view, using a perspective camera no loops closure are usually detected. This is due to the perspective distortion of the camera that, for different points of view, projects the same scene in different ways.

A possible solution is to employ an omni-directional camera, which provides a rotational invariant snapshot of the surrounding scene in a single frame. If an omni-directional camera is not available, a solution is to remove the perspective distortion from the images, at least where the three dimensional (3D) structure of the scene allows this process. An example is shown in Fig. 1, where two images of the same room are captured from opposite points of view (the black bounding boxes represent the same plane into the two views). Even if the two images represent the same place, they cannot be used directly as input for common place recognition techniques, since they are captured with very different view angles.

Fig. 1c and Fig. 1d show the two planes highlighted in the input images after removing the perspective distortion: they can be used for robust place recognition


(a) (b) (c) (d)

Fig. 1: a) and b) two images of a room captured from opposite points of view. The black boxes highlight a plane seen from both views. c) and d) the two planes highlighted in black after removing the perspective distortion.

purposes. The capability to recognize planar surfaces in the environment enables to obtain a canonical, point of view invariant, visual representation of the places of interest.

In this work, we present a robust and fast plane recognition method that works with single images. We exploit the Manhattan World assumption [3], considering three mutually orthogonal directions as possible orientations. This hypothesis is well justified by the fact that most indoor environments meet this assumption. In order to extract the planes, the vanishing points are calculated from the line segments detected in the images. This represents a starting point for successive information extraction about the 3D structure of the scene. The image is then segmented to obtain three regions, representing mutually orthogonal directions. Finally, contextual information and a region growing technique are used to refine the result. From the extracted planes, it is possible to recover the related homographies that enable to re-project the plane in a canonical way (i.e., to remove the perspective distortion), independently from the initial point of view (see Fig. 1). These planes can be used to solve the place recognition problem by using conventional methods. Quantitative experiments show that our method outperforms the line-sweep approach presented by Lee *et al.* in [13].

The remainder of the paper is organized as follows. Related work is discussed in Section 2. Section 3 describes the key modules of the proposed algorithm and

quantitative experimental results are reported in Section 4. Finally, conclusions and future directions are given in Section 5.

## 2 Related Work

The problem of recovering a model for indoor scenes from images is an extensively studied problem. Existing methods can be grouped according to the nature of the input data that are processed to compute the model. In particular, three categories can be identified: single image, Multi-View Stereo (MVS) images, and RGB-D data.

***Single image.*** One of the most popular single image approach has been presented by Lee *et al.* in [13]. The Indoor Manhattan World assumption and a set of rules, describing geometric constraints between groups of segments, is exploited to generate a scene hypothesis with the most plausible interpretation. Despite its apparent simplicity, this method is efficient and can correctly reconstruct a great variety of indoor scenes.

The method described in [13] has been improved by Flint *et al.* in [6] by employing a dynamic programming approach: This algorithm exhibits linear computational complexity in both model complexity and image size, and it works also with partially occluded scenes.

Saxena *et al.* in [16] propose to segment the image into small homogeneous regions (i.e., superpixels). Their orientations and 3D positions are then inferred inside a Markov Random Field framework.

Hoiem *et al.* in [11] use learning appearance-based models of geometric classes to estimate planes in the scene structure, providing also a confidence value for each inferred geometric label.

***Multi-View Stereo.*** Baker *et al.* in [1] propose a MVS approach that starts from a photo consistency estimate and then refines it by using a re-synthesis algorithm, which takes into account both occlusions and mixed pixels. The output represents the scene as a collection of approximately planar layers.

Fukurawa *et al.* in [8] use stereo pairs to reconstruct textured regions. Dominant plane directions are extracted, generating a set of plane hypotheses. Per-view depth maps are finally recovered by using Markov Random Fields.

All the above-listed MVS algorithms often require textured regions in order to give accurate results. Therefore, they can work poorly for many architectural scenes (e.g., for building interiors with textureless regions, painted walls). Actually, they give no relevant improvements with respect to a monocular set up.

***RGB-D data.*** Silberman *et al.* [17] propose to use a Conditional Random Field (CRF) based model to evaluate a range of different representations for depth information and a novel prior on 3D location. In Guan *et al.* [10], the image is segmented in an initial number of planes, followed by a pixel-to-plane assignment. Plane equations are iteratively refined.

