Robust Manhattan Frame Estimation from a Single RGB-D Image

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

Abstract

This paper proposes a new framework for estimating the Manhattan Frame (MF) of an indoor scene from a single RGB-D image. Our technique formulates this problem as the estimation of a rotation matrix that best aligns the normals of the captured scene to a canonical world axes. By introducing sparsity constraints, our method can simultaneously estimate the scene MF, the surfaces in the scene that are best aligned to one of three coordinate axes, and the outlier surfaces that do not align with any of the axes. To test our approach, we contribute a new set of annotations to determine ground truth MFs in each image of the popular NYUv2 dataset. We use this new benchmark to experimentally demonstrate that our method is more accurate, faster, more reliable and more robust than the methods used in the literature. We further motivate our technique by showing how it can be used to address the RGB-D SLAM problem in indoor scenes by incorporating it into and improving the performance of a popular RGB-D SLAM method.

1. Introduction

The representation of indoor scenes using the Manhattan world assumption [4] has been widely used in applications of computer vision and robotics. These applications take advantage of this assumption, by simplifying the representation of objects with respect to the whole 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 are interested in determining the MF of an indoor scene. Our motivation is that an accurate estimation of the MF can assist in addressing a variety of 3D problems, such as RGB-D SLAM and 3D understanding of objects and their spatial relations, as well as, speedup the pipeline of indoor scene understanding methods.

In this work, we propose an accurate, fast, reliable, and robust method to estimate the MF of an indoor scene using a single RGB-D image. In order to evaluate the properties of our method, we introduce a new evaluation benchmark that comprises ground truth MFs (logged as rotation matrices) for the NYUv2 dataset [19]. Using this new benchmark, we compare our method against several popular MF algorithms in the literature, and show that our approach outperforms state-of-the-art techniques in terms of accuracy and speed. Furthermore, we perform controlled tests to evaluate the repeatability and robustness of our method in a variety of scenarios. We also show how our method can be used in addressing the RGB-D SLAM problem. We do this by incorporating it into and improving the performance of a popular RGB-D SLAM method. Finally, we present a framework of assessment for any MF estimation method, and make our code and results publicly available for future use. To the best of our knowledge, no prior work provides such detailed evaluation and benchmarking for the task of MF estimation.

Figure 1 presents the pipeline of our approach. Given an RGB-D image pair, we first compute normals at every pixel. Given the point normals and the Manhattan assumption on the scene, we formulate an optimization problem to estimate a rotation matrix that transforms *most* of the normals to be aligned with a coordinate direction (i.e. either the x, y, or z axis). This process is equivalent to finding a rotation that converts the original normals into the sparsest set of directions in 3D. We use this sparsity intuition in our solution, and show how it yields accurate results in a fast and robust manner. With this solution, we are also able to estimate outliers in the scene, i.e., points with normals that are not aligned with one of the principal directions of the scene. We present the details of our method in Section 3 and a comparative evaluation in Section 4.

2. Related Work

We now review the most common approaches in the literature that estimate MFs in indoor scenes. Here, we note that most previous work usually presents comprehensive approaches that focus on a higher-level end task (e.g. indoor scene semantic labeling of objects, free-space and support surface estimation, etc.), however, all of these methods formulate a module to estimate the MF of the scene, and use

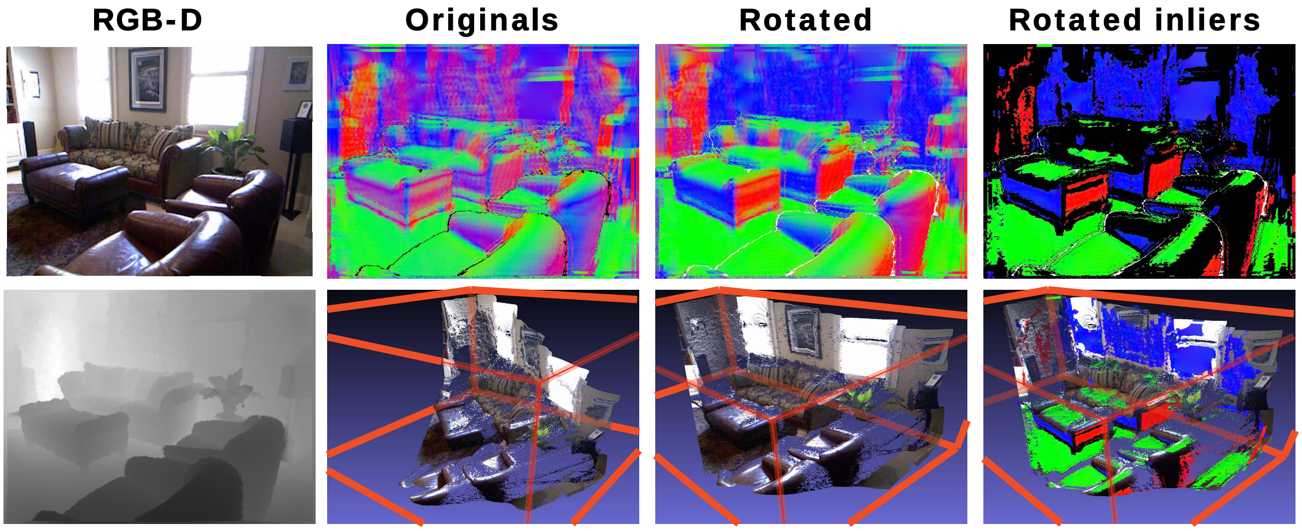

Figure 1: Overview of our method. **RGB-D** Top: Original RGB image. Bottom: Inpainted depth image from NYUv2 dataset. **Originals** Top: Original normals. Bottom: Original 3D point cloud. **Rotated** Top: Normals after alignment with our method. Bottom: Aligned 3D point cloud, where the wall, sofas, 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. All figures are best viewed in color and are included in the **supplementary material**.

it as an essential part of their process. We categorize the relevant MF estimation techniques based on whether their inputs are RGB or RGB-D images.

