2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Leveraging Structural Regularity of Atlanta World for Monocular SLAM

Haoang Li, Yazhou Xing, Ji Zhao, Jean-Charles Bazin, Zhe Liu, and Yun-Hui Liu

***Abstract*— A wide range of man-made environments can be abstracted as the Atlanta world. It consists of a set of Atlanta frames with a common vertical (gravitational) axis and multiple horizontal axes orthogonal to this vertical axis. This paper focuses on leveraging the regularity of Atlanta world for monocular SLAM. First, we robustly cluster image lines. Based on these clusters, we compute the local Atlanta frames in the camera frame by solving polynomial equations. Our method provides the global optimum and satisfies inherent geometric constraints. Second, we define the posterior probabilities to refine the initial clusters and Atlanta frames alternately by the maximum a posteriori estimation. Third, based on multiple local Atlanta frames, we compute the global Atlanta frames in the world frame using Kalman filtering. We optimize rotations by the global alignment and then refine translations and 3D line-based map under the directional constraints. Experiments on both synthesized and real data have demonstrated that our approach outperforms state-of-the-art methods.**

I. INTRODUCTION

Simultaneous localization and mapping (SLAM) is a crucial technology for robot navigation. Among various SLAM strategies, monocular SLAM is based on a single camera and has gained popularity due to low price, compactness and flexibility [1]. Existing state-of-the-art monocular SLAM methods [2] [3] are reliable in highly textured scenes. However, in *structural* scenes (typically man-made environments) with less textures and repetitive patterns, the accuracy of these methods may decrease. Structural scenes exhibit particular regularity like parallelism and orthogonality, and recent papers showed that it can improve SLAM accuracy [4] [5].

Fig. 1 shows three typical structural scenes, i.e. Manhattan world (MW) [6], mixture of Manhattan frames (MMF) [7], and Atlanta world (AW) [8]. Each scene corresponds to specific regularity. MW has single frame (denoted by Manhattan frame (MF)), which holds for scenes with three mutually orthogonal dominant directions; MMF consists of multiple independent MFs and represents more general layout than MW; AW is composed of a set of frames sharing a common vertical axis (denoted by Atlanta frames (AFs)). Among these models, the MW regularity has been leveraged to improve SLAM accuracy [9]. However, the MW regularity-based SLAM is not applicable to many structures like non-orthogonal walls. MMF represents more general scenes, but each MF is estimated independently, which results in non-aligned 3D structures and limited effect of camera pose optimization [10]. In contrast, AW inherently encodes not only the *intra-AF* constraint (i.e. axes of each AF are mutually orthogonal), but also the *inter-AF* constraint (i.e. AFs share a common vertical axis). This paper investigates the AW regularity-based SLAM.

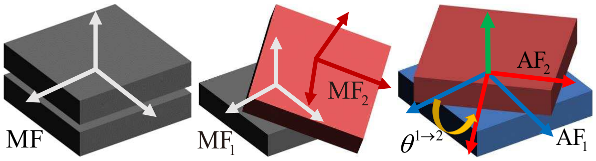

Fig. 1. Typical structural scenes. Left: Manhattan world [6] with a Manhattan frame (MF). Middle: mixture of MFs [7]. Right: Atlanta world [8] with multiple Atlanta frames (AFs) sharing a common vertical axis.

To leverage the AW regularity, three main challenges exist. First, existing AF estimation methods are inefficient for a large number of AFs [11], sensitive to thresholds [12], or prone to getting stuck in the local optimum [13]. Second, the accuracy of AFs is inevitably affected by noise. AF refinement is difficult due to their inherent constraints, i.e. intra- and inter-AF constraints. Third, each image corresponds to local AFs in the camera frame independently, leading to the uncertainty of global AFs in the world frame. How to compute the global AFs and apply the directional constraints for SLAM optimization remain open questions.

Our approach overcomes the above three challenges, as we will show in the experiments. The main contributions of this paper are summarized as follows.

- We propose a robust image line clustering method and an accurate AF computation approach that provides the global optimum in a non-iterative way;
- We design a novel optimization strategy based on the maximum a posteriori estimation to refine the initial clusters and AFs alternately;
- We compute reliable global AFs based on multiple local AFs and also optimize camera poses and 3D map under the directional constraints of AW.

II. RELATED WORK

Several papers exploited the structure of Manhattan world. For example, the vanishing point (VP) [14] encodes the parallelism and orthogonality of 3D lines. The VP estimation is related to clustering image lines w.r.t. unknown-but-sought VPs. It has been solved by RANSAC [15], consensus set maximization [16] [17], sampling [18] or J-Linkage [19]. The 3D plane normals also reflect the MW regularity, i.e.

This work is supported in part by the Natural Science Foundation of China under Grant U1613218, in part by the Hong Kong ITC under Grant ITS/448/16FP, and in part by the VC Fund 4930745 of the CUHK T Stone Robotics Institute.

H. Li, Z. Liu and Y.-H. Liu are with the T Stone Robotics Institute and Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, China. Y. Xing is with HKUST, Hong Kong, China. J. Zhao is with TuSimple, Beijing, China. J.-C. Bazin is with KAIST, Daejeon, South Korea. Email: {yhliu, hali}@mae.cuhk.edu.hk.

978-1-5386-6026-3/19/$31.00 ©2019 IEEE

they are parallel to the MW dominant directions. Straub et al. [20] modeled the parallelism efficiently to estimate the MW structure in real time. In addition, the MW regularity has been leveraged for robot navigation. For example, Li et al. [5] used VPs to reduce the error accumulation of SLAM. Kim et al. [21] proposed a low-drift visual odometry system by aligning 3D plane normals to MW dominant directions. The limitation of these methods is that they only work well under the specific MW assumption. Compared with above MW-based research, work on MMF is limited since the unknown number of MFs makes MF inference more challenging. Straub et al. [7] proposed a representative MMF estimation approach. It clusters 3D plane normals by a sampling-based strategy, but is sensitive to the initial value.