A common limitation of the depth based algorithms is that they suffer in presence of crowded places. In addition, depth sensors (e.g., Kinect) can have a limited field of view.

Our aim is to extract planes in a robust way, avoiding corruptions provided by obstacles near walls or people in common environment, and in real-time. Given the limitations of MVS and RGB-D algorithms, we have decided to follow a single camera approach.

## 3 Indoor Manhattan World Plane Extraction

Given a single image of a scene, our goal is to calculate the orientation of each pixel. A perfectly uncluttered indoor environment can be represented exactly by an indoor Manhattan model, though, in general, we expect to encounter clutter and, in such cases we aim to recovering the orientation of the environment in spite of this distraction. We aim to ignore completely all the objects within the environment and to reconstruct only the main structure of the scene, in contrast to most previous approaches that aim to reconstruct the entire scene. This choice is due to our intention of using the models as input for successive higher-level reasoning steps.

The Manhattan World assumption, as described in [3], states that world surfaces are oriented in one of three mutually orthogonal directions. In case of indoor scenes, it also states that the environment consists of a floor plane, a ceiling plane, and a set of walls extending vertically between them. In addition, indoor environments usually have a single floor plane and a single ceiling plane with constant ceiling height.

With these simplifying hypothesis an indoor scene can often be fully represented by corners, thus geometric constraints on corners will guarantee the entire structure to be valid.

Indoor Manhattan models are very interesting because they can represent many indoor and outdoor environments, with an adeguate level of precision. Indeed, with this kind of geometric model, it is possible to reconstruct the indoor world, as done in [13], or make a decomposition of the scene, as described in [6,7].

The two above cited approaches require an initial orientation hypothesis, on which the method proposed in this paper rests. The entire pipeline of our approach to reconstruct indoor Manhattan environment is shown in Fig. 2. The computer steps are given in Algorithm 1.

### 3.1 Vanishing Point Extraction

We extract the three vanishing points from the detected lines. This extraction process is highly influenced by the illumination of the environment. In order to reduce this effect, i.e., for recovering edge and contours in the part of the image where there are variations in the illumination conditions, we carry out an image contrast enhancement based on two main steps: First, by using a histogram

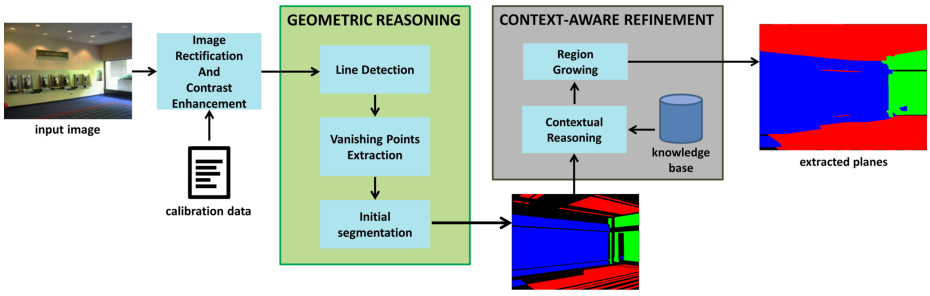



Fig. 2: Overall functional architecture of the proposed algorithm for plane extraction from single images.

**Algorithm 1:** Plane extraction from single image.

**input** : $m \times n$ RGB image $I$
**data** : $m \times n$ RGB images $R$, $E$, and $T$; vector of lines $D$; vectors of points $CurrModel$ and $BestModel$; scalar $Score$
**output**: $m \times n$ RGB image $S$

$R \leftarrow$ `ImageRectification(`$I$`)`
$E \leftarrow$ `Enhancing(`$R$`)`
$D \leftarrow$ `LineExtraction(`$E$`)`
**initialize** $BestScore$
**for** $i = 1$ **to** $max\_iter$ **do**
    D$\leftarrow$`LineSampling(`$D$`)`
    $Score \leftarrow$`ScoringFunction(`$CurrModel$`)`
    **if** $Score \geq BestScore$ **then**
        $BestModel = CurrModel$
        $BestScore = Score$
$T \leftarrow$ `SweepRegion(`$BestModel$`)`
$S \leftarrow$ `OrientationRefinment(`$T$`)`

equalization and then, by increasing the contrast of the image. In particular, the contrast is increased by mapping the values of the input intensity image to new values such that, by default, 1% of the data are saturated at low and high intensities of the input data. The lines are then extracted by using a standard Canny edge detector [2] with a probabilistic Hough lines extractor [14]. The line segments that belong to the same lines are linked together, while short lines are filtered out, since they often corrupt the plane estimation process.

