A Mixture of Manhattan Frames: Beyond the Manhattan World

Julian Straub Guy Rosman Oren Freifeld John J. Leonard John W. Fisher III
Massachusetts Institute of Technology
{jstraub,rosman,freifeld,jleonard,fisher}@csail.mit.edu

Abstract

Objects and structures within man-made environments typically exhibit a high degree of organization in the form of orthogonal and parallel planes. Traditional approaches to scene representation exploit this phenomenon via the somewhat restrictive assumption that every plane is perpendicular to one of the axes of a single coordinate system. Known as the Manhattan-World model, this assumption is widely used in computer vision and robotics. The complexity of many real-world scenes, however, necessitates a more flexible model. We propose a novel probabilistic model that describes the world as a mixture of Manhattan frames: each frame defines a different orthogonal coordinate system. This results in a more expressive model that still exploits the orthogonality constraints. We propose an adaptive Markov-Chain Monte-Carlo sampling algorithm with Metropolis-Hastings split/merge moves that utilizes the geometry of the unit sphere. We demonstrate the versatility of our Mixture-of-Manhattan-Frames model by describing complex scenes using depth images of indoor scenes as well as aerial-LiDAR measurements of an urban center. Additionally, we show that the model lends itself to focal-length calibration of depth cameras and to plane segmentation.

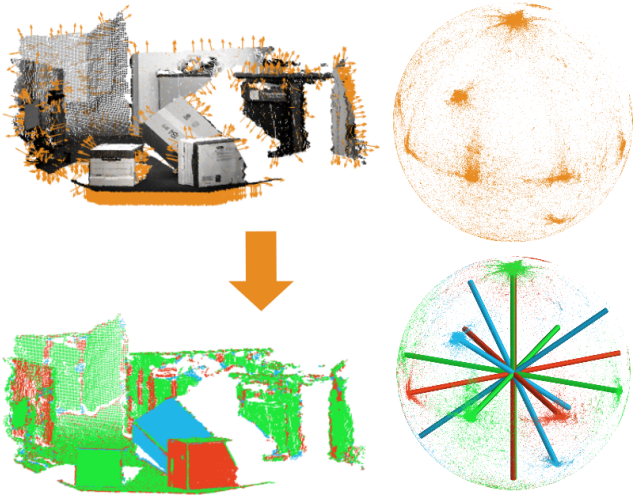

Figure 1: Surface normals in a man-made environment (top left), tend to form clusters on the unit sphere (top right) such that these clusters can be divided into subsets which we call Manhattan Frames (MF). Each MF explains clusters of normals aligned with the six signed axes of a common coordinate system. Our algorithm infers 3 distinct MFs, shown in different colors on the sphere (bottom right) and the scene (bottom left).

1. Introduction

Simplifying assumptions about the structure of the surroundings facilitate reasoning about complex environments. On a wide range of scales, from the layout of a city to structures such as buildings, furniture and many other objects, man-made environments lend themselves to a description in terms of parallel and orthogonal planes. This intuition is formalized as the Manhattan World (MW) assumption [10] which posits that most man-made structures may be approximated by planar surfaces that are parallel to one of the three principal planes of a common orthogonal coordinate system.

At a coarse level, this assumption holds for city layouts, most buildings, hallways, offices and other man-made environments. However, the strict Manhattan World assumption cannot represent many real-world scenes: a rotated desk, a half-open door, complex city layouts (as opposed to planned cities like Manhattan). While parts of the scene can be modeled as a MW, the entire scene cannot. This suggests a more flexible description of a scene as a mixture of *Manhattan Frames* (MF). Each Manhattan Frame in itself defines a Manhattan World of a specific orientation.

Our contributions include the formulation of a novel model for scene representation that describes scene surface normals as a mixture of orthogonally-coupled clusters on the unit sphere – we refer to this model as a *Mixture of Manhattan Frames* (MMF). We formulate a probabilistic Bayesian model for the MMF and propose a Gibbs-sampling-based inference algorithm. Using Metropolis-Hastings [18] split/merge proposals [26], the inference al-

gorithm adapts the number of MFs to that of the distribution of normals in the scene. Additionally, we propose an approximation to the posterior distribution over MF rotations that utilizes a gradient-based optimization of a robust cost function which exploits the geometry of both the unit sphere as well as the group of rotation matrices $\mathrm{SO}(3)$.

We demonstrate the advantages of our model in several applications including plane segmentation and single-shot RGB-D camera depth-focal-length calibration. Furthermore, we show its versatility by inferring MFs from not only depth images, but also large-scale aerial LiDAR data of an urban center.

2. Related Work

The connection between vanishing points (VPs) in images and 3D MW structures has been used to infer dense 3D structure from a single RGB image by Delage *et al*. [11] and from sets of images by Furukawa *et al*. [15]. This is done via projective geometry [17]. More specifically, Furukawa *et al*. employ a greedy algorithm for a single-MF extraction from normal estimates that works on a discretized sphere, while Neverova *et al*. [24] integrate RGB images with associated depth data from a Kinect camera to obtain a 2.5D representation of indoor scenes under the MW assumption.