Some researchers also conducted work related to the Atlanta world. Schindler and Dellaert [8] exploited the expectation-maximization algorithm to infer AFs. However, their approach assumes the known vertical direction of AW. Antunes et al. [22] formulated the AF estimation as a multi-model fitting problem solved by a message passing-based strategy. It improves the accuracy at the cost of increasing the computational complexity. Joo et al. [11] recently proposed a globally optimal AF estimation method. They maximized the inlier set of 3D plane normals associated with the same AF. While it is accurate, its efficiency is not appropriate for robotic applications (around 6 seconds per image with two AFs). We will compare these methods with our approach in the experiments. Beyond the above AF estimation work, the AW structure has been applied to image parsing [23], camera calibration [24] and 3D reconstruction [25]. In the context of SLAM, the system of Lee et al. [4] uses VPs of multiple AFs for pose optimization. However, its accuracy dramatically decreases when the number of AFs is relatively large, as will be shown in the experiments. To the best of our knowledge, it is the only existing SLAM system for AW, and there has not been a mature SLAM method that effectively improves the accuracy by the AW regularity. In contrast, we estimate reliable AFs and significantly improve SLAM accuracy based on the directional constraints of AW.

III. PROBLEM FORMULATION

Given an image captured at time t, we use image lines to estimate local AFs in the camera frame. Then we exploit multiple local AFs estimated at times $\{t, t{-}1, \cdots\}$ to compute global AFs in the world frame, followed by refining camera poses and 3D map. In the following, we introduce the problem formulation.

A. Atlanta Frame Parameterization

We denote the number of AFs by M and concisely parameterize AFs by their inherent constraints (cf. Section I). Based on the intra-AF constraint, we express each AF by a 3×3 orthogonal matrix $\mathbf{A}^m$ $(m=1\cdots M)$ whose columns represent frame axes. We denote the n-th $(n=1,2,3)$ axis of the m-th AF by $[\mathbf{A}^m]^n$. Without loss of generality, we select the first AF $\mathbf{A}^1$ as the "reference AF". In the camera frame, we express $\mathbf{A}^1$ by applying a rotation $\mathbf{R}^1$ parameterized by the quaternion $\mathbf{q}^1$ to the axes of camera frame, i.e. $\mathbf{A}^1=\mathbf{R}^1$. We then express the other $M-1$ AFs based on the inter-AF constraint. We rotate $\mathbf{A}^1$ by the angle $\theta^{1\mapsto m}$ around the vertical axis (shown in Fig. 1) to parameterize AFs $\{\mathbf{A}^m\}_{m=2}^M$, i.e. $\mathbf{A}^m=\mathbf{R}^{1\mapsto m}\mathbf{A}^1$ where $\mathbf{R}^{1\mapsto m}$ is a 1 DOF rotation composed of the angle $\theta^{1\mapsto m}$. Therefore, for M AFs, our unknown parameter set is $\{\mathbf{q}^1, \{\theta^{1\mapsto m}\}_{m=2}^M\}$.

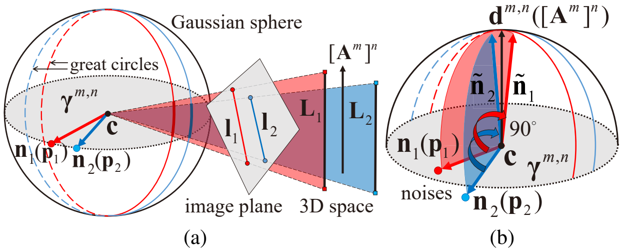

Fig. 2. Geometric relations on sphere. (a) For image lines $\mathbf{l}_1$ and $\mathbf{l}_2$ whose corresponding 3D lines $\mathbf{L}_1$ and $\mathbf{L}_2$ are parallel to the AF axis $[\mathbf{A}^m]^n$, their associated great circle normals $\mathbf{n}_1$ and $\mathbf{n}_2$ lie on the same plane $\boldsymbol{\gamma}^{m,n}$. (b) The normal $\mathbf{d}^{m,n}$ of $\boldsymbol{\gamma}^{m,n}$ is ideally parallel to $[\mathbf{A}^m]^n$. We rotate $\mathbf{n}_1$ $(\mathbf{n}_2)$ contaminated by noise within the plane formed by $\mathbf{n}_1$ $(\mathbf{n}_2)$ and $[\mathbf{A}^m]^n$ by 90°, generating the direction $\tilde{\mathbf{n}}_1$ $(\tilde{\mathbf{n}}_2)$ that deviates from $[\mathbf{A}^m]^n$.

B. Geometric Relations on Gaussian Sphere

We choose the Gaussian sphere [26] (hereinafter called the "sphere") as the AF inference space since it avoids searching in the infinite area of image plane [21]. As shown in Fig. 2(a), the sphere center is located at the camera center $\mathbf{c}$ and the radius is 1. The image line $\mathbf{l}_k$ is the projection of 3D line $\mathbf{L}_k$ parallel to the AF axis $[\mathbf{A}^m]^n$. The 3D plane formed by the center $\mathbf{c}$ and the line $\mathbf{l}_k$ intersects the sphere, forming a great circle with plane normal $\mathbf{n}_k$. The image line $\mathbf{l}_k$ is thus associated with unique vector $\mathbf{n}_k$. For image lines $\{\mathbf{l}_k\}$ whose corresponding 3D lines are parallel to the same AF axis $[\mathbf{A}^m]^n$, their associated vectors $\{\mathbf{n}_k\}$ lie on the same plane $\boldsymbol{\gamma}^{m,n}$ passing the center $\mathbf{c}$. For simplicity, we use the point $\mathbf{p}_k$ on sphere to represent the unit vector $\mathbf{n}_k$. The coplanarity of points $\{\mathbf{p}_k\}$ provides a cue to cluster them.

As shown in Fig. 2(b), we rotate the direction $\mathbf{n}_k$ by 90° within the plane formed by $\mathbf{n}_k$ and the AF axis $[\mathbf{A}^m]^n$, generating the direction $\tilde{\mathbf{n}}_k$. Due to the affect of noise, the consistency between directions $[\mathbf{A}^m]^n$ and $\tilde{\mathbf{n}}_k$ cannot be strictly satisfied. To model this direction deviation, we adopt the Bingham distribution [27] that conditions a trivariate Gaussian distribution to the sphere. The probability density of the Bingham distribution is

$$p_{\text{b}}(\tilde{\mathbf{n}}_k;[\mathbf{A}^m]^n,\mathbf{k})=\frac{1}{f(\mathbf{k})}\exp\left\{\sum_{n=1}^{3}\kappa_n\left(\tilde{\mathbf{n}}_k^\top[\mathbf{A}^m]^n\right)^2\right\}, \quad (1)$$

where $f(\mathbf{k})$ is the normalization constant and $\mathbf{k}$ is composed of the concentration parameters $\{\kappa_n\}_{n=1}^3$. Larger κ_n represents higher extent that Eq. (1) peaks at the direction $[\mathbf{A}^m]^n$.