RGB Methods. Single image RGB methods rely on perspective cues such as vanishing points and lines, and associate each line to one of 3 main vanishing points. In the work by Furukawa *et al.* [6], the authors plot normalized coordinates of the line segments in a hemisphere, and find three orthogonal directions as the 3 main clusters in the hemisphere plot. Hedau *et al.* [9] propose an algorithm to estimate the MF parameters, by using structured learning to predict the best solution based on global perspective cues. Lee *et al.* [12] estimate MF by creating hypotheses based on vanishing points and line segments, and geometric reasoning on how corners and wall intersections are present in real scenes. A similar approach is presented by Del Pero *et al.* [15], with added geometric assumptions for robustness. Schwing *et al.* [17, 18] present the same framework as [9] but with a more efficient structured learning approach. The works in [10, 11, 16] are extensions of [9, 12, 17] respectively, where the authors use enhanced hypotheses to simultaneously estimate the MF of the scene and its objects. Del Pero *et al.* [14] propose a statistical model to compute MF that integrates camera parameters, room layout, and object hypotheses. Finally, the work of Chao *et al.* [1, 3] follows the procedure of MF estimation as [9], but enhances vanishing point estimation by considering humans in the scene.

RGB-D Methods. More recent approaches use RGB-D images to obtain better MF estimates. Silberman *et al.* [19] propose a method where 3D perspective cues are complemented by point normals to estimate the principal directions of a scene. They are motivated by the fact that most surfaces in a Manhattan environment are aligned with one of the 3 principal directions. This approach exhaustively searches among a set of candidates and chooses the best one according to a scoring heuristic. This incurs a substantial computational burden. The MF results are used to do object segmentation and support relation estimation. Taylor *et al.* [21] use depth information to estimate room geometry and layout given that the scene has a large number of visible walls and floor. Zhang *et al.* [25] present an approach similar to [16], where scene objects are modeled as clutter and are estimated with the help of depth information. Gupta *et al.* [8] address the problem of semantic segmentation in indoor scenes. The authors do not compute a complete MF of the scene, instead, they only estimate a gravity vector (y-axis) by taking candidates from point normal estimated from the depth image. Recently, Straub *et al.* [20] propose a method to estimate a Mixture of Manhattan Frame (MMF) model. In this method, the authors work on the idea that Manhattan scenes can be represented by multiple frames, and propose a technique to estimate all these frames. This idea relaxes the constraint that all planar surfaces are aligned to one set of orthogonal directions, and allows instead for a mixture of sets; however, our experiments show that this method lacks accuracy when computing the MF of an indoor scene.

This brief review shows the diversity of applications that estimate the MF of an indoor scene. Most RGB methods rely on structured learning approaches, rendering their estimation slow. RGB-D methods have the advantage of 3D information and are expected to perform better. Unfortunately, the accuracy and robustness of all these methods is unclear from the literature. To the best of our knowledge, there is no available benchmark evaluation in the literature to assess indoor MF estimation methods.

Contributions. The contributions of this work are threefold. **(i)** We present a robust and efficient algorithm to estimate the MF of an indoor scene from a single RGB-D image, using concepts of sparsity and convex optimization. **(ii)** We present an extensive comparison of most MF estimation techniques, by contributing ground truth MFs for the NYUv2 dataset [19] and comparing these methods against it. We also perform sensitivity analysis to gauge repeatability and robustness. We make all our data and code publicly available for use with future methods **(iii)** We show how our MF estimation method can be used to enhance the performance of RGB-D SLAM.

3. Methodology

In this section, we give a detailed description of the RGB-D MF estimation problem and our proposed method for robust MF estimation.

3.1. Problem Statement

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, i.e. where most surfaces are oriented along the dominant directions. In Manhattan scenes, these directions form an orthonormal system that is assumed without loss of generality to be right-handed. Therefore, estimating the MF becomes equivalent to computing the *best* 3D rotation matrix $\mathbf{R} \in \mathrm{SO}(3)$ that transforms surface normals (and line directions if available) in the scene to the three unit directions or their reflections about the center (i.e. $[\pm 1, 0, 0]^T$, $[0, \pm 1, 0]^T$, and $[0, 0, \pm 1]^T$). Note that $\mathrm{SO}(3)$ defines the rotation group such that each element $\mathbf{R} \in \mathrm{SO}(3)$ satisfies the following properties: $\mathbf{R}^{-1} = \mathbf{R}^T$ and $\det(\mathbf{R}) = +1$. In fact, the rows of $\mathbf{R}$ define the dominant directions of the scene in the original coordinate system. These rows are exactly the vectors $\mathbf{v}1$, $\mathbf{v}2$, and $\mathbf{v}3$ estimated in [19].