The MW assumption was used to estimate orientations within man-made environments for the visually impaired by Coughlan *et al*. [10] and for robots by Bosse *et al*. [5]. In the application of Simultaneous Localization and Mapping (SLAM), the MW assumption has been used to impose constraints on the inferred map [25, 29].

While the MW model has also been useful in applications of RGB-camera calibration and metric rectification [6, 8], we are unaware of calibration schemes for depth sensors that exploit the MW or similar scene priors. Besides calibrating the IR camera of the depth sensors using standard monocular camera techniques [17], there is work by Herrera *et al*. [19] on the joint calibration of RGB and depth of an RGB-D sensor. Teichman *et al*. [32] follow a different approach for depth-camera intrinsic and distortion calibration within a SLAM framework.

A popular alternative to the MW model describes man-made structures by individual planes with no constraints on their relative normal directions. Such plane-based representations of 3D scenes have been used in scene segmentation [20], localization [31], optical flow [27], as well as other computer-vision applications. Triebel *et al*. [33] extract the main direction of planes in a scene using a hierarchical Expectation-Maximization (EM) approach. Using the Bayesian Information Criterion (BIC) they infer the number of main directions. The plane-based approach does not exploit important orthogonality relations between planes that are common in man-made structures. In such cases, independent location and orientation estimates of planes will be less robust, especially for planes that have few measurements or are subject to increased noise.

Due to the tight coupling of VP estimation and the MW assumption, the depth-based approach presented herein is similar in spirit to recent work on estimating multiple sets of VPs in images. The Atlanta World (AW) model of Schindler *et al*. [30] assumes that the world is composed of multiple MFs sharing the same z-axis. This facilitates inference from RGB images as they only have to estimate a single angle per MF as opposed to a full 3D rotation. Note, however, that common indoor scenes (*e.g*., see Fig. 4c) break the Atlanta World assumption. The approach by Antunes *et al*. [1] is more general than the AW model in that it does not assume a common axis for the Manhattan Frames. However, it is formulated in the image domain and does not estimate the rotation of the underlying 3D structure. Our MMF model can be seen as a generalization of both this model and the AW model.

Finally, our approach, which utilizes the unit sphere, is related to early work on VP estimation. There, one is interested in finding great circles and their intersections because these constitute VPs. While Barnard [2] discretizes the sphere to extract the VP, Collins *et al*. [9] formulate the VP inference as a parameter-estimation problem for the Bingham distribution [3]. To preclude discretization artifacts we opt to avoid an approach similar to Barnard. We eschew the Bingham distribution as the proposed probabilistic mixture model is straightforwardly incorporated within a Bayesian framework.

3. A Mixture of Manhattan Frames (MMF)

In this section, we explain our MMF framework, starting with its *mathematical representation*. Next, we define a *probabilistic model* for the MMF and conclude with a *statistical-inference* scheme for this model. Note that while our approach is probabilistic, the representation may still be used within a purely-deterministic approach. Similarly, though we suggest a specific probabilistic model as well as an effective inference method, one may adopt alternative probabilistic models and/or inference schemes for MMF.

3.1. MMF: Mathematical Representation

Let R be a 3D rotation matrix, which by construction defines an orthogonal coordinate system of $\mathbb{R}^3$. We define the MF associated with R as the 6 unit-length 3D vectors which coincide, up to a sign, with one of the columns of R. That is, the MF, denoted by M, can be written as a 3-by-6 matrix: $M = [R, -R]$, where we may regard the jth column $[M]_j$ of M as a *signed axis*, $j \in \{1, \ldots, 6\}$; see Fig. 2. If a 3D scene consists of only planar surfaces such that the set of their surface normals is contained in the set $\{[M]_j\}_{j=1}^{6}$, then M captures all possible orientations in the scene – the scene obeys the MW assumption. In our

MMF representation, however, scenes consist of K MFs, $\{M_1, \ldots, M_K\}$ which jointly define $6K$ signed axes. Note that for $K = 1$, the MMF coincides with the MW.

The MMF representation is aimed at describing surface normals. In practice, as is common in many 3D processing pipelines (*e.g.*, in surface fairing or reconstruction [21, 22]), the observed unit normals are estimated from noisy measurements (in our experiments, these are depth images or LiDAR data). The unit normals live on S^2 (the unit sphere in $\mathbb{R}^3$), a 2D manifold whose geometry is well understood.

Specifically, let $q_i \in S^2$ denote the i-th observed normal. Each q_i has two levels of association. The first, $c_i \in \{1, \ldots, K\}$, assigns q_i to a specific MF. The second, $z_i \in \{1, \ldots, 6\}$, assigns q_i to a specific signed axis within the MF M_{c_i}. We let $[M_{c_i}]_{z_i}$ denote the z_i-th column of M_{c_i}; *i.e.*, $[M_{c_i}]_{z_i}$ is the signed axis associated with q_i. In real observed data, q_i may deviate from its associated signed axis. This implies that the angle between these two unit vectors, q_i and $[M_{c_i}]_{z_i}$, may not be zero. As we will see in later sections, it will be convenient to model these deviates not on S^2 (the unit sphere in $\mathbb{R}^3$) directly but in a *tangent plane*. To explain this concept, we now touch upon some differential-geometric notions.