IV. ATLANTA FRAME ESTIMATION

Given a set of images lines, this section explains how we estimate AFs in the camera frame. We first infer initial AFs by data clustering and then refine these AFs iteratively.

A. Initial Atlanta Frame Inference

We map image lines to points on sphere and cluster these points by co-planarity (cf. Section III-B). We employ the T-Linkage [28] to cluster points without the prior knowledge of the number of clusters. Our method is more *robust* than existing work. T-Linkage generates several hypothesized models by random sampling. Each data is evaluated by all models in terms of the fitting degree and associated with a descriptor. Then the data with similar descriptors are clustered together. In our context, the sphere center and two points randomly sampled from the point set form a 3D plane $\boldsymbol{\pi}_s$. We sample points S times, generating S planes $\{\boldsymbol{\pi}_s\}_{s=1}^{S}$. We compute the fitting degree between point $\mathbf{p}_k$ and planes $\{\boldsymbol{\pi}_s\}_{s=1}^{S}$ by

$$\psi_k(\boldsymbol{\pi}_s) = \begin{cases} e^{-d_\perp(\mathbf{p}_k,\boldsymbol{\pi}_s)/\tau}, & \text{if } d_\perp(\mathbf{p}_k,\boldsymbol{\pi}_s) < 5\tau, \\ 0, & \text{if } d_\perp(\mathbf{p}_k,\boldsymbol{\pi}_s) \geqslant 5\tau. \end{cases} \tag{2}$$

where $d_\perp(\cdot,\cdot)$ denotes the vertical distance from 3D point to 3D plane; τ is the time constant of exponential function. It is reasonable to set the inlier threshold by 5τ since for $d_\perp > 5\tau$, $e^{-d_\perp/\tau}$ approximates to a small constant (curve is flat). With S planes, each point $\mathbf{p}_k$ is associated with a S-dimension vector descriptor whose elements $\{\psi_k(\boldsymbol{\pi}_s)\}_{s=1}^{S}$ are in the *continuous* space $[0,1]$. In contrast, traditional J-Linkage [19] used for MF inference [12] generates binary vector descriptors whose elements are in the *discrete* space $\{0,1\}$. We use the exponential decline to replace the direct truncation (0 or 1), so our point descriptor is less sensitive to the threshold τ, as will be shown in the experiments.

Based on our robust descriptors, we agglomerate points that are originally assigned into independent clusters. We merge two clusters whose descriptors have the smallest Tanimoto distance [19] $d_{\mathrm{T}} \in [0,1]$ until d_{T} between any two clusters equals to 1. In addition, we robustly discard clusters formed by outliers. In brief, we leverage the fact that the cluster with lower point cardinality has higher probability of clustering outliers by coincidence [29].

Then we compute AFs based on the obtained point clusters. We exploit each cluster of roughly co-planar points to compute the least-squares solution of the plane normal. Among these normals, we first find a special one $\mathbf{d}^{\mathrm{v}}$ that is nearly orthogonal to others. If $\mathbf{d}^{\mathrm{v}}$ does not exist, it is computed by the cross-product of any other two normals. For the rest of plane normals, we find 1) each two nearly orthogonal ones $\mathbf{d}_1^{\mathrm{h}}$ and $\mathbf{d}_2^{\mathrm{h}}$ and combine them with $\mathbf{d}^{\mathrm{v}}$ as $\mathcal{D}_m = \{\mathbf{d}^{\mathrm{v}}, \mathbf{d}_1^{\mathrm{h}}, \mathbf{d}_2^{\mathrm{h}}\}$, and 2) each independent normal $\mathbf{d}_1^{\mathrm{h}}$ not orthogonal to any others and combine it with $\mathbf{d}^{\mathrm{v}}$ as $\mathcal{D}_m = \{\mathbf{d}^{\mathrm{v}}, \mathbf{d}_1^{\mathrm{h}}\}$. Ideally, $\mathbf{d}^{\mathrm{v}}$ is parallel to the common vertical AF axis and $\mathbf{d}_1^{\mathrm{h}}$ and $\mathbf{d}_2^{\mathrm{h}}$ are parallel to two mutually orthogonal horizontal AF axes of a certain AF. Under the presence of noise, each set $\mathcal{D}_m$ corresponds to a *quasi*-AF that might not strictly satisfy the intra-AF constraint. A straightforward way is to project all quasi-AFs to $SO(3)$ [20], but it fails to satisfy the inter-AF constraint due to respective vertical axes. We propose a novel method to compute AFs satisfying inherent constraints.

A plane normal corresponding to the AF axis $[\mathbf{A}^m]^n$ is associated with W roughly coplanar points $\{\mathbf{p}_w^{m,n}\}_{w=1}^{W}$. As each N ($N=2$ or 3) *roughly* mutually orthogonal normals are combined into the set $\mathcal{D}_m$ above, their associated N clusters of points are combined as $\mathcal{P}_m = \{\{\mathbf{p}_w^{m,n}\}_{w=1}^{W}\}_{n=1}^{N}$. We use each $\mathcal{P}_m$ to compute an AF $\mathbf{A}^m$ by

$$\arg\min_{\mathbf{A}^m} \sum_{n=1}^{N} \left(\sum_{w=1}^{W} \left([\mathbf{A}^m]^{n\top} \mathbf{p}_w^{m,n} \right)^2 \right). \tag{3}$$

We parameterize Eq. (3) by our AF parameterization that encodes inherent constraints of AFs (cf. Section III-A). To avoid the local optimum of Eq. (3) obtained by the *iterative* method [13], we construct polynomial equations and obtain the globally optimal AF in a *non-iterative* way. For the reference AF, Eq. (3) is parametrized by the quaternion $\mathbf{q}^1$. We take partial derivatives w.r.t. the r-th component q_r^1 ($r = 1 \cdots 4$) of $\mathbf{q}^1$ respectively, obtaining four cubic multivariate polynomial equations. We solve this polynomial system by the Gröbner basis [30] to obtain $\{q_r^1\}_{r=1}^{4}$. For the other AFs, we reformulate Eq. (3) as a polynomial w.r.t. $\sin(\theta^{1\rightarrow m})$ and $\cos(\theta^{1\rightarrow m})$. We derive a univariate polynomial equation and solve the angle $\theta^{1\rightarrow m}$ easily.