In this work, we determine the MF of a Manhattan scene using a single RGB-D pair of images $\mathbf{I}_C$ and $\mathbf{I}_D$ that are generated by a calibrated RGB-D sensor (e.g. KINECT). Similar to previous work [19], we use conventional methods (e.g. local plane fitting using RANSAC) to compute 3D surface normals. We ensure that surface normals are oriented in such a way that their surfaces are visible by the camera, whose image plane and optical center define the orthonormal coordinate system used to represent these normals. We concatenate all the unit-norm normals in a single matrix $\mathbf{N} \in \mathbb{R}^{3\times m}$. Due to sensor noise/limitations, errors and noise in surface normal (e.g. at depth discontinuities), and the presence of non-Manhattan outliers in the scene (e.g. planar surfaces that do not align with the scene's dominant directions), the problem of reliable scene MF estimation is challenging. To showcase the difficulty of the problem, we plot vectors in $\mathbf{N}$ of an image from the NYUv2 dataset [19] (refer to Figure 2). In an ideal Manhattan scene, these vectors should cluster around at most 5 points on the unit sphere corresponding to the five faces of a box, where the sixth face is not visible to the camera. Clearly, the columns of $\mathbf{N}$ shown in Figure 2 do not follow this ideal setup. Although there are regions of the sphere where there is a reasonable density of points, applying conventional clustering methods to localize the dominant directions of a scene would fail in general due to the significant amount of noise, the lack of cluster compactness, and the significant number of outliers.

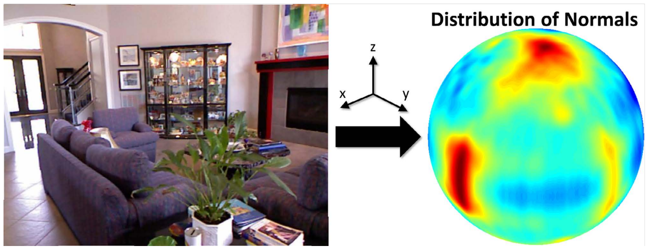

Figure 2: Distribution of normals (on the unit sphere) from a sample image in the NYUv2 dataset. In an ideal Manhattan scene with objects aligned with the main directions, point normals should cluster in 5 locations, at most. However, due to noise and outliers, point normals are distributed across much of the unit sphere. This property of point normals in real-world scenes renders MF estimation quite challenging.

3.2. Proposed Solution: Robust Manhattan Frame Estimation (RMFE)

As stated before, estimating the MF of an indoor scene is equivalent to computing a rotation $\mathbf{R}$ that transforms surface normals into the three unit directions or their reflections about the center. Therefore, applying $\mathbf{R}$ to matrix $\mathbf{N}$ 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), we incorporate the above observation to formulate the MF Estimation (MFE) problem as in Eq (1). Here, the first term penalizes reconstruction error, while the second term serves as a sparse reg-

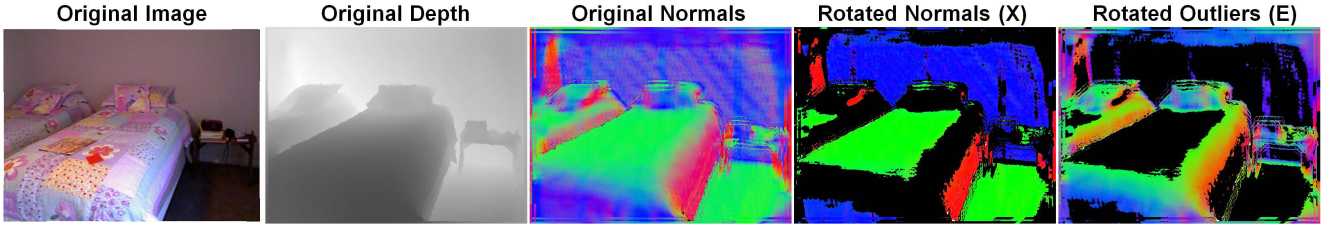

Figure 3: Result of solving RMFE on a sample RGB-D image pair. Note that after the optimal rotation $\mathbf{R}$ is applied, the original normals are transformed so they are aligned to the dominant directions in the scene: x-axis (normal to the side wall; *red*), y-axis (normal to the floor; *green*), and z-axis (normal to the front wall; *blue*). The columns in $\mathbf{X}$ are colored coded to indicate these directions. Pixels in black designate outliers (non-zero columns of $\mathbf{E}$) and their corresponding normals are shown on the right. Clearly, the edges and corners of the bed as well as the pillow are detected as outliers. Only the sparsest rotated normals are included in $\mathbf{X}$.

ularizer. Note that any $\ell_{p,q}$ matrix norm of $\mathbf{A} \in \mathbb{R}^{n\times m}$ is defined as: $\|\mathbf{A}\|_{p,q} = \left(\sum_{i=1}^{n} \|\mathbf{A}_i\|_p^q\right)^{\frac{1}{q}}$, where $\|\mathbf{A}_i\|_p$ is the ℓ_p norm of the i^{th} row of $\mathbf{A}$. Thus, the regularizer is simply the sum of the ℓ_1 norms of the columns in $\mathbf{X}$. This sparsity inducing ℓ_1 regularizer has been used successfully in other applications (e.g. face recognition [23], tracking [13], and image classification [24].) to provide robustness against noise and overfitting. For the MFE problem, this sparsity term is a crucial and characteristic prior on the dominant directions of the scene.