Let p be a point in S^2 and T_pS^2 denote the tangent space to S^2 at point p; namely,

$$T_pS^2 = \{x : x \in \mathbb{R}^3 \; ; \; x^T p = 0\} \, . \tag{1}$$

While S^2 is nonlinear, T_pS^2 is a 2-dimensional linear space; see Fig. 2. This linearity of S^2 is what simplifies probabilistic modeling and statistical inference. The *Riemannian logarithm* (w.r.t. the point of the tangency p), $\mathrm{Log}_p : S^2 \setminus \{-p\} \to T_pS^2$, enables us to map points on the sphere (except the antipodal point: $-p$) to T_pS^2. Likewise, the *Riemannian exponential map*, $\mathrm{Exp}_p : T_pS^2 \to S^2$, maps T_pS^2 onto S^2. Note these two maps depend on the point of tangency p. Finally, if p and q are two points on S^2, then the geodesic distance between p and q is simply defined to be the angle between them: $d_G(p, q) = \arccos(p^T q)$. It can be shown that $d_G(p, q) = \left\|\mathrm{Log}_p(q)\right\|$. See our supplemental material or [12] for formulas and additional details.

Let us now return to the issue of the (angular) deviation of q_i from $[M_{c_i}]_{z_i}$: as long as q_i and $[M_{c_i}]_{z_i}$ are not antipodal points (see above), their deviation can be computed as $d_G([M_{c_i}]_{z_i}, q_i) = ||\mathrm{Log}_{[M_{c_i}]_{z_i}}(q_i)||$.

Given N observed normals, $\{q_i\}_{i=1}^N$, the sought-after parameters of an MMF are: K, $\{M_k\}_{k=1}^K$, $\{c_i\}_{i=1}^N$, and $\{z_i\}_{i=1}^N$. In order to fit these parameters, one would seek to penalize the deviates $\{||\mathrm{Log}_{[M_{c_i}]_{z_i}}(q_i)||\}_{i=1}^N$. While, in principle, this can be formulated as a deterministic optimization, we adopt a probabilistic approach.

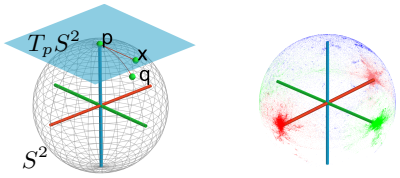

Figure 2: The signed axes of an MF displayed within S^2 (the unit sphere). The blue plane (left) illustrates T_pS^2, the tangent space to S^2 at $p \in S^2$ (here p is taken to be the north pole). A tangent vector $x \in T_pS^2$ is mapped to $q \in S^2$ via Exp_p; see text for details. The MF on the right is shown with its associated data (*i.e.*, normals viewed as points on S^2) whose colors indicate normal-to-axis assignments.

3.2. MMF: Probabilistic Model

In practice, scene representations may be comprised of multiple intermediate representations, which may include MMFs, to facilitate higher level reasoning. As such, adopting a probabilistic model allows one to describe and propagate uncertainty in the representation. Furthermore, it allows one to incorporate prior knowledge in a principled way, model inherent measurement noise, and derive tractable inference since conditional independence facilitates drawing samples in parallel.

Figure 3 depicts a graphical representation of the probabilistic MMF model. It is a Bayesian finite mixture model that takes into account the geometries of both S^2 and $\mathrm{SO}(3)$. In this probabilistic model, the MMF parameters are regarded as random variables.

The MF assignments c_i are assumed to be distributed according to a categorical distribution with a *Dirichlet distribution* prior with parameters α:

$$c_i \sim \mathrm{Cat}(\pi); \; \pi \sim \mathrm{Dir}(\alpha) \, . \tag{2}$$

Let $R_k \in \mathrm{SO}(3)$ denote the rotation associated with M_k. Making no assumptions about which orientation of M_k is more likely than others, R_k is distributed uniformly:

$$R_k \sim \mathrm{Unif}(\mathrm{SO}(3)) \, . \tag{3}$$

See the supplemental material for details.

At the second level of association, the z_i's are assumed to be distributed according to a categorical distribution w_{c_i} with a *Dirichlet distribution* prior parameterized by γ:

$$z_i \sim \mathrm{Cat}(w_{c_i}); \; w_{c_i} \sim \mathrm{Dir}(\gamma) \, . \tag{4}$$

The deviations of the observed normals from their signed axis are modeled by a 2D zero-mean Gaussian distribution in the tangent space to that axis:

$$p(q_i; [M_{c_i}]_{z_i}, \Sigma_{c_i z_i}) = \mathcal{N}(\mathrm{Log}_{[M_{c_i}]_{z_i}}(q_i); 0, \Sigma_{c_i z_i}) \, , \tag{5}$$

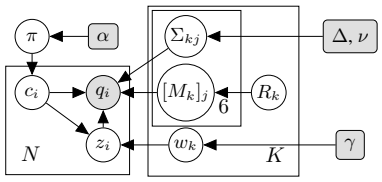

Figure 3: Graphical model for a mixture of K MFs.