B. Atlanta Frame and Point Cluster Refinement

The accuracy of the initial AFs obtained in Section IV-A may be affected by some mistakenly clustered points. This is attributed to the point descriptors that are inevitably contaminated by noise. To refine these initial point clusters, we leverage the maximum a posteriori (MAP) estimation [31]. MAP obtains the optimal parameters that maximize the posterior probability $p(\Omega|\mathbf{x})$, i.e. $\Omega^* = \arg\max_\Omega p(\Omega|\mathbf{x})$ where Ω is the parameter set and $\mathbf{x}$ are observations. Based on the Bayesian rule [31], the posterior probability is computed by $p(\Omega|\mathbf{x}) \propto p(\Omega) \cdot L(\mathbf{x};\Omega)$ where $p(\Omega)$ is the prior probability and $L(\mathbf{x};\Omega)$ is the likelihood. In our context, each point $\mathbf{p}_k$ belongs to a cluster and corresponds to an AF axis. We associate $\mathbf{p}_k$ with AF by 1) the *frame label* $c_k = m \in \{1 \cdots M\}$ to indicate $\mathbf{p}_k$ corresponds to the m-th AF; 2) the *axis label* $z_k = n \in \{1 \cdots N\}$ to indicate $\mathbf{p}_k$ corresponds to the n-th axis of the c_k-th AF. For each point $\mathbf{p}_k$, we find the optimal labels c_k and z_k maximizing the posterior probability that $\mathbf{p}_k$ belongs to certain cluster (or corresponds to certain AF axis). We define the posterior probability by 1) the *prior probability* related to the descriptors of initial clusters, and 2) the *likelihood* related to the geometric relations on sphere.

We first define the prior probability. For points from the same initial cluster $\{\mathbf{p}_w^{m,n}\}_{w=1}^{W}$, their assigned descriptors $\mathcal{V} = \{\mathbf{v}_w^{m,n}\}_{w=1}^{W}$ are similar (cf. Section IV-A). We apply SVD [26] to all descriptors of $\mathcal{V}$ and select the vector with the largest singular value as the *representative* descriptor $\bar{\mathbf{v}}^{m,n}$ of $\mathcal{V}$. We measure the similarity between the descriptor $\mathbf{v}_k$ of point $\mathbf{p}_k$ and the representative descriptor $\bar{\mathbf{v}}^{m,n}$ of descriptor set $\mathcal{V}$ by the Tanimoto distance [19] that is denoted by $t_k^{m,n}$. Then we define the "similarity ratio" between $\mathbf{v}_k$ and $\bar{\mathbf{v}}^{m,n}$ as $r_k^{m,n} = (1 - t_k^{m,n}) \times 100$. Intuitively, for the point $\mathbf{p}_k$ corresponding to the AF axis $[\mathbf{A}^m]^n$ rather than $[\mathbf{A}^u]^v$, the ratio $r_k^{m,n}$ is higher than $r_k^{u,v}$. Based on $r_k^{m,n}$, we design two "similarity vectors": M-dimension similarity vector $\boldsymbol{\alpha} =$

$[\max\{\{r_k^{1,n}\}_{n=1}^N\},\cdots,\max\{\{r_k^{M,n}\}_{n=1}^N\}]^\top$ for the label c_k, and N-dimension similarity vector $\boldsymbol{\beta}=[r_k^{c_k,1},\cdots,r_k^{c_k,N}]^\top$ for the label z_k. We use $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to model the prior probabilities $p(c_k=m)$ and $p(z_k=n)$ respectively as follows.

The AF label c_k may take more than two possible values ($M>2$), so it follows the Multinoulli distribution (MD) [31] that is the generalization of the Bernoulli distribution. Based on the probability mass of MD, the prior probability $p(c_k=m;\boldsymbol{\mu})$ equals to μ_m with the given probability vector $\boldsymbol{\mu}=[\mu_1,\cdots,\mu_M]^\top$ ($\mu_m\in(0,1)$, $\sum_{m=1}^M\mu_m=1$). However, exact $\boldsymbol{\mu}$ is unknown in our problem and thus is treated as a random variable. Intuitively, there is a positive correlation between the probability vector $\boldsymbol{\mu}$ and the similarity vector $\boldsymbol{\alpha}$ in terms of the magnitude of corresponding element. We thus model $\boldsymbol{\mu}$ by the Dirichlet distribution (DD) [31] parameterized by the similarity vector $\boldsymbol{\alpha}$. DD describes the probability that the probability vector $\boldsymbol{\mu}$ occurs, and its probability density is

$$f_\text{d}(\boldsymbol{\mu};\boldsymbol{\alpha})=\frac{1}{b(\boldsymbol{\alpha})}\prod_{m=1}^M(\mu_m)^{\alpha_m-1}, \tag{4}$$

where $b(\boldsymbol{\alpha})$ is the normalization constant and α_m is the m-th element of $\boldsymbol{\alpha}$. We obtain the approximation $\tilde{\mu}_m$ of μ_m by sampling Eq. (4) and set the prior probability $p(c_k=m;\boldsymbol{\mu})$ as $\tilde{\mu}_m$. Similarly, the axis label z_k follows MD. The probability that the probability vector $\boldsymbol{\omega}^{c_k}=[\omega_1^{c_k},\cdots,\omega_N^{c_k}]^\top$ occurs follows $f_\text{d}(\boldsymbol{\omega}^{c_k};\boldsymbol{\beta})$ parameterized by the similarity vector $\boldsymbol{\beta}$ (cf. Eq. (4)). We thus set the prior probability $p(z_k=n;\boldsymbol{\omega}^{c_k})$ as the approximation $\tilde{\omega}_n^{c_k}$ sampled from $f_\text{d}(\boldsymbol{\omega}^{c_k};\boldsymbol{\beta})$.