We adopt the method proposed by Rother [15] to find three orthogonal vanishing points. Two pairs of lines are randomly sampled in a RANSAC fashion and the intersection of each pair of lines generates a candidate vanishing point. A voting scheme, with a cost function that takes into account the number of classified lines, is used for choosing the best model. The third vanishing point is computed to be orthogonal to the two previous calculated vanishing points.

(a) (b)

Fig. 3: An example of vanishing line extraction. a) Input image. b) Detected, mutually orthogonal, vanishing lines.

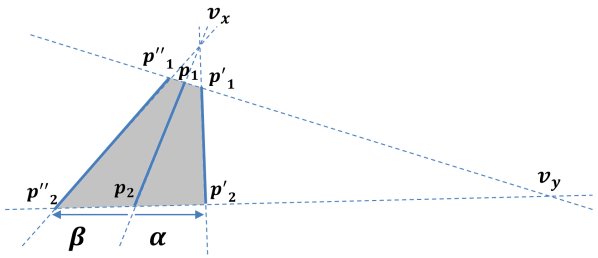



Fig. 4: The shaded area denotes the sweep $S(l_{x,i}, v_y, \alpha)$ of a line $l$ towards and backwards vanishing point $v_y$ by amount $\alpha$ and $\beta$. It potentially supports the region to be orthogonal to $v_x$ and $v_y$.

From the three vanishing points, we can recover the orientation of the three principal axis of the 3D structure. All the extracted planes will have one of these three orientation.

Is important to note that the proposed technique can fail in some particular cases, e.g., when there are no lines in one of the three direction, or when many lines don't belong to the principal directions. To avoid these problems, we simply discard the estimated model if the error is greater than a fixed threshold. An example of the final line labelling is shown in Fig. 3.

### 3.2 Initial Building Orientation

Once the vanishing points have been estimated, we generate an initial segmentation, also denoted as the building hypothesis, which represents an initial estimate of the indoor structure. For this initial estimation, we use the approach described in [13], called geometric reasoning, due to its robustness and efficiency. The geometric reasoning approach , in fact, uses only the detected lines because they

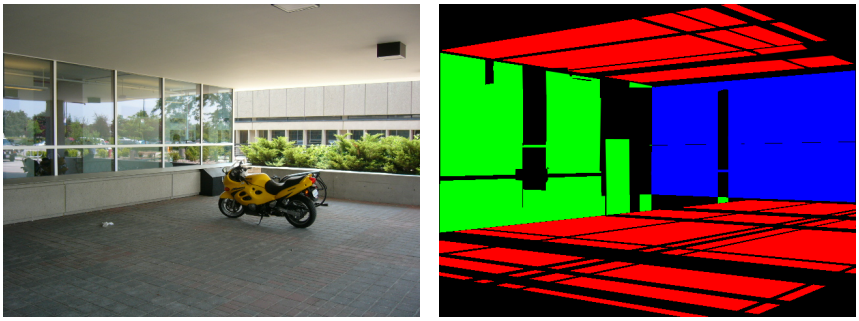


Fig. 5: Example of a building model extracted from the York Urban Line Segments Database [5] as described by Lee *et al.* in [13].

give a strong indication about local orientation in the image. In other words, if a pixel is supported by two line segments with different orientation, there is a strong probability that it has a perpendicular orientation with respect to two line segments. This assumption is true in a world with all mutually orthogonal lines too, usually producing accurate orientation map, except around occluding objects or people. More formally, let $L_x = \{l_{x,1}, l_{x,2}, ..., l_{x,n}\}$ the set of extracted lines, where $x \in \{1, 2, 3\}$ is one of the three possible orientations among the calculated vanishing points and $n$ represents the total number of the extracted lines. The "sweep" $S(l_{x,i}, v_y, \alpha)$ of a line $l_{x,i}$ is a pixel region with the normal along the third vanishing point $z$ and extends up to a the nearest line with the same orientation, going toward the vanishing point $v_y$. $\alpha$ is the the width of this region, called "amount". For the same line $l_{x,i}$ is computed another "sweep" away from the vanishing point $v_y$, in this case with amount $\beta$. An example is shown in Fig. 4. The total set of pixels supported by $L_x$ swept towards $v_y$ is:

$$\bigcup_{l_{x,i} \in L_x} S(l_{x,i}, v_y, \hat{\alpha}_{x,i}) \cup S(l_{x,i}, v_y, \hat{\beta}_{x,i}).$$