where $\mathrm{Log}_{[M_{c_i}]_{z_i}}(q_i) \in T_{[M_{c_i}]_{z_i}} S^2$. In other words, we evaluate the probability density function (pdf) of $q_i \in S^2$ by first mapping it into $T_{[M_{c_i}]_{z_i}} S^2$ and then evaluating it under the Gaussian distribution with covariance $\Sigma_{c_i z_i} \in \mathbb{R}^{2\times 2}$. The pdf of the normals over the nonlinear S^2 is then induced by the Riemannian exponential map:

$$q_i \sim \mathrm{Exp}_{[M_{c_i}]_{z_i}}\left(\mathcal{N}(0, \Sigma_{c_i z_i})\right);\ \Sigma_{c_i z_i} \sim \mathrm{IW}(\Delta, \nu)\ . \quad (6)$$

Note that the range of Log_p is contained within a disk of finite radius (π) while the Gaussian distribution has infinite support. Consequently, we use an *inverse Wishart* (IW) prior that favors small covariances resulting in a probability distribution that, except a negligible fraction, is within the range of Log_p and concentrated about the respective axis.

We now explain how we choose the (hyper-parameters) α and γ. We set $\alpha < 1$ to favor models with few MFs, as expected for man-made scenes. To encourage the association of equal numbers of normals to all MF axes, we place a strong prior $\gamma \gg 1$ on the distribution of axis assignments z_i. Intuitively, this means that we want an MF to explain several normal directions and not just a single one.

3.3. MMF: Metropolis-Hastings MCMC Inference

We perform inference over the probabilistic MMF model described in Sec. 3.2 using Gibbs sampling with Metropolis-Hastings [18] split/merge proposals [26]. Specifically, the sampler iterates over the latent assignment variables $\mathbf{c} = \{c_i\}_{i=1}^N$ and $\mathbf{z} = \{z_i\}_{i=1}^N$, their categorical distribution parameters π and $\mathbf{w} = \{w_k\}_{k=1}^K$, as well as the covariances in the tangent spaces around the MF axes $\mathbf{\Sigma} = \{\{\Sigma_{kj}\}_{j=1}^6\}_{k=1}^K$ and the MF rotations $\mathbf{R} = \{R_k\}_{k=1}^K$. We first explain all posterior distributions needed for Gibbs sampling before we outline the algorithm.

3.3.1 Posterior Distributions for MCMC Sampling

The posterior distributions of both mixture weights are:

$$p(\pi|\mathbf{c}; \alpha) = \mathrm{Dir}(\alpha_1 + N_1, \ldots, \alpha_K + N_K) \quad (7)$$

$$p(w_k|\mathbf{c}, \mathbf{z}; \gamma) = \mathrm{Dir}(\gamma_1 + N_{k1}, \ldots, \gamma_{k6} + N_{k6})\,, \quad (8)$$

where $N_k = \sum_{i=1}^N \mathbb{1}_{[c_i=k]}$ is the number of normals assigned to the kth MF and $N_{kj} = \sum_{i=1}^N \mathbb{1}_{[c_i=k]}\mathbb{1}_{[z_i=j]}$ is the number of normals assigned to the jth axis of the kth MF. The indicator function $\mathbb{1}_{[a=b]}$ is 1 if $a = b$ and 0 otherwise.

Evaluating the likelihood of q_i as described in Eq. (5), the posterior distributions for labels c_i and z_i are given as:

$$p(c_i = k|\pi, q_i, \Theta) \propto \pi_k \sum_{j=1}^{6} w_{kj}\, p(q_i; [M_k]_j, \Sigma_{kj}) \quad (9)$$

$$p(z_i = j|c_i, q_i, \Theta) \propto w_{c_i j}\, p(q_i; [M_{c_i}]_j, \Sigma_{c_i j})\,, \quad (10)$$

where $\Theta = \{\mathbf{w}, \mathbf{\Sigma}, \mathbf{R}\}$. We compute $x_i = \mathrm{Log}_{[M_{c_i}]_{z_i}}(q_i)$, the mapping of q_i into $T_{[M_{c_i}]_{z_i}} S^2$, to obtain the scatter matrix $S_{kj} = \sum_i^N \mathbb{1}_{[c_i=k]}\mathbb{1}_{[z_i=j]} x_i x_i^T$ in $T_{[M_k]_j} S^2$. Using S_{kj} the posterior distribution over covariances Σ_{kj} is:

$$p(\Sigma_{kj}|\mathbf{c}, \mathbf{z}, \mathbf{q}, \mathbf{R}; \Delta, \nu) = \mathrm{IW}\left(\Delta + S_{kj}, \nu + N_{kj}\right)\ . \quad (11)$$

Since there is no closed-form posterior distribution for an MF rotation given axis-associated normals, we approximate it as a narrow Gaussian distribution on $\mathrm{SO}(3)$ around the optimal rotation $R_k^\star$ under normal assignments $\mathbf{z}$ and $\mathbf{c}$:

$$p(R_k|\mathbf{z}, \mathbf{c}, \mathbf{q}) \approx \mathcal{N}(R_k; R_k^\star(R_k^0, \mathbf{z}, \mathbf{c}, \mathbf{q}), \Sigma_{\mathrm{so}(3)})\,, \quad (12)$$

where $\Sigma_{\mathrm{so}(3)} \in \mathbb{R}^{3\times 3}$ and R_k^0 is set to R_k from the previous Gibbs iteration. Refer to the supplemental material for details on how to evaluate and sample from this distribution.