We then define the likelihood. As introduced in Section III-B, we generate the direction $\tilde{\mathbf{n}}_k$ using the point $\mathbf{p}_k$. Intuitively, smaller direction deviation between $\tilde{\mathbf{n}}_k$ and the AF axis $[\mathbf{A}^m]^n$ represents higher likelihood that the point $\mathbf{p}_k$ corresponds to $[\mathbf{A}^m]^n$. To model this deviation, we adopt the Bingham distribution [27] w.r.t. the AF axis $[\mathbf{A}^m]^n$ and the vector $\tilde{\mathbf{n}}_k$. Based on Eq. (1), we define the likelihood $L_m(\tilde{\mathbf{n}}_k;[\mathbf{A}^m]^n,\mathbf{k},\tilde{\omega}_n^m)$ for the frame label c_k:

$$L_m=\sum_{n=1}^N\left(\tilde{\omega}_n^m\cdot p_\text{b}(\tilde{\mathbf{n}}_k;[\mathbf{A}^m]^n,\mathbf{k})\right), \tag{5}$$

and the likelihood $L_n(\tilde{\mathbf{n}}_k;[\mathbf{A}^{c_k}]^n,\mathbf{k})$ for the axis label z_k:

$$L_n=p_\text{b}(\tilde{\mathbf{n}}_k;[\mathbf{A}^{c_k}]^n,\mathbf{k}). \tag{6}$$

After defining the prior probabilities and likelihoods, we compute the posterior probabilities as $p(c_k=m|\tilde{\mathbf{n}}_k)\propto\tilde{\mu}_m\cdot L_m$ and $p(z_k=n|\tilde{\mathbf{n}}_k)\propto\tilde{\omega}_n^{c_k}\cdot L_n$. We maximize these two posterior probabilities to obtain the optimal labels c_k and z_k that indicate which cluster the point $\mathbf{p}_k$ belongs to. We adjust clusters accordingly and then optimize AFs by Eq. (3), which in turn refines clusters by Eq. (5) and Eq. (6). We optimize clusters and AFs alternately until convergence.

V. SLAM Optimization

Single image corresponds to the *local* AFs in the camera frame. Given an image sequence, we use multiple local AFs to compute the *global* AFs in the world frame. Then we exploit the directional constraints for SLAM optimization.

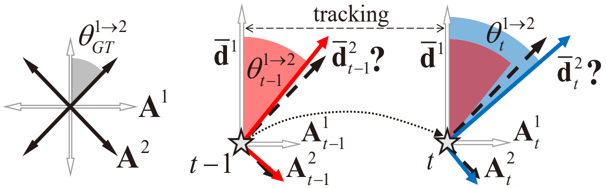

Fig. 3. Top view of an AW composed of the reference AF $\mathbf{A}^1$ (white) and another AF $\mathbf{A}^2$ (black). We estimate the local AFs $\{\mathbf{A}_{t-1}^1,\mathbf{A}_{t-1}^2\}$ at time $t-1$ and $\{\mathbf{A}_t^1,\mathbf{A}_t^2\}$ at time t independently. Assume that the estimation of $\mathbf{A}_{t-1}^2$ (red) and $\mathbf{A}_t^2$ (blue) is affected by noise. Angles $\theta_{t-1}^{1\to2}$ and $\theta_t^{1\to2}$, which are equal in theory, both deviate from the ground truth one $\theta_{GT}^{1\to2}$.

A. Computing Global Atlanta Frames

For the camera P_1 (time $t=1$), we set its AF associated with the most points on sphere (the least affected by noise) as the reference AF $\mathbf{A}_{t=1}^1$ and track it along the image sequence. We set the frame of camera P_1 as the world frame, so the local reference AF $\mathbf{A}_{t=1}^1$ serves as one of global AFs. Remaining global AFs are *initialized* as other local AFs $\{\mathbf{A}_{t=1}^m\}_{m=2}^M$ of camera P_1 and *refined* by multiple local AFs estimated at different times $\{t,t\text{–}1,\cdots\}$. As shown in Fig. 3, the difference between angles $\theta_{t-1}^{1\to2}$ and $\theta_t^{1\to2}$ results in the uncertainty of the global AF axis $\bar{\mathbf{d}}^2$ w.r.t. the global AF axis $\bar{\mathbf{d}}^1$ defined by the reference AF $\mathbf{A}_{t=1}^1$ above. We refine the angle $\theta^{1\to2}$ by Kalman filter [32]. At time t, we treat angle $\theta_t^{1\to2}$ as the *observation* and use angle $\theta_{t-1}^{1\to2}$ to compute the *prediction* $\tilde{\theta}_t^{1\to2}$. The refined angle $\hat{\theta}_t^{1\to2}$ is given by

$$\hat{\theta}_t^{1\to2}=\tilde{\theta}_t^{1\to2}+G\cdot(\theta_t^{1\to2}-\tilde{\theta}_t^{1\to2}), \tag{7}$$

where G is the Kalman gain. Then we rotate the preset axis $\bar{\mathbf{d}}^1$ by the refined angle $\hat{\theta}_t^{1\to2}$, obtaining the global AF axis $\bar{\mathbf{d}}_2$. We perform filtering continuously and leverage these global AFs for SLAM optimization.

B. Camera Pose and 3D Map Optimization

Aligning the *local* AFs to the *global* AFs provides the basis of rotation optimization. Let $\bar{\mathbf{d}}^{m,n}$ be the global AF axis in the world frame. We use the rotation $\mathbf{R}_i$ to transform $\bar{\mathbf{d}}^{m,n}$ to the frame of camera P_i by $\mathbf{d}^{m,n}=\mathbf{R}_i\bar{\mathbf{d}}^{m,n}$. Since $\mathbf{d}^{m,n}$ is ideally parallel to the local AF axis $[\mathbf{A}^m]^n$ of camera P_i, we define the cost function to optimize $\mathbf{R}_i$ as

$$\arg\max_{\mathbf{R}_i}\sum_{m=1}^M\left(\sum_{n=1}^N[\mathbf{A}^m]^{n\top}\mathbf{R}_i\bar{\mathbf{d}}^{m,n}\right), \tag{8}$$

To avoid the local optimum of the iterative method [5], we represent the rotation $\mathbf{R}_i$ by the unit quaternion and solve the optimal closed-form solution of $\mathbf{R}_i$ by SVD [33]. Since we align the local AFs of each camera to the global AFs independently, the error accumulation of rotation is reduced.

Then for the camera P_i, we optimize its translation and observed 3D lines by fixing its optimized rotation. We represent a 3D line in the world frame by the Plücker coordinates [26] as $\bar{\boldsymbol{\xi}}=[\bar{\boldsymbol{\eta}}^\top,\bar{\boldsymbol{\nu}}^\top]^\top$ where $\bar{\boldsymbol{\eta}}$ is the 3D normal of the plane passing this line and the origin of frame; $\bar{\boldsymbol{\nu}}$ is the 3D line direction. Let $\bar{\boldsymbol{\xi}}$ be a 3D line to optimize. In AW, each 3D line is ideally parallel to a certain AF axis. We thus constrain the line direction $\bar{\boldsymbol{\nu}}$ as the direction along the AF

TABLE I
ACCURACY AND EFFICIENCY EVALUATION OF AF ESTIMATION.

methods (**-AE**)	precision (%)		recall (%)		ADA (deg)		time (s)
	mean	median	mean	median	mean	median	mean
JL [19]	83.21	86.95	77.34	79.19	5.57	4.84	0.13
MP [22]	91.62	92.10	96.68	96.14	2.44	2.29	0.37
TL (our)	96.57	96.93	97.75	98.06	0.69	0.65	0.29
TL-MAP (our)	98.52	98.89	97.87	98.13	0.58	0.57	0.45

axis $\bar{\mathbf{d}}^{m,n}$ that has the smallest angle with $\bar{\boldsymbol{\nu}}$. Then we focus on the 3D line parameter $\bar{\boldsymbol{\eta}}$ and translation.