A pixel is labeled with orientation $z$ when two lines of different orientation $x$ and $y$ exclusively support it, i.e., the pixel can have a possible orientation only. The complete initial building model is:

$$R_z = P_{x,y,z} \cap P_{y,x,z}$$
$$O_z = R_z \cap R_x \cap R_y.$$

An example of building model generation is shown in Fig. 5.

### 3.3 3D Structure Hypothesis Refinement

The initial structure model can result inaccurate when too few lines can be extracted or if the scene contains many occluding objects. This negatively affects

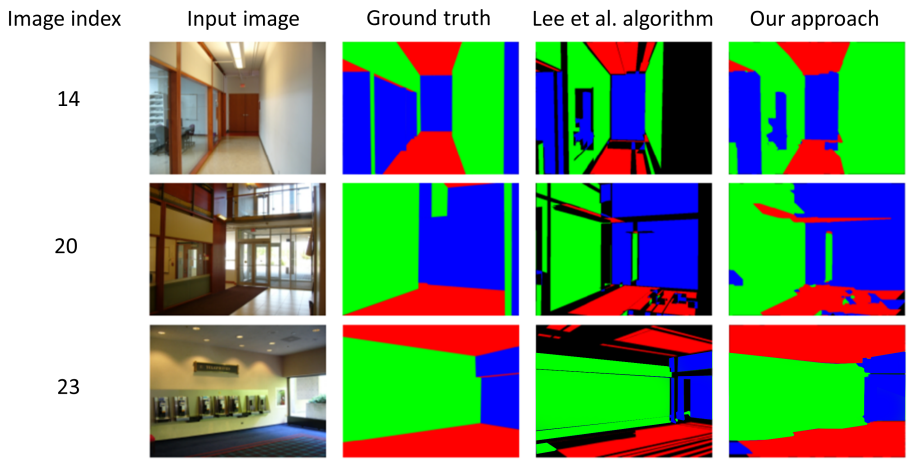



Fig. 6: Some sample images with the pixels orientation estimated by our method. Each row contains four images: the input image, the ground truth image, the initial orientation estimate by lee et al. [13], and the final output generated by our method.

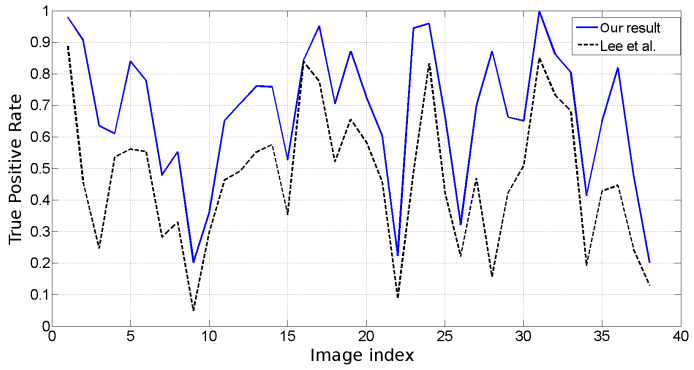
directly also higher-level processes, like navigation, scene recognition or 3D indoor scene reconstruction. Actually, the line extraction process highly depends on the lighting conditions and the resolution of the camera.

As a consequence, in many cases of indoor scenes the lines on the floor and on the ceiling are badly extracted. Thus, the sweep regions $S(l_{x,i}, v_y, \alpha)$, where $x$ is the orientation of the floor or ceiling lines and $v_y$ is the direction of the vanishing point orthogonal to $v_x$ and to the vertical orientation, support only few pixels and the amount of pixel with orientation $R_z$ will be very small.

