Robust Manhattan Frame Estimation from a Single RGB-D Image

Bernard Ghanem[1], Ali Thabet[1], Juan Carlos Niebles[2], Fabian Caba[1]
[1]King Abdullah University of Science and Technology (KAUST), Saudi Arabia. [2]Universidad del Norte, Colombia.

The representation of indoor scenes using the Manhattan world assumption [1] has been widely used in computer vision and robotics applications, which take advantage of this assumption to simplify object representations w.r.t. the scene layout. This simplification states that most objects in an indoor scene are composed of planar surfaces aligned to one of three orthogonal directions. This set of orthogonal directions is referred to as the Manhattan Frame (MF) of the scene. In this paper, we wish to effectively and efficiently determine the MF of an indoor scene in the presence of noise and outliers. Our motivation is that an accurate MF estimate can assist in a variety of problems, such as RGB-D SLAM and 3D object understanding.

In this work, **(1)** we propose an accurate, fast, reliable, and robust method to estimate the MF of an indoor scene using a single RGB-D image. **(2)** In order to evaluate the properties of our method, we introduce a new evaluation benchmark that comprises ground truth MFs for the popular NYUv2 dataset [3]. We also compare our method against several MF algorithms in the literature, and show that our approach outperforms state-of-the-art techniques in terms of accuracy and speed. **(3)** We perform controlled tests to evaluate the repeatability and robustness of our method in challenging scenarios. **(4)** We show how our method can be used in addressing a popular vision problem, namely RGB-D SLAM, where our algorithm is shown to improve the performance of a popular SLAM method.

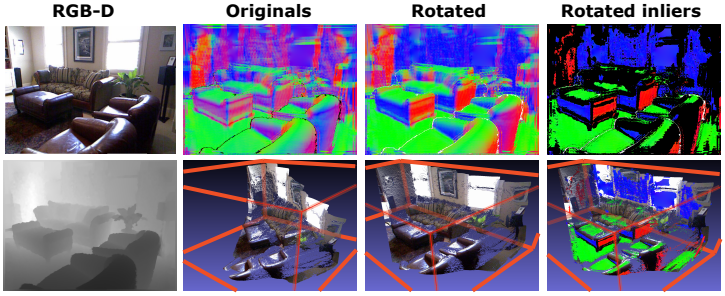

Figure 1: Overview of our method. **RGB-D** Top: Original RGB image. Bottom: Inpainted depth image from NYUv2 dataset. **Originals** Top: Original normals. Bottom: Original 3D point cloud. **Rotated** Top: Normals after alignment with our method. Bottom: Aligned 3D point cloud, where the wall, sofas, and tables are well aligned with the MF of the scene. **Rotated Inliers** Top: Our algorithm estimates as inliers those normals that can be aligned to one of the coordinate axes. Here, we color-code inlier normals according to the axis they are aligned to; black pixels are outliers. Bottom: Aligned 3D point cloud with color-coded inliers; outliers (non-planar objects, surfaces that cannot be aligned) retain their original RGB color.

The aim of indoor scene MF estimation is to determine the three *dominant* directions, along which most surfaces and possibly lines are oriented. Similar to previous work, we study indoor scenes that have an inherent Manhattan structure. Therefore, estimating the MF becomes equivalent to computing the *best* 3D rotation matrix $\mathbf{R}$ that transforms surface normals (and line directions if available) in the scene to the three unit directions or their reflections about the center. In fact, the rows of $\mathbf{R}$ define the dominant directions of the scene in the original coordinate system.

Problem Formulation: Applying a scene's MF $\mathbf{R}$ to the matrix of scene normals $\mathbf{N} \in \mathbb{R}^{3\times m}$ should lead to a matrix $\mathbf{X}$, whose columns are sparse. In the absence of noise, $\mathbf{X}$ should be the sparsest possible matrix such that $\|\mathbf{X}\|_0 = \|\mathbf{X}\|_{1,1} = m$. Equality holds here because the columns of $\mathbf{X}$ have unit norm. This observation establishes the basis of our proposed solution. In the presence of noise (e.g. due to noisy depth measurements and normal computation) and outliers (e.g. non-Manhattan surfaces in the scene), we incorporate the above observation to formulate the Robust MF Estimation (RMFE) problem in Eq (1). The first term penalizes reconstruction error, while the second term serves as a sparse regularizer.

$$(RMFE): \quad \min_{\mathbf{R}\in SO(3),\mathbf{X},\mathbf{E}} \ \|\mathbf{E}^T\|_{2,1} + \lambda\|\mathbf{X}\|_{1,1} \qquad (1)$$
$$\text{subject to: } \mathbf{RN} = \mathbf{X} + \mathbf{E}$$

The RMFE problem above can be solved efficiently using alternating optimization and IALM (or ADMM), where the solution is achieved after iterations of closed-form update steps of the primal and dual variables of Eq (1). For more details, please refer to the paper and the supplementary material.