$$(MFE): \quad \min_{\mathbf{R}\in SO(3),\mathbf{X}} \ \frac{1}{2}\|\mathbf{RN}-\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_{1,1} \tag{1}$$

Solving Eq (1). Since SO(3) is not a convex set, then Eq (1) is in general non-convex. Although a global minimum is not guaranteed, a local minimum is achievable via alternating optimization, which iterates between updating the current solution for one of the two variables ($\mathbf{X}$ and $\mathbf{R}$) while keeping the other fixed. Given the current estimate of $\mathbf{R}$, the current estimate of $\mathbf{X}$ is updated in closed form according to the identity in Eq (2), where $\mathcal{S}_\lambda(\mathbf{A}_{ij}) = \text{sign}(\mathbf{A}_{ij})\max(0, |\mathbf{A}_{ij}| - \lambda)$ is the well-known soft-thresholding operator. Following [22], the current estimate of $\mathbf{R}$ is updated according to the identity in Eq (3). If $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors of $\mathbf{XN}^T$, then $\mathcal{K}(\mathbf{N},\mathbf{X}) = \mathbf{U}\left[\text{diag}(1,1,\text{sign}(\det(\mathbf{XN}^T)))\right]\mathbf{V}^T$.

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} \ \frac{1}{2}\|\mathbf{X}-\mathbf{RN}\|_F^2 + \lambda\|\mathbf{X}\|_{1,1} = \mathcal{S}_\lambda(\mathbf{RN}) \tag{2}$$

$$\mathbf{R}^* = \arg\min_{\mathbf{R}\in SO(3)} \ \|\mathbf{RN}-\mathbf{X}\|_F^2 = \mathcal{K}(\mathbf{N},\mathbf{X}) \tag{3}$$

Initializing variables to reasonable estimates is important in alternating optimization, so as to avoid undesirable local minima and expedite convergence. In our case, we initialize $\mathbf{R}$ to identity, thus, testing the hypothesis that the depth sensor is oriented along the dominant directions first. This initial hypothesis is valid, since most images are taken from reasonable viewpoints (e.g. the floor and/or ceiling are similarly oriented with the second dominant direction).

Handling Outliers. Since the Frobenius norm (equivalently the ℓ_2 norm of a vector) in the reconstruction error term tends to be sensitive to outliers, the rotation estimated by Eq (1) might be affected by non-Manhattan surface normals and line directions. This occurs, for example, when a boxy object in the scene is *not* aligned with at least two of the dominant directions or when this object is not boxy in the case of a ball, vase, or cup. Most MF estimation methods enforce the Manhattan assumption on the whole scene, so they tend to be sensitive to outliers (as we show empirically in Section 4.3). To handle non-Manhattan outliers that do not yield a 1-sparse column in $\mathbf{X}$, we refine Eq (1) to explicitly represent the error as a column-sparse matrix $\mathbf{E} \in \mathbb{R}^{3\times m}$. As such, we assume that only a sparse number of outliers exist in the scene and they are identified as the non-zero columns of $\mathbf{E}$. This robust refinement of the MFE problem (denoted as RMFE) is formulated in Eq (4), where the $\ell_{2,1}$ norm is used to encourage column sparsity on $\mathbf{E}$. In essence, Eq (4) is the same as Eq (1) but with a different robust penalty for the reconstruction error.

$$(RMFE): \quad \min_{\mathbf{R}\in SO(3),\mathbf{X},\mathbf{E}} \ \|\mathbf{E}^T\|_{2,1} + \lambda\|\mathbf{X}\|_{1,1} \tag{4}$$
$$\text{subject to: } \ \mathbf{RN} = \mathbf{X} + \mathbf{E}$$

Adding an explicit error term to handle sparse outlier error has been successfully used in other problems, such as robust face recognition [23] and image registration [7]. Moreover, the $\ell_{2,1}$ norm has been used extensively in coupling different tasks in a multi-task learning framework [2, 27, 26].

Solving Eq (4). Similar to the MFE problem, we use alternating optimization to reach a desirable local minimum. Eq (3) is used to update the current estimate of $\mathbf{R}$. Updating $\mathbf{X}$ and $\mathbf{E}$ is done jointly by applying the conventional Inexact Augmented Lagrange Multiplier (IALM) method on the resulting convex but non-smooth optimization problem. IALM is an iterative method that augments the traditional Lagrangian function with quadratic penalty terms and has been shown to have attractive convergence properties. The iterative update of $\mathbf{X}$ makes use of the identity in Eq (2). Using the result in [2], the i^{th} row of $\mathbf{E}$ is updated using the identity in Eq (5). The tradeoff parameter $\tilde{\lambda}$ is a function of the sparsity coefficient λ and the increasing IALM parameter μ. Due to space limitations, we leave the optimization details for the **supplementary material**.

$$\begin{aligned} \mathbf{E}_i &= \arg\min_{\mathbf{Y}_i} \frac{1}{2}\|\mathbf{Y}_i - \mathbf{A}_i\|_2^2 + \tilde{\lambda}\|\mathbf{Y}_i\|_2 \\ &= \max\left(0, 1 - \frac{\tilde{\lambda}}{\|\mathbf{A}_i\|_2}\right)\mathbf{A}_i \end{aligned} \tag{5}$$

Since solving RMFE is computationally more expensive than MFE, we initialize the rotation matrix in RMFE using the final estimate of $\mathbf{R}$ in MFE. In other words, we assume that LE quickly converges to a reasonable rotation and RMFE simply refines this rotation.