Assume that the 3D line $\bar{\boldsymbol{\xi}}$ is observed by the camera P_i, i.e. it is associated with an image line $\mathbf{l}_i$ of P_i. We transform $\bar{\boldsymbol{\xi}}$ to the camera frame of P_i as $\boldsymbol{\xi}_i = [\boldsymbol{\eta}_i^\top, \boldsymbol{\nu}_i^\top]^\top = \mathbf{M}_i\bar{\boldsymbol{\xi}}$ where $\mathbf{M}_i$ is the motion matrix [34] composed of rotation $\mathbf{R}_i$ and translation $\mathbf{t}_i$. Then we project the 3D line $\boldsymbol{\xi}_i$ by the camera P_i as $\tilde{\mathbf{l}}_i = \tilde{\mathbf{K}}\boldsymbol{\eta}_i$ where $\tilde{\mathbf{K}}$ is the known line projection matrix [34]. The projected image line $\tilde{\mathbf{l}}_i$ is thus formulated by the function w.r.t. the 3D line parameter $\bar{\boldsymbol{\eta}}$ and translation $\mathbf{t}_i$. For the 3D line $\bar{\boldsymbol{\xi}}$ observed by I cameras $\{P_i\}_{i=1}^{I}$ with translations $\mathcal{T} = \{\mathbf{t}_i\}_{i=1}^{I}$, we define the re-projection error-based cost function w.r.t. the *projection* $\tilde{\mathbf{l}}_i$ and the endpoints $\mathbf{s}_i$ and $\mathbf{e}_i$ of the *observation* $\mathbf{l}_i$ by

$$\arg\min_{\bar{\boldsymbol{\eta}},\mathcal{T}} \sum_{i=1}^{I} \left(d^2(\mathbf{s}_i, \tilde{\mathbf{l}}_i) + d^2(\mathbf{e}_i, \tilde{\mathbf{l}}_i) \right), \tag{9}$$

where $d(\cdot,\cdot)$ represents the distance from point to line in the image. We solve Eq. (9) by the Levenberg-Marquardt (LM) method [26] to optimize the 3D line parameter $\bar{\boldsymbol{\eta}}$ and camera translations $\mathcal{T}$. Note that Eq. (9) can be easily extended for multiple 3D lines. Compared with traditional algorithm neglecting the geometric priors of structural scenes [3], our method optimizes SLAM by the AW regularity and achieves higher accuracy, as will be shown in the experiments.

VI. EXPERIMENTS

To validate the proposed methods for AF estimation and SLAM optimization, we conduct experiments on synthesized data and real images. We compare our method with state-of-the-art approaches in terms of accuracy and efficiency. The supplementary material and datasets are available at `https://sites.google.com/view/haoangli/projects/atlanta-slam`.

We list state-of-the-art AF estimation and SLAM methods (cf. Sections I and II) to compare with our approaches.

a) AF estimation (AE): We condition the J-Linkage [19] to estimate AFs and denote it by **JL-AE**. We also test the message-passing-based strategy [22] denoted by **MP-AE**. We compare them with our T-Linkage-based approach (cf. Section IV-A) denoted by **TL-AE**. Besides, we report the result of **TL-AE** optimized by our MAP-based strategy (cf. Section IV-B), which is denoted by **TL-MAP-AE**.

b) SLAM: **ORB-SLAM** [2] is a point-based method. Built on ORB-SLAM, the point and line-based **PL-SLAM** [3] adds an independent line-based module composed of localization, mapping and bundle adjustment. We denote this line-based module by **Line-SLAM**. We also test **VP-Line-SLAM** [4] that uses VPs for optimization. Our SLAM optimization strategy (cf. Section V) can be integrated into most existing SLAM methods that fully or partially use lines. We report the representative result obtained by the integration of our AW regularity-based optimization strategy and above Line-SLAM [3], which is denoted by **AWR-Line-SLAM**.

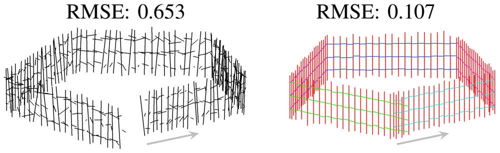

Fig. 4. Reconstructed 3D line-based maps on the synthesized image sequence. Left: **Line-SLAM** [3]. Right: our **AWR-Line-SLAM** (3D lines aligned to different AF axes are shown in respective colors). For **VP-Line-SLAM** [4], please refer to the supplementary material.

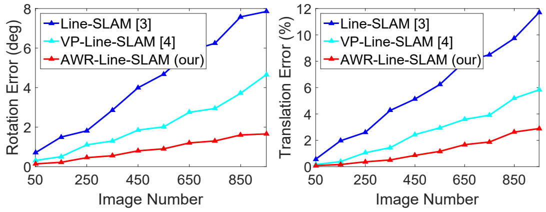

Fig. 5. Pose error accumulation w.r.t. the image number on the synthesized image sequence. Left: rotation. Right: translation.