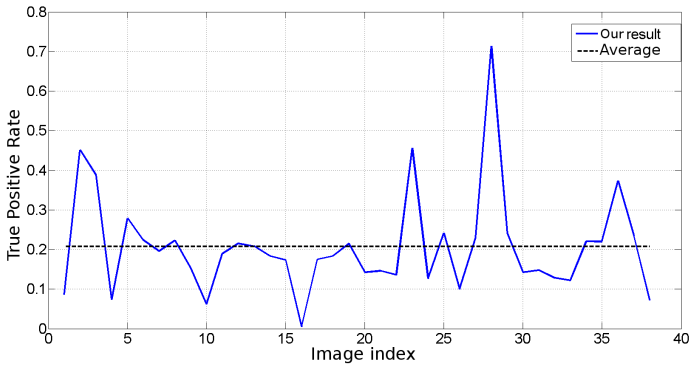
We propose a solution for this problem, that uses contextual information. In particular the indoor Manhattan scene has exactly one floor and one ceiling plane, both with orientation $z$. First, we find the unlabeled region of the initial segmentation that can probably belong to the floor or ceiling regions. For this purpose, a floor or a ceiling always has as boundary at least two vertical walls. If these regions are supported by one of the two sweeps of $R_z$, we classify them as a floor or a ceiling. The second step we propose is a region growing algorithm. The aim is to classify the pixels that are not yet labeled with an orientation, e.g. to occluding objects. Formally, for every unlabeled pixel in the image with position $(u, v)$ we define the sum of its neighborhood as:

$$\beta_{u,v,i} = f(u \pm 1, v \pm 1)$$

where $f$ is a linear filter and the index $i = \{x, y, z\}$ is the orientation along the three vanishing points in which the sum is computed. Repeating the above operation for the three orientation images $R_x$, $R_y$, $R_z$ , we are able to classify the unlabeled pixels with the following rule:




(a)




(b)

Fig. 7: a) Comparison of the true positive rate for the pixel with orientation $z$. b) Difference of the true positive rate between the two approaches.

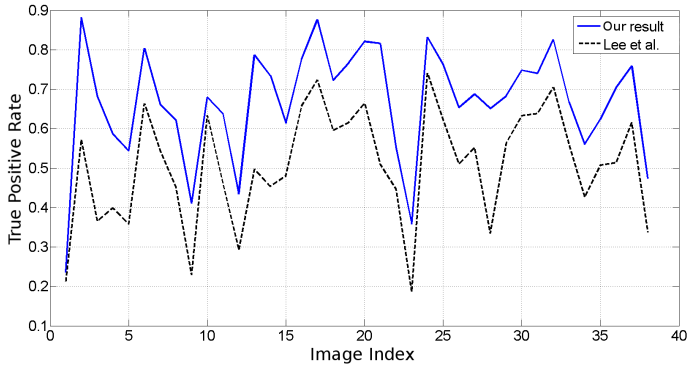
$$d(u,v) = \arg\max_{u,v}\{\beta_{u,v,x}, \beta_{u,v,y}, \beta_{u,v,z}\}$$

$$l_{u,v} = \begin{cases} x \;\; if \;\; d(u,v) > 3 \;\; \wedge \;\; d(u,v) = \beta_{u,v,x} \\ y \;\; if \;\; d(u,v) > 3 \;\; \wedge \;\; d(u,v) = \beta_{u,v,y} \\ z \;\; if \;\; d(u,v) > 3 \;\; \wedge \;\; d(u,v) = \beta_{u,v,z} \end{cases}$$

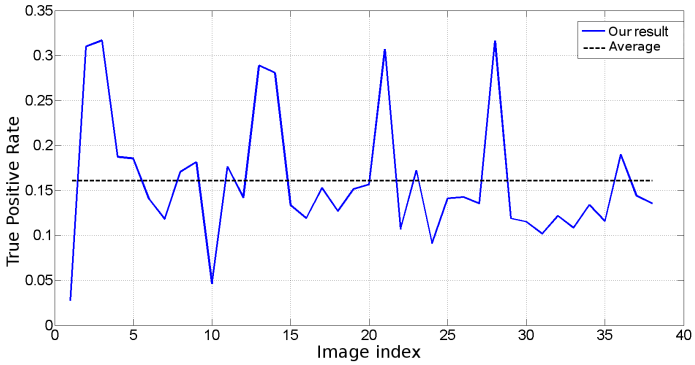
where $l_{u,v}$ is the orientation assigned to the unlabeled pixel at hand in position $(u,v)$.