Implementation Details. In all our experiments, we select $\lambda = 0.3$, which leads to an empirically viable tradeoff between sparsity and error sensitivity. We use a conventional stopping criterion (i.e. relative change in current estimate) to determine when alternating optimization has converged. A tolerance of 10^{-4} is used. When $m = 480 \times 640$ (size of a KINECT frame), RMFE converges on average in 0.9 seconds on a 3GHz workstation running MATLAB. We compute normals as a preprocessing step, which takes 0.1 seconds. In our experiments, we do *not* compute line directions (as in other methods), since they do not improve MF accuracy. In Figure 3, we plot RMFE results for a sample RGB-D image from the NYUv2 dataset [19]. Here, we mention that a box layout can be estimated from the inliers of the RMFE estimate (i.e. the non-zero columns of $\mathbf{X}$). To do this, we center the box at the average 3D position of the 3D point cloud and orient it along the estimated dominant directions. The extent of this box is determined by fitting a minimum volume box, which encloses all points in the point cloud having inlier normals in $\mathbf{X}$. In this paper, we depict the box layout using red borders as in Figure 1.

4. Experimental Results

We assess our MF estimation methodology under three complimentary perspectives. Unfortunately, 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. First, we create a new benchmark framework for evaluating algorithms that estimate MF for indoor scenes from RGB-D images by generating MF ground truth for the entire NYUv2 dataset[19]. This dataset is comprised of 1449 images of indoor scenes with a variety of difficulty in terms of clutter and noise. Our new annotations include a ground truth rotation matrix for every image in the dataset. We use this new benchmark to quantitatively compare the performance of our method against the algorithms available in the literature. Second, 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. Finally, we show how our method can be used in RGB-D SLAM and how it improves a popular RGB-D SLAM method

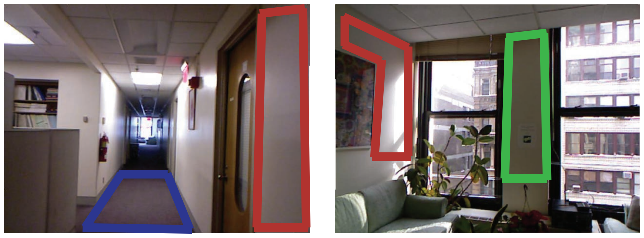

Figure 4: Images from the NYUv2 dataset and their corresponding ground truth annotation. In the left image, the floor is visible, so we annotate it first. Since the image on the right has no visible floor, we choose the two walls and assign them to the x or z-direction, ensuring the resulting y-direction to be aligned with the floor.

4.1. An Indoor Scene MF Estimation Benchmark

To compare and evaluate MF estimation in indoor scenes, we introduce a new evaluation benchmark that consists of MF ground truth for the NYUv2 dataset. Given an RGB-D image pair from this dataset, our annotation consists of the three main directions of the captured scene. Since all the scenes contain planes aligned with the principal directions, we select two regions corresponding to planar areas that are orthogonal in 3D. Using the depth image and KINECT's calibration matrix, we estimate the 3D points of the selected regions, fit a plane to each, and compute their normals. We associate every computed normal to one of the main directions of the scene. We only select two planes, as the third one can be inferred as a cross product to guarantee orthonormality.

To be consistent in all the scenes, we align the floor direction with the y-axis. To do this, we select an image region aligned with the floor whenever possible; if no such regions exist, we make sure the annotation of x and z preserve the desired floor direction. An example of this annotation is presented in Figure 4. On the left, one of the selected regions corresponds to the actual floor (blue selection) and we therefore assign its normal to the positive y-direction. The second annotation can be assigned to either x or z. On the right, there are no visible surfaces aligned with the floor, so we select regions from the two visible walls. To maintain consistency in the y-direction, we choose the normal of the red portion to point in the $+x$-direction and the $+z$-direction is assigned to the second region.

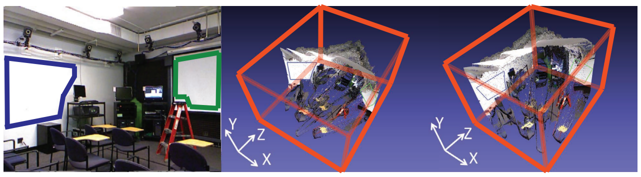

Figure 5: Left: RGB image with user annotation. Middle: Original point cloud with user annotated region outlines (axis-aligned bounding box in red). Right: Rotated point cloud using ground truth; we see how both annotated surfaces align with the sides of the axis aligned box.

Annotating an image results in two normalized directions, with the third being their cross product. Next, we make sure that the normals are orthogonal, since this might not always be the case due to sensor noise or small plane fitting errors. Given three normal vectors $\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$, where $\mathbf{v}_3 = \mathbf{v}_1 \times \mathbf{v}_2$, we re-compute $\mathbf{v}_1 = \mathbf{v}_2 \times \mathbf{v}_3$, this guarantees the 3 vectors to form an orthonormal set. Finally, we create a rotation matrix $\mathbf{R} = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]^\top$. Note that applying this rotation aligns the scene, such that the normal to the floor points in the $+y$-direction. Figure 5 shows an example of a scene before and after ground truth alignment, with an axis-aligned bounding box around each two point clouds.