We evaluate the accuracy of AF estimation in terms of image line clustering and AF axis computation. For clustering, we define 1) the *precision* by $\delta^{\mathrm{c}}/(\delta^{\mathrm{c}}+\delta^{\mathrm{w}})$ where δ^{c} and δ^{w} are the numbers of correctly and wrongly clustered lines respectively; 2) the *recall* by $\delta^{\mathrm{c}}/(\delta^{\mathrm{c}}+\delta^{\mathrm{m}})$ where δ^{m} is the number of missing lines. For AF axis computation, we calculate the average deviation angle (ADA) between the computed and ground truth axes. In addition, we evaluate the accuracy by the criteria used by [35], [36]. Given the ground truth rotation $\mathbf{R}^{\mathrm{GT}}$ and translation $\mathbf{t}^{\mathrm{GT}}$, we compute 1) the *rotation error* $E_{\mathbf{R}}(\mathrm{deg}) = \max_{n=1}^{3}\{\arccos\left((\mathbf{r}_{\mathrm{n}}{}^{\top}\mathbf{r}_{\mathrm{n}}^{\mathrm{GT}})\right) \times 180/\pi\}$ where $\mathbf{r}_n$ and $\mathbf{r}_n^{\mathrm{GT}}$ are the n-th columns of $\mathbf{R}$ and $\mathbf{R}^{\mathrm{GT}}$, respectively; 2) the *translation error* $E_{\mathbf{t}}\,(\%) = \|\mathbf{t}^{\mathrm{GT}} - \mathbf{t}\|/\|\mathbf{t}\| \times 100$; 3) the *Hausdorff distance* d_{H} between two point sets sampled along the computed and ground truth 3D lines respectively. We report the root mean squared error (RMSE) of $\{d_{\mathrm{H}}\}$.

A. Synthesized Data

We generate a 3D five-side fence whose top view is a regular pentagon to synthesize an AW with five AFs. Each side is composed of 80 3D line segments aligned to two AF axes (3×20 horizontal and 20 vertical ones). The endpoints are within the interval $[-2,2]\times[-2,2]\times[4,8]$. We define 1000 cameras uniformly distributed around the fence. The focal length is 800 pixels and the principle point is located at the center of image whose size is 640×480 pixels. We project the fence by these cameras to generate an image sequence. We perturb the endpoints of the projected line segments by zero-mean Gaussian noise whose standard deviation is 2 pixels.

a) Single Image: We use each image of the above sequence to estimate AFs independently and the results are shown in Table I. In terms of accuracy, **JL-AE** has unsatisfactory performance on both precision and recall of

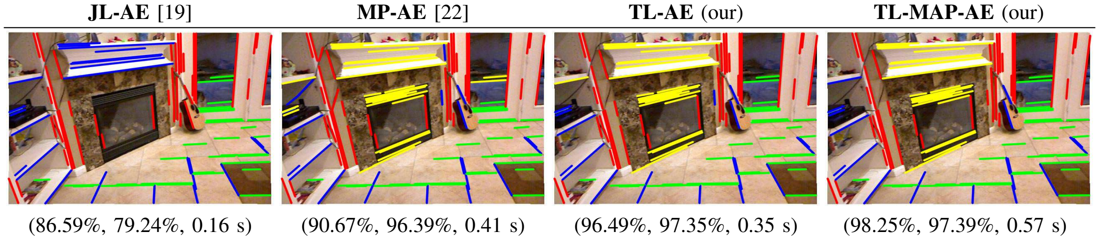

Fig. 6. AF estimation on the *NYU-V2* dataset [37]. Different image line clusters are shown in respective colors. The triplet of numbers below each image represents the precision and recall of image line clustering, and the total time of AF estimation.

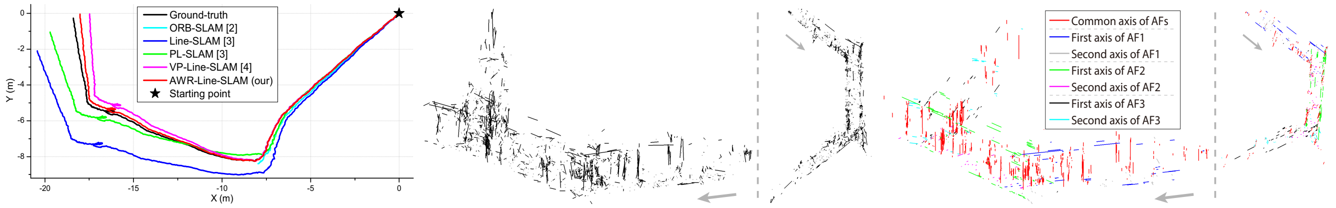

Fig. 7. Experimental results on our *Corridor* dataset. Left: estimated camera trajectories (top view). Middle: 3D map of **PL-SLAM** [3] (points and lines are shown in gray and black respectively). Right: 3D map of our **AWR-Line-SLAM** (lines aligned to different AF axes are shown in respective colors).

image line clustering. Besides, it obtains inaccurate AFs. **MP-AE** improves the recall but the precision is relatively low since it is sensitive to noise. In contrast, our **TL-AE** robustly clusters image lines and computes accurate AFs. Our **TL-MAP-AE** further improves the accuracy by the proposed optimization strategy. In terms of efficiency, **JL-AE** is the fastest method and **MP-AE** takes almost three times longer. Compared with **JL-AE**, our **TL-AE** is more time-consuming. The additional time cost is mainly attributed to more complex descriptors. Our **TL-MAP-AE** sacrifices partial efficiency to improve the accuracy, but its running time is still acceptable.

b) Image Sequence: We evaluate the accuracy of the line-based SLAM methods in terms of 3D map (shown in Fig. 4) and camera pose (shown in Fig. 5). Without the structural constraint-based optimization, **Line-SLAM** generates a disordered map and its trajectory has significant drift error. **VP-Line-SLAM** obtains more accurate result, but at fence corners, the error increases due to unreliable estimated VPs. In contrast, our **AWR-Line-SLAM** achieves the smallest error accumulation thanks to our SLAM optimization strategy.

B. Real Images

We conduct experiments on the *NYU-V2* dataset [37] and a self-collected dataset.

a) NYU-V2 dataset: We manually select a set of images captured in the scene with at least two AFs. Fig. 6 shows representative AF estimation results. In terms of accuracy, both precision and recall of **JL-AE** are relatively low since it is sensitive to the inlier threshold. **MP-AE** infers AFs more completely. However, it is vulnerable to outliers and noises, leading to many mistakenly clustered lines. In contrast, our **TL-AE** achieves high precision and recall thanks to our reliable descriptors and outlier rejection strategy. In addition, our **TL-MAP-AE** further improves the accuracy by the proposed MAP-based optimization strategy. In terms of efficiency, **JL-AE** is the fastest method. **MP-AE** and our **TL-AE** also show relatively high efficiency. **TL-MAP-AE** is more time-consuming due to the iteration for AF optimization, but its iteration quickly converges thanks to reliable initial value provided by our **TL-AE**.

b) Self-collected dataset: We collect an image sequence in a hotel corridor composed of three AFs and denote it by the *Corridor* dataset. The sequence is captured by a hand-held single camera and it consists of 1859 images of 960×540 pixels. We obtain the ground truth trajectory (32.51 m) by the Hokuyo UTM-30LX laser scanner.