## 4 Experimental results

For evaluating our method we used the publicly available York Urban Line Segment Database [5]. It is a compilation of 102 images (45 indoor, 57 outdoor) of urban environments. Most of them consist of scenes from the campus of York




(a)




(b)

Fig. 8: a) True positive rate for all the three orientations. b) Difference of the true positive rate between the two approaches.

University and downtown Toronto, Canada. The images are 640 x 480 in size and have been taken with a calibrated Panasonic Lumix DMC-LC80 digital camera.

We have manually labeled every pixel with three orientation values in all the indoor scenes, ignoring the occluding object, in order to create the ground truth images. Qualitative results on three sample frames from the York Urban data set are shown in Fig. 6. The results of our method are compared with the output generated by the algorithm of Lee et al. [13] with respect to corresponding ground truth images. As quality metric, we have used the True Positive Rate between the labeled pixel with respect to the ground truth. Fig. 7 shows the high percentage of correctly labeled pixel with orientation $z$ of our approach, due to the indoor assumptions. The average percentage of improvement is around 20% and, in a few cases, it is higher than the 50%.

Fig. 8 shows the True Positive rate among all labeled pixel of the input images. Also in this case, the percentage of correctly labeled pixel is increased significantly with an average of 16% by our method.

## 5 Conclusions

We have proposed a novel method for interpreting a collection of line segments for recovering a model of indoor scenes. Starting from an initial estimate we have shown that, by using additional prior knowledge and a region growing mechanism, it is possible to increase the accuracy of the computed model.

Quantitative experiments have been carried out on a publicly available dataset, containing indoor images. As future work, we intend to use the computed model for reconstructing the 3D structure of the captured indoor environment.

## References


1. Baker, S., Szeliski, R., Anandan, P.: A layered approach to stereo reconstruction. In: CVPR. pp. 434–441 (1998)
2. Canny, J.: A computational approach to edge detection 8, 679–697 (1986)
3. Coughlan, J., Yuille, A.: Manhattan world: compass direction from a single image by bayesian inference. In: CVPR. pp. 941–947 vol.2 (1999)
4. Cummins, M., Newman, P.: Appearance-only SLAM at large scale with FAB-MAP 2.0. The International Journal of Robotics Research (2010)
5. Denis, P., Elder, J.H., Estrada, F.J.: Efficient edge-based methods for estimating manhattan frames in urban imagery. In: Proceedings of the 10th European Conference on Computer Vision: Part II. pp. 197–210. ECCV '08 (2008)
6. Flint, A., Mei, C., Murray, D., Reid, I.: A dynamic programming approach to reconstructing building interiors. In: ECCV. pp. 394–407. Springer (2010)
7. Flint, A., Murray, D., Reid, I.: Manhattan scene understanding using monocular, stereo, and 3d features. In: ICCV. pp. 2228–2235 (Nov 2011)
8. Furukawa, Y., Curless, B., Seitz, S., Szeliski, R.: Manhattan-world stereo. CVPR (2009)
9. Galvez-Lopez, D., Tardos, J.: Real-time loop detection with bags of binary words. In: IROS. pp. 51–58 (Sept 2011)
10. Guan, L., Yu, T., Tu, P., Lim, S.N.: Simultaneous image segmentation and 3d plane fitting for rgb-d sensors x2014; an iterative framework. In: CVPR Workshops. pp. 49–56 (2012)
11. Hoiem, D., Efros, A., Hebert, M.: Geometric context from a single image. In: ICCV. pp. 654–661 Vol. 1 (2005)
12. Labbe, M., Michaud, F.: Appearance-based loop closure detection for online large-scale and long-term operation. Robotics, IEEE Transactions on 29(3), 734–745 (June 2013)
13. Lee, D., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: CVPR. pp. 2136–2143 (2009)
14. Matas, J., Galambos, C., Kittler, J.: Robust detection of lines using the progressive probabilistic hough transform. CVIU (Apr 2000)
15. Rother, C.: A new approach for vanishing point detection in architectural environments. In: BMVC. pp. 382–391 (2000)
16. Saxena, A., Sun, M., Ng, A.: Make3d: Learning 3d scene structure from a single still image. PAMI pp. 824–840 (2009)
17. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: ICCV Workshops. pp. 601–608 (2011)