With this ground truth, our benchmark evaluates the MF estimation error of any MF algorithm. We focus our comparison on representative techniques for MF estimation. For instance, [15, 14] use the method introduced in [9], so we use [9] as a proxy for evaluating all of these methods. We are also constrained by the lack of publicly available data and code with some of the methods. This left us with 5 techniques to compare against [9, 12, 19, 21, 20]. Some of the methods use RGB images as input and therefore do not provide a 3D MF estimate; however, they output 3 clusters of 2D line segments each corresponding to one of the principal directions of the scene. In order to adapt them for a fair comparison, we compute the equivalent 3D line segments using the depth image. For every clustered set of 3D line segments, we compute an average direction by only considering parallel segments. These directions can be used to create rotation matrices for comparison.

4.2. Evaluation of MF Estimation

For each RGB-D image, we have a set of rotation matrices $\{\mathbf{R}_i\}_{i=1}^5$ corresponding to the MF methods to be compared against our rotation matrix $\mathbf{R}_{our}$. Using the ground truth rotation $\mathbf{R}_{gt}$ as baseline, we compute the error in rotation angles $\Theta = [\theta_x, \theta_y, \theta_z]$ between each $\mathbf{R}_i$ and $\mathbf{R}_{gt}$.

Since the only alignment restriction is to bring the floor normal to the $+y$-axis, different alignments of the other two directions can still be valid even if they do not match the ground truth. For example, if a wall is aligned with the ground truth $+x$-axis, aligning it with the $+z$-axis would still create a valid rotation, given that the floor still points in the $+y$-direction. In terms of rotation angles, the difference of θ_x and θ_z to ground truth corresponds to changes in the floor's orientation, while differences in θ_y represent different orientations of the walls only. To allow for such diversity, we create a set $\mathcal{R}_{GT}$ containing several versions of $\mathbf{R}_{gt}$ where θ_y has been rotated by 0^o, $\pm 90^o$, and $\pm 180^o$. These rotations will create all possible alignments for a given scene, without distorting the aligned floor orientation. We evaluate all 6 rotations by comparing each one with its closest ground truth rotation.

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.

Ground Truth Comparison and Runtime

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 1 summarizes the comparison-to-ground-truth results for the RGB-based and RGB-D based techniques. The RGB methods are based on Vanishing Points (VP) [9], and Vanishing Points with Geometric Context (VPGC) [12]. The RGB-D methods are based on Exhaustive Search (ES) [19], Main Plane Estimation (MPE) [21], and Mixture of Manhattan Frames (MMF) [20]. Since the MMF method computes a mixture of frames, we compare against all the estimated frames and report the best results. Furthermore, we analyze the angular errors around each of the axes independently and plot error histograms in Figure 6. The results show how our RMFE method results in small errors when compared to ground truth. As expected, the RGB methods underperform due to the limitations that arise from fusing 3D information only at a later stage. Among the RGB-D methods, MPE is the least competitive, since it assumes that a large portion of the floor is already visible, which is not the case in many of the NYUv2 images, and it does not account well for scene clutter and outliers. The results also show MMF being significantly less accurate than ES and our method. ES performs comparably well to our technique, but its runtime is 20 times slower. We present a comparative runtime study for all methods in Table 1.

Since our method has a slight advantage over ES on NVUv2, we conduct two additional experiments to further assess their performance. The first one looks at the repeatability of the methods with varying initializations of the input scene, while the second gauges the robustness of both methods under different scene configurations. We discuss the robustness experiment in this section and provide details of the repeatability analysis in the **supplementary material.**