We evaluate the SLAM accuracy as shown in Fig. 7. In terms of localization, **ORB-SLAM** fails to cover the whole distance since its point-based tracking module collapses at the corner where the detected points are scarce. **Line-SLAM** obtains a complete trajectory, but its drift error is significant. Without the structural constraints, the line-based localization is sensitive to noise. **PL-SLAM** uses points and lines simultaneously to improve the number of observations for noise compensation. However, the accuracy improvement is limited. **VP-Line-SLAM** only obtains accurate trajectory in the straightway with single AF. At the corner composed of more AFs, it estimates inaccurate VPs that affect the pose optimization. In contrast, our **AWR-Line-SLAM** obtains the most accurate trajectory thanks to our optimization strategy. In terms of mapping, for **PL-SLAM**, its 3D point-based map is too sparse to identify the corridor structure; its 3D line-based map sketches the spatial layout better, but the accuracy is affected by unreliable camera poses and the error of image line detection. In contrast, our **AWR-Line-SLAM** generates accurate 3D line-based map. The line directions are regular and the endpoints of line segments are refined to the optimal positions. Due to limited space, for **Line-SLAM** and **VP-Line-SLAM**, please refer to the supplementary material.

VII. CONCLUSION

We proposed a novel algorithm that leverages the AW regularity for monocular SLAM. We estimate accurate AFs and also optimize camera poses and 3D map under the directional constraints. Experiments showed our method outperforms state-of-the-art ones in terms of accuracy and robustness.

References

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, 2015.

[3] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "PL-SLAM: Real-time monocular visual SLAM with points and lines," in *IEEE International Conference on Robotics and Automation*, 2017.

[4] T. Lee, C. Kim, and D. Cho, "A monocular vision sensor-based efficient SLAM method for indoor service robots," *IEEE Transactions on Industrial Electronics*, 2018.

[5] H. Li, J. Yao, J.-C. Bazin, X. Lu, Y. Xing, and K. Liu, "A monocular SLAM system leveraging structural regularity in Manhattan world," in *IEEE International Conference on Robotics and Automation*, 2018.

[6] J. M. Coughlan and A. L. Yuille, "Manhattan world: Compass direction from a single image by Bayesian inference," in *IEEE International Conference on Computer Vision*, 1999.

[7] J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher, "A mixture of Manhattan frames: Beyond the Manhattan world," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[8] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[9] H. Zhou, D. Zou, L. Pei, R. Ying, P. Liu, and W. Yu, "StructSLAM: Visual SLAM with building structure lines," *IEEE Transactions on Vehicular Technology*, 2015.

[10] J. Straub, O. Freifeld, G. Rosman, J. J. Leonard, and J. W. Fisher, "The Manhattan frame model—Manhattan world inference in the space of surface normals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[11] K. Joo, T.H. Oh, I.S. Kweon, and J.-C. Bazin, "Globally optimal inlier set maximization for Atlanta frame estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[12] J.-P. Tardif, "Non-iterative approach for fast and accurate vanishing point detection," in *IEEE International Conference on Computer Vision*, 2009.

[13] P. Denis, J. H. Elder, and F. J. Estrada, "Efficient edge-based methods for estimating Manhattan frames in urban imagery," in *European Conference on Computer Vision*, 2008.

[14] X. Lu, J. Yao, H. Li, and Y. Liu, "2-line exhaustive searching for real-time vanishing point estimation in Manhattan world," in *IEEE Winter Conference on Applications of Computer Vision*, 2017.

[15] J.-C. Bazin and M. Pollefeys, "3-line RANSAC for orthogonal vanishing point detection," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[16] J.-C. Bazin, Y. Seo, and M. Pollefeys, "Globally optimal consensus set maximization through rotation search," in *Asian Conference on Computer Vision*, 2012.

[17] J.-C. Bazin, Y. Seo, C. Demonceaux, P. Vasseur, K. Ikeuchi, I. Kweon, and M. Pollefeys, "Globally optimal line clustering and vanishing point estimation in Manhattan world," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[18] J.-C. Bazin, C. Demonceaux, P. Vasseur, and I. Kweon, "Rotation estimation and vanishing point extraction by omnidirectional vision in urban environment," *International Journal of Robotics Research*, 2012.

[19] R. Toldo and A. Fusiello, "Robust multiple structures estimation with J-Linkage," in *European Conference on Computer Vision*, 2008.

[20] J. Straub, N. Bhandari, J. J. Leonard, and J. W. Fisher, "Real-time Manhattan world rotation estimation in 3D," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

[21] P. Kim, B. Coltin, and H.J. Kim, "Low-drift visual odometry in structured environments by decoupling rotational and translational motion," in *IEEE International Conference on Robotics and Automation*, 2018.

[22] M. Antunes and J. P. Barreto, "A global approach for the detection of vanishing points and mutually orthogonal vanishing directions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[23] E. Tretyak, O. Barinova, P. Kohli, and V. Lempitsky, "Geometric image parsing in man-made environments," *International Journal of Computer Vision*, 2012.

[24] H. Lee, E. Shechtman, J. Wang, and S. Lee, "Automatic upright adjustment of photographs with robust camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[25] G. Schindler, F. Dellaert, and S. B. Kang, "Inferring temporal order of images from 3D structure," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[26] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, second edition, 2003.

[27] C. Bingham, "An antipodally symmetric distribution on the sphere," *The Annals of Statistics*, 1974.

[28] L. Magri and A. Fusiello, "T-Linkage: A continuous relaxation of J-Linkage for multi-model fitting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[29] C. V. Stewart, "MINPRAN: a new robust estimator for computer vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

[30] Z. Kukelova, M. Bujnak, and T. Pajdla, "Automatic generator of minimal problem solvers," in *European Conference on Computer Vision*, 2008.

[31] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[32] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, The MIT Press, 2005.

[33] B. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America A*, 1987.

[34] A. Bartoli and P. Sturm, "The 3D line motion matrix and alignment of line reconstructions," *International Journal of Computer Vision*, 2004.

[35] H. Li, J. Yao, X. Lu, and J. Wu, "Combining points and lines for camera pose estimation and optimization in monocular visual odometry," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

[36] H. Li, J. Zhao, J.-C. Bazin, L. Luo, J. Wu, and J. Yao, "Robust camera pose estimation via consensus on ray bundle and vector field," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018.

[37] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *European Conference on Computer Vision*, 2012.