We now formulate the optimization procedure that yields a (locally-) optimal rotation $R_k^\star \in \mathrm{SO}(3)$ of MF M_k given a set of N_k assigned normals $\mathbf{q} = \{q_i\}_{i:c_i=k}$ and their associations z_i to one of the six axes $[M_k]_{z_i}$.

We find the optimal rotation as $R_k^\star = \arg\min_{R_k} F(R_k)$ where our cost function, $F : \mathrm{SO}(3) \to \mathbb{R}_+$, penalizes the geodesic deviation of a normal from its associated MF axis:

$$F(R_k) = \frac{1}{N_k} \sum_{i:c_i=k} \rho(d_G(q_i, [M_k]_{z_i}))\,. \quad (13)$$

To achieve robustness against noise and model outliers, instead of taking the non-robust $\rho : x \mapsto x^2$, we use the Geman-McClure robust function [4, 16]: $\rho_{\mathrm{GMC}} : x \mapsto x^2/(x^2 + \sigma^2)$.

Note that F is defined over $\mathrm{SO}(3)$, a nonlinear space. In order to ensure that not only the minimizer will be in $\mathrm{SO}(3)$ but also that the geometry of that space will be fully exploited, it is important to use appropriate tools from optimization over manifolds. Specifically, we use the conjugate-gradient algorithm suggested in [13]. We have found this successfully minimizes the cost function and converges in only a few iterations.

3.3.2 Metropolis-Hastings MCMC Sampling

The Gibbs sampler with Metropolis-Hastings split/merge proposals is outlined in Algorithm 1. For K MFs and

N normals the computational complexity per iteration is $O(K^2N)$. To let the order of the model adapt to the complexity of the distribution of normals on the sphere, we implement Metropolis-Hastings-based split/merge proposals. In the following we give a high level description of split and merge moves. A detailed derivation can be found in the supplemental material.

On a high level, a merge of MFs M_k and M_l consists of the steps: (1) assign all normals of M_l to M_k to obtain $\hat{M}_k$, remove M_l, and resample axis assignments $\hat{z}_i$ of the normals in $\hat{M}_k$; (2) sample the rotation of $\hat{M}_k$ as described in Sec. 3.3.1; and (3) sample $\{\hat{\Sigma}_{kj}\}_{j=1}^6$ under the new rotation.

A split of MF M_k into MFs $\hat{M}_l$ and $\hat{M}_m$ consists of sampling associations $\hat{c}_i$ to MFs $\hat{M}_l$ and $\hat{M}_m$ for all normals previously assigned to M_k and sampling axis assignments $\hat{z}_i$ for $\hat{M}_l$ and $\hat{M}_m$. Conditioned on the new assignments, new rotations and axis covariances are sampled.

Algorithm 1 One Iteration of the MMF Inference

1: Draw $\pi \mid \mathbf{c}; \alpha$ using Eq. (7)
2: Draw $\mathbf{c} \mid \pi, \mathbf{q}, \mathbf{R}, \Sigma$ in parallel using Eq. (9)
3: **for** $k \in \{1, \ldots, K\}$ **do**
4: Draw $w_k \mid \mathbf{c}, \mathbf{z}; \gamma$ using Eq. (8)
5: Draw $\mathbf{z} \mid \mathbf{c}, \mathbf{w}, \mathbf{q}, \mathbf{R}, \boldsymbol{\Sigma}$ in parallel using Eq. (10)
6: Draw $R_k | \mathbf{z}, \mathbf{c}, \mathbf{q}; \Sigma_{\mathrm{so}(3)}$ using Eq. (12)
7: Draw $\{\Sigma_{kj}\}_{j=1}^6 \mid \mathbf{c}, \mathbf{z}, \mathbf{q}, \mathbf{R}; \Delta, \nu$ using Eq. (11)
8: **end for**
9: Propose splits for all MFs
10: Propose merges for all MF combinations

4. Results and Applications

We now describe results for MMF inference from both depth images and a large scale LiDAR scan of a part of Cambridge, MA, USA. Additionally, we demonstrate the applicability and usefulness of the MMF to plane segmentation and depth camera calibration.

4.1. Computation of the Depth-Image Normal Map

As the MMF model relies on the structured and concentrated pattern of surface normals of the 3D scene, an accurate and robust estimation of normals is key. In a first step, our algorithm estimates the normal map[1] by extracting the raw normals as $q(u, v) = \frac{X_u \times X_v}{\|X_u \times X_v\|}$. Computed using forward finite-differences, X_u and X_v are the derivatives of the observed 3D surface patch w.r.t. its local parameterization as implied by the image coordinate system [20].

Since the depth image is subject to noise, we regularize the normal map in a way that preserves discontinuities to avoid artifacts at the edges of objects. This is done by total-variation (TV) regularization of the normal field, as specified in [28]. The total-variation of the map from the image domain into a matrix manifold (in our case the unit sphere) is minimized using a fast augmented-Lagrangian scheme. The resulting map indeed has a concentrated distribution of normals as can be seen in Fig. 1. We observe that inclusion of this regularization generally leads to better MMF models.