Benchmark: We assess our RMFE method from three perspectives. Due the lack of rigorous evaluation of MF algorithms, there is no standardized dataset and ground truth available in the literature. To fill this gap, we contribute two sets of annotated data for evaluation. **(i)** We create a new benchmark framework for evaluating MF estimation algorithms from RGB-D images by generating MF ground truth (rotation matrix) for the entire NYUv2 dataset [3]. We use this new benchmark to quantitatively compare the performance of our method against the algorithms available in the literature (refer to Table 1). **(ii)** We perform a sensitivity analysis to gauge repeatability and robustness in the presence of varying amounts of scene rotation, noise and object misalignment in the scene (refer to Figure 2). **(iii)** We show how our method can be used in RGB-D SLAM and how it improves the performance of a popular SLAM method (refer to Table 2).

Table 1: Average angular error in degrees and runtime in seconds for 6 MF methods. Our method and ES outperform all other methods, with our method having a slight advantage in θ_y and θ_z. As for runtime, our method is significantly faster than its closest competitor.

Category	RGB		RGB-D			
Method	VP	VPGC	MPE	MMF	ES	Ours
θ_x	7.2°	21.4°	26.3°	8.1°	2.3°	2.3°
θ_y	9.7°	35.7°	18.1°	19.6°	5.6°	4.7°
θ_z	24.1°	20.5°	18.2°	9.8°	2.9°	2.8°
Runtime (s)	17.2	9.6	2.8	0.1	21.4	0.9

Table 2: Columns 1 - 3: Performance of RGB-D SLAM method [2] with no pre-rotation, ES [3], and our pre-rotation. Columns 4 - 5: SLAM performance (with only translation computed), with ES and our rotation as input. Our method improves runtime without compromising on accuracy, since our estimated rotations are a very good prior to the final rotations estimated by SLAM.

Method	SLAM R+T			SLAM T	
Pre-rotation	None	ES	Ours	ES	Ours
Trans RMSE	0.103m	0.113m	0.107m	0.125m	0.108m
Rot RMSE	3.41°	3.39°	3.37°	22.3°	4.61°
Runtime	145s	141s	112s	141s	112s

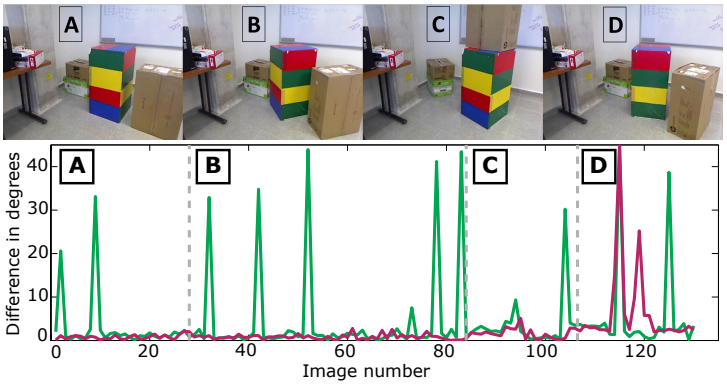

Figure 2: **First row:** Images from the new dataset for robust MF evaluation. There are four different categories of scenes with increasing difficulty. **(A):** all objects are aligned with the scene, **(B):** some objects are aligned, **(C):** all objects are equally unaligned, **(D):** all objects are unaligned at different orientations. **Second row:** Estimation error around the y-axis. Red curve: Ours. Green curve: ES [3]. Our method performs consistently better in **A-C**. Category **D** consists of more difficult images and we see the downgrade in performance on both methods.

[1] James M. Coughlan and Alan L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *ICCV*, pages 941–947, 1999.

[2] Felix Endres, Jürgen Hess, Nikolas Engelhard, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. An evaluation of the rgb-d slam system. In *ICRA*, pages 1691–1696, 2012.

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.