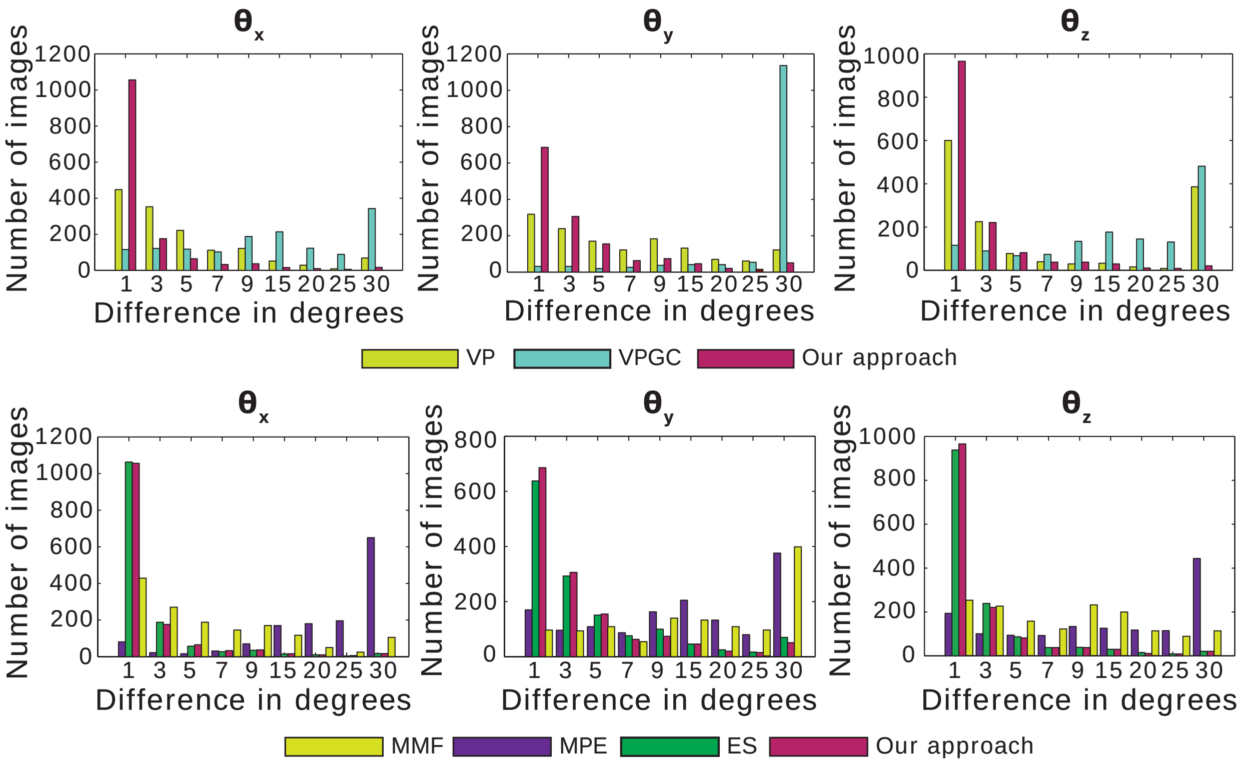

Figure 6: **First row:** Angular error histograms of our method compared to RGB based algorithms. Left: θ_x. Middle: θ_y. Right θ_z. Yellow corresponds to VP [9], Cyan to VPGC [11], and Magenta to ours. **Second row:** Angular error histograms of our method compared to RGB-D based algorithms. Left: θ_x. Middle: θ_y. Right θ_z. Yellow corresponds to MMF [20], Purple to MPE [21], Green to ES [19], and Magenta to ours. MPE performs poorly due to its tight restrictions on scene type and floor visibility. ES performs comparably to ours; we have a slight advantage with more values in the small error bins. Also, our RMFE method is faster, more repeatable and robust, as shown later.

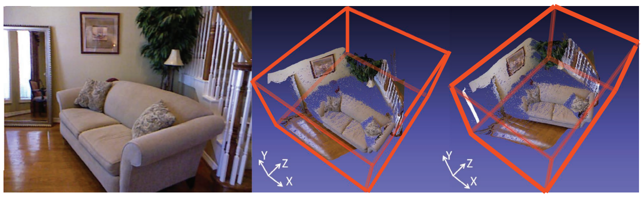

Figure 7: An example from the NYUv2 dataset with a misaligned object. Left: Original RGB Image. Middle: Point cloud aligned using ES. Right: Point cloud aligned using our method. Since the sofa is diagonally aligned with the walls, it causes ES to align the scene to the sofa and not the actual walls of the room. Our method is more robust to these outliers and aligns the left wall to its proper position.

4.3. Robustness to Outliers

Here, we are interested in evaluating the robustness of ES and our RMFE in non-ideal Manhattan scenes, i.e. scenes with non-Manhattan elements in them. Figure 7 is an NYUv2 example of such scenes. We see how ES performs poorly in such a scenario. The comparable results of ES with RMFE on NYUv2 is due to the fact that there are very few scenes like Figure 7 in the NYUv2 dataset. To gauge the robustness of MF estimation, we compile a new set of RGB-D images captured in a room with a varying number of boxes at varying orientations. Examples are presented in the first row of Figure 8. This set provides 4 categories of images: **(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. A total of 131 images were recorded. The difficulty of these scenes, in terms of MF accuracy, increases from **A-D**. Figure 8 shows comparative results w.r.t. ground truth on the new dataset. We notice that our method incurs much less angular error than ES over nearly the entire dataset. It is consistently better in every image of categories **A-C**. We expect a degradation in performance on images of category **D**, since the amount of unaligned normals is substantially high. The ES method is not as reliable, since it incurs huge errors even at the 'easier" categories **A-B**, with the angular error reaching as high as 45° in some cases. It is obvious that our RMFE method is substantially more robust than

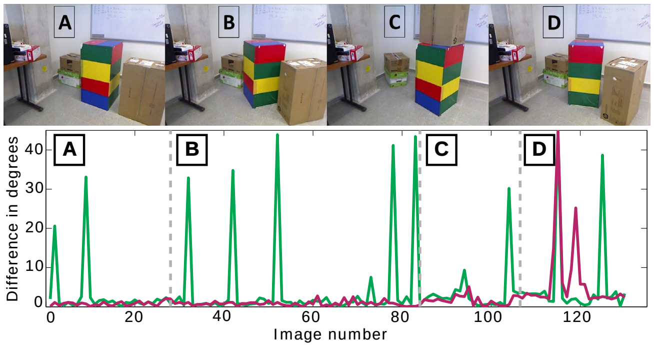

Figure 8: **First row:** Images from the new dataset for robustness assessment. There are four different categories of scenes with increasing difficulty. **(A):** all objects are aligned with the scene, **(B):** some objects are aligned, **(C):** all objects are equally unaligned, **(D):** all objects are unaligned at different orientations. **Second row:** Estimation error around the y-axis. Red curve: Ours. Green curve: ES. 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.

ES to misalignment of scene objects, since it is designed to handle non-Manhattan outliers.