[1] If the entire scene happens to be a smooth surface then this coincides with the *Gauss map* [12], restricted to the observed portion of the surface.

4.2. MMF Inference from Depth Images

We infer an MMF in a coarse-to-fine approach. First, we down-sample to 120k normals and run the algorithm for $T = 80$ iterations proposing splits and merges throughout as described in Sec. 3.3. Second, using the thus obtained MMF, we sample labels for the full set of normals.

We use the following parameters for the inference of MMFs in all depth images: $\Sigma_{\mathrm{so}(3)} = (2.5°)^2\ \mathrm{I}_{3\times3}$ and $\sigma = 15°$. The hyper-parameters for the MMF were set to $\alpha = 0.01$, $\gamma = 120$, $\nu = 12k$, and $\Delta = (15°)^2\nu\ \mathrm{I}_{2\times2}$.

We first highlight different aspects and properties of the inference using the 3-box scene depicted in Fig. 1. For this scene, we initialized the number of MFs to $K = 6$. The algorithm correctly infers $K = 3$ MFs as displayed in Fig. 1 on the sphere and in the point cloud. The three extracted MFs correspond to the three differently rotated boxes in the depth image. While the blue MF consists only of the single box standing on one corner, the green and red MFs contain planes of the surrounding room in addition to their respective boxes. This highlights the ability of our model to pool normal measurements from the whole scene. On a Core i7 laptop, this inference takes our unoptimized single thread Python implementation 9 min on average. This could be sped up significantly by proposing splits and merges less frequently or by employing a sub-cluster approach for splits and merges as introduced by Chang and Fisher [7].

To evaluate the performance of the MMF inference algorithm, we ran it on the NYU V2 dataset [23] which contains 1449 RGB-D images of various indoor scenes. For each scene, we compare the number of MFs the algorithm infers to the number of MFs a human annotator perceives. We find that in 80% (for initial $K = 3$ MFs) and 81% (for initial $K = 6$ MFs) of the scenes our algorithm converged to the hand-labeled number of MFs. Qualitatively, the inferred MF orientations were consistent with the human-perceived layout of the scenes. Besides poor depth measurements due to reflections, strong ambient light, black surfaces, or range limitations of the sensor, the inference converged to the wrong number of MFs mainly because of close-by round objects or significant clutter in the scene. The latter failure cases violate the Manhattan assumption and are hence to be expected. However, we observe that the algorithm fails gracefully approximating round objects with several MFs or adding a "noise MF" to capture clutter. Hence, to eliminate

(a) 1 MF (b) 1 MF (c) 2 MFs (d) 2 MFs (e) 2 MFs (f) 2 MFs (g) 3 MFs

Figure 4: We show the RGB images of various indoor scenes in the 1st row and the inferred MMF model in the 2nd row. Fig. 4b, 4e, and 4f were taken from the NYU V2 depth dataset [23]. For the single-MF scenes to the left we color-code the assignment to MF axes (brighter colors designate opposing axes). For the rest of the scenes we depict the assignments to MFs in orange, blue and pink. Areas without depth information are colored black. In the 3rd row we show the log likelihood of the normals under the inferred MMF (see colorbar to the right). Plane segmentations are depicted in the last row.

"noise MFs", we consider only MFs with more than 15% of all normals. Over all scenes the algorithm converged to the same number of MFs in 90% of the scenes when initialized $K = 3$ MFs and $K = 6$ MFs. For these scenes the hand-labeled number of MFs was correctly inferred in 84% of the cases. These statistics show that the inference algorithm can handle a wide variety of indoor scenes and is not sensitive to the initial number of MFs.

In Fig. 4 we show several typical indoor scenes of varying complexity and the inferred MFs in the 2nd row. The inference algorithm was started with six MFs in all cases. For scenes 4a and 4b, the inference yielded a single MF each. We display the assignment to MF axes in red, green and blue, where opposite MF axes are distinguished by a weaker tone of the respective color. The algorithm infers $K = 2$ MFs for the scenes in Fig. 4c to 4f and $K = 3$ MFs for the scene in Fig. 4g. For those scenes we display assignment of normals to the different MFs in orange, blue and pink. The gray color stems from a mixture of blue and orange which occurs if MFs share an axis direction.

Given the inferred MMF parameters, we can evaluate the likelihood of a normal using Eq. (5). The log-likelihood for each normal is displayed in the 3rd row of Fig. 4: planar surfaces have high probability (black) while corners, round objects and noisy parts of the scene have low probability (yellow) under the inferred model. The likelihood is valuable to remove noisy measurements for further processing.

4.3. MMF Inference from LiDAR Data

To demonstrate the versatility of our model, we show the extraction of MFs from a large-scale LiDAR scan of a part of Cambridge, MA, USA. The point cloud generated from the scan has few measurements associated with the sides of buildings due to reflections of the glass facades. Additionally, the point cloud does not have homogeneous density due to overlapping scan-paths of the airplane. This explains the varying density of points in Fig. 5.