4.4. Application: RGB-D SLAM

In this section, we show how MF estimation can be used to improve RGB-D SLAM methods. The aim of SLAM is to map a scene by imaging it at different locations using an agent (usually a robot), while keeping track of the agent's location. This is usually achieved by mounting an imaging sensor on the robot and using the frame-to-frame information to estimate the structure of the scene and the motion of the agent. This motion is generally expressed as a frame-to-frame rotation and translation. When using an RGB-D sensor, SLAM can better estimate motion with the help of 3D information provided by the depth frames. The performance of SLAM depends on the misalignment of consecutive frames. If the robot is moving inside an indoor scene, we can use RMFE to align every depth image before applying SLAM. This means that the input depth frames of SLAM will be aligned to a big extent, and most computational effort will go towards estimating the frame-to-frame translations. To illustrate this advantage, we consider a popular RGB-D SLAM method by Endres *et al.* [5], where the authors evaluate using different indoor scenes. Performance is evaluated by runtime, translation Root Mean Squared Error (RMSE), and rotation RMSE. We choose a sequence from [5] that covers an entire scene, with a total of 745 frames. We perform two different experiments. (i) We pre-rotate every depth frame using ES and RMFE and then run SLAM. (ii) We pre-rotate every depth frame using ES and RMFE and run a simplified SLAM where only the translations are computed. We present the results of these experiments in Table 2. Pre-rotating the scenes does not improve the accuracy of the SLAM method, however, our method decreases runtime significantly (by 23%), without degrading performance much. For experiment (ii), we modify the motion estimation module of SLAM to compute only the frame-to-frame translations. Given a proper initial alignment of all the frames, we expect this modified SLAM version to still perform competitively with its complete counterpart, since all the frames are from the same scene and therefore share the same MF. We see how in this experiment our method yields similar accuracy, while ES performs poorly. This is mainly due to the fact that RMFE robustly and consistently estimates the MF of each frame, while ES fails in several cases.

Table 2: Evaluation of RGB-D SLAM. Columns 1 - 3: Results of method [5] with no pre-rotation, ES, and our pre-rotation. Columns 4 - 5: Results of modified SLAM to compute translation only, 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.

Performance of RGB-D 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	$145s$	$141s$	$112s$	$141s$	$112s$

The results presented in this section suggest that our MF estimation method can be used as a strong initialization for RGB-D SLAM methods on indoor scenes, as it yields similar accuracy, but decreases runtime significantly.

5. Conclusion

In this paper, we present a new formulation for estimating the Manhattan frame of indoor scenes from an RGB-D image. The main advantages of our technique are its robustness, speed, reliability and accuracy, which we evaluate experimentally with our contributed evaluation benchmark. We show how our method can be used in an RGB-D SLAM framework, where it enhances its performance. We contribute our code to the community, encouraging interested authors to incorporate it as a preprocessing module for indoor scene understanding and RGB-D SLAM.

For future work, we are interested in extending our proposed formulation to multiple Manhattan frames in the same scene. We also would like to incorporate our MF estimation into an end-to-end scene understanding pipeline.

Acknowledgments: This work was supported by competitive research funding from King Abdullah University of Science and Technology (KAUST). JCN is supported by a Microsoft Research Faculty Fellowship.

References

[1] Y.-W. Chao, W. Choi, C. Pantofaru, and S. Savarese. Layout estimation of highly cluttered indoor scenes using geometric and semantic cues. In *Proceedings of the International Conference on Image Analysis and Processing*, pages 489–499, 2013.

[2] X. Chen and W. P. J. T. K. J. G. Carbonell. Accelerated gradient method for multi-task sparse learning problem. In *International Conference on Data Mining*, pages 746–751, 2009.

[3] W. Choi, Y. W. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3d geometric phrases. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 33–40, 2013.

[4] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 941–947, 1999.

[5] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard. An evaluation of the rgb-d slam system. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1691–1696, 2012.

[6] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1429, 2009.

[7] B. Ghanem, T. Zhang, and N. Ahuja. Robust video registration applied to field-sports video analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[8] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013.

[9] V. Hedau, D. Hoiem, and D. A. Forsyth. Recovering the spatial layout of cluttered rooms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1849–1856, 2009.

[10] V. Hedau, D. Hoiem, and D. A. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *Proceedings of European Conference on Computer Vision*, pages 224–237, 2010.

[11] D. C. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *Advances in Neural Information Processing Systems*, pages 1288–1296, 2010.

[12] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2136–2143, 2009.

[13] X. Mei and H. Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2259–2272, 2011.

[14] L. D. Pero, J. Bowdish, D. Fried, B. Kermgard, E. Hartley, and K. Barnard. Bayesian geometric modeling of indoor scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[15] L. D. Pero, J. Guan, E. Brau, J. Schlecht, and K. Barnard. Sampling bedrooms. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[16] A. G. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[17] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction for 3d indoor scene understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[18] A. G. Schwing and R. Urtasun. Efficient exact inference for 3d indoor scene understanding. In *Proceedings of European Conference on Computer Vision*, 2012.

[19] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of European Conference on Computer Vision*, 2012.

[20] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher. A mixture of manhattan frames: Beyond the manhattan world. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3770–3777, 2014.

[21] C. J. Taylor and A. Cowley. Parsing indoor scenes using rgb-d imagery. In *Robotics: Science and Systems*, 2012.

[22] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991.

[23] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[24] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[25] J. Zhang, C. Kan, A. G. Schwing, and R. Urtasun. Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[26] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via multi-task sparse learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2042–2049, 2012.

[27] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via structured multi-task sparse learning. *International Journal of Computer Vision*, 101(2):367–383, 2013.