In order to handle noisy and unevenly sampled LiDAR data, we implement a variant of robust moving-least-squares normal estimation [14]. The local plane is estimated using RANSAC, based on a preset width that defines outliers of the plane model. The normal votes are averaged for each point from neighboring estimates based on a Gaussian weight w.r.t. the Euclidean distance from the estimator. We count only votes whose estimation had sufficient support in the RANSAC computation in the nearby point set.

Figure 5 shows the point cloud colored according to MF assignment of the normals on top of a gray street-map. We do not show the normals associated with upward pointing MF axes to avoid clutter in the image. Interestingly, the inferred MFs have clear directions associated with them: blue is the direction of Boston, green is the direction of Harvard and red is aligned with the Charles river waterfront. The fact that the inference converges to this MMF demonstrates

Figure 5: Inferred MMF from the LiDAR scanned urban scene on top of a gray street map. There is a clear separation into three MFs colored red, green and blue with the orientations indicated by the axes in the top-left corner. These MFs share the upward direction without imposing any constraints. Normals associated with upward axes are hidden to reveal the composition of the scene more clearly. Note that the underlying point cloud has varying density due to the scan-paths of the airplane.

the descriptive power of our model to capture large scale organizational structure in man-made environments.

4.4. Depth Camera Calibration

Our MMF provides us with associations of normals to MF axes which are assumed to be orthogonal to each other. We can exploit this to find the focal length f of a depth camera since q_i is influenced by f through the computation of the normals q_i as expressed in Sec. 4.1 and the inverse projection relationship between a point $(x, y, z)^T$ in 3D and a point $(u, v)^T$ in the image: $\begin{pmatrix} x \\ y \end{pmatrix} = \frac{z}{f}\begin{pmatrix} u-u_c \\ v-v_c \end{pmatrix}$, where $(u_c, v_c)^T$ is the image center.

This process, however, is nonlinear and does not have a closed-form solution for its derivative w.r.t. f. Therefore, we resort to exhaustive search to find the minimum of the cost function in Eq. (13) where we fix the MMF but introduce a dependency on f:

$$F(f) = \frac{1}{N}\sum_{i=1}^{N} \rho(d_G(q_i(f), [M_{c_i}]_{z_i})) \; . \tag{14}$$

Given a reasonable initialization of f (*i.e.* the factory calibration) we can determine f uniquely, without concerns of local minima, as shown in Fig. 6.

The angular deviation of about 4° in corners of point clouds vanishes after calibrating the focal length with our method. The calibration algorithm determines the focal length of our ASUS Xtion PRO depth camera to be $f = 540$ px whereas the factory calibration is $f = 570.3$ px.

While this can be viewed as the first step of an alternating minimization of both f and the MMF parameters, in practice, one update of f usually suffices. This provides us with a way of calibrating a depth scanner from a single depth image of any scene exhibiting MMF structure. Compared to other techniques [17, 19, 32] our proposed calibration procedure is much simpler.

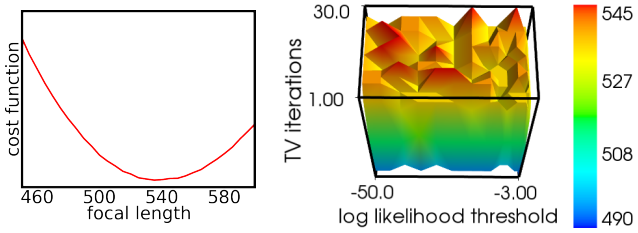

Figure 6: Left: the cost function $F(f)$ for a specific MMF. Right: estimated focal length as a function of the number of TV iterations and the log-likelihood threshold for normal selection.

4.5. Plane Segmentation

For a given scene the MMF provides us with the orientation of all planes. The normals of different planes with the same orientation contribute to the same MF axis. However, we can separate the planes by their offset in space along the respective MF axis.

After removing low-likelihood normals and combining MF axes pointing in the same direction (such as the normals of the floor in Fig. 1), we perform the plane segmentation for each MMF axis in two steps: First we project all 3D points, associated with a certain axis through their normal, onto the respective axis. Next, we bin these values, remove buckets under a certain threshold n_{bin} and collect points in consecutive bins into sets that constitute planes. We keep only planes that contain more than n_{plane} normals.

We found thresholds of $n_{\text{bin}} = 100$, $n_{\text{plane}} = 1000$ and a bin size of 3 cm to work well across all scenes in our evaluation. Fig. 4 shows the plane segmentation for several common indoor scenes in the 4th row. Despite the fact that our model does not utilize spatial regularity, we are able perform dense plane segmentation.

5. Conclusion

Motivated by the observation that the commonly-made Manhattan-World assumption is easily broken in man-made environments, we have proposed the *Mixture-of-Manhattan-Frames* model. Our inference algorithm, a manifold-aware Gibbs sampler with Metropolis-Hastings split/merge proposals, allows adaptive and robust inference of MMFs. This enables us to describe both complex small-scale-indoor and large-scale-urban scenes. We have shown the usefulness of our model by providing algorithms for plane segmentation and depth-camera focal-length calibra-

tion. Moreover, we have demonstrated the versatility of our model by extracting MMFs not only from 1.5k indoor scenes but also from aerial LiDAR data of Cambridge, MA.

Future work should incorporate color information into the estimation process. We expect that this will facilitate more robust MF inference because we will be able reason about parts of the scene that are too remote for the depth sensor. Another avenue of research would be to utilize the model to obtain robust rotation estimation in buildings for visual odometry. Due to the flexibility and robustness our framework in modeling real-world man-made environments, we envision many applications for it.

Acknowledgments. We thank J. Chang for his help with the split/merge proposals and R. Cabezas for his help with the Cambridge dataset. J.S., O.F., J.L. and J.F. were partially supported by the Office of Naval Research Multidisciplinary Research Initiative program, award N00014-11-1-0688. J.S., O.F. and J.F. were also partially supported by the Defense Advanced Research Projects Agency, award FA8650-11-1-7154. G.R. was partially funded by the MIT-Technion Postdoctoral Fellowships Program.

References

[1] M. Antunes and J. P. Barreto. A global approach for the detection of vanishing points and mutually orthogonal vanishing directions. In *CVPR*, pages 1336–1343. 2013.

[2] S. T. Barnard. Interpreting perspective images. *Artif. Intell.*, 21(4):435–462, 1983.

[3] C. Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225, 1974.

[4] M. J. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *IJCV*, 19(1):57–91, 1996.

[5] M. Bosse, R. Rikoski, J. Leonard, and S. Teller. Vanishing points and three-dimensional lines from omni-directional video. *The Visual Computer*, 19(6):417–430, 2003.

[6] B. Caprile and V. Torre. Using vanishing points for camera calibration. *IJCV*, 4(2):127–139, 1990.

[7] J. Chang and J. W. Fisher III. Parallel sampling of DP mixture models using sub-clusters splits. In *NIPS*, 2013.

[8] R. Cipolla, T. Drummond, and D. P. Robertson. Camera calibration from vanishing points in image of architectural scenes. In *BMVC*, vol. 99, pages 382–391, 1999.

[9] R. T. Collins and R. S. Weiss. Vanishing point calculation as a statistical inference on the unit sphere. In *ICCV*, pages 400–403, 1990.

[10] J. M. Coughlan and A. L. Yuille. Manhattan world: Compass direction from a single image by Bayesian inference. In *ICCV*, vol. 2, pages 941–947. 1999.

[11] E. Delage, H. Lee, and A. Y. Ng. Automatic single-image 3D reconstructions of indoor Manhattan world scenes. In *Robotics Research*, pages 305–321. 2007.

[12] M. P. Do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

[13] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIMAX*, 20(2):303–353, 1998.

[14] S. Fleishman, D. Cohen-Or, and C. T. Silva. Robust moving least-squares fitting with sharp features. In *SIGGRAPH*, pages 544–552, 2005.

[15] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *ICCV*, pages 80–87. 2009.

[16] S. Geman and D. E. McClure. Statistical methods for tomographic image reconstruction. In *Proc. of ISI*, 1987.

[17] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, 2004.

[18] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[19] D. Herrera C., J. Kannala, and J. Heikkil. Joint depth and color camera calibration with distortion correction. *TPAMI*, 34(10):2058–2064, 2012.

[20] D. Holz, S. Holzer, and R. B. Rusu. Real-Time Plane Segmentation using RGB-D Cameras. In *Proc. of the RoboCup Symposium*, 2011.

[21] A. E. Johnson and M. Hebert. Surface registration by matching oriented points. In *3DIM*, pages 121–128. 1997.

[22] M. Kazhdan. Reconstruction of solid models from oriented point sets. In *SGP*, page 73. 2005.

[23] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012.

[24] N. Neverova, D. Muselet, and A. Trémeau. 2 1/2D scene reconstruction of indoor scenes from single RGB-D images. In *CCIW*, pages 281–295. 2013.

[25] B. Peasley, S. Birchfield, A. Cunningham, and F. Dellaert. Accurate on-line 3D occupancy grids using manhattan world constraints. In *IROS*, pages 5283–5290. IEEE, 2012.

[26] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *JRSS: series B*, 59(4):731–792, 1997.

[27] G. Rosman, S. Shemtov, D. Bitton, T. Nir, G. Adiv, R. Kimmel, A. Feuer, and A. M. Bruckstein. Over-parameterized optical flow using a stereoscopic constraint. In *SSVM*, vol. 6667, pages 761–772, 2011.

[28] G. Rosman, Y. Wang, X.-C. Tai, R. Kimmel, and A. M. Bruckstein. Fast regularization of matrix-valued images. In *ECCV*, vol. 7574, pages 173–186. 2012.

[29] O. Saurer, F. Fraundorfer, and M. Pollefeys. Homography based visual odometry with known vertical direction and weak manhattan world assumption. *ViCoMoR*, 2012.

[30] G. Schindler and F. Dellaert. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *CVPR*, vol. 1, pages I–203. 2004.

[31] J. Stückler and S. Behnke. Orthogonal wall correction for visual motion estimation. In *ICRA*, pages 1–6. 2008.

[32] A. Teichman, S. Miller, and S. Thrun. Unsupervised intrinsic calibration of depth sensors via SLAM. In *RSS*, 2013.

[33] R. Triebel, W. Burgard, and F. Dellaert. Using hierarchical em to extract planes from 3d range scans. In *ICRA*, pages 4437–4442. 2005.