© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Robust Global Structure from Motion pipeline with Parallax on manifold Bundle Adjustment and Initialization

Liyang Liu, Teng Zhang, Brenton Leighton, Liang Zhao, Shoudong Huang and Gamini Dissanayake[1]

***Abstract*—In this paper we present a novel global Structure from Motion (SfM) pipeline that is particularly effective in dealing with low-parallax scenes and camera motion collinear with the features that represent the environment structure. It is therefore particularly suitable in Urban SLAM, in which frequent road-facing motion poses many challenges to conventional SLAM algorithms. Our pipeline includes a recently explored bundle adjustment (BA) method that exploits a feature parameterization using parallax angle between on-manifold observation rays (PMBA). It is demonstrated that this BA stage has a consistently stable optimization configuration for features with any parallax and therefore low-parallax features can stay in reconstruction without pre-filtering. To allow practical usage of PMBA, we provide a compatible initialization stage in the SfM to initialize all camera poses simultaneously, exhibiting friendliness to collinear motion. This is achieved by simplifying PMBA into a hybrid graph problem of high connectivity yet small node set size, solved using a robust linear programming technique. Using simulations and a series of publicly available real datasets including "KITTI" and "Bundle Adjustment in the Large", we demonstrate the robustness of the position initialization stage in handling collinear motion and outlier matches, superior convergence performance of the BA stage in presence of low-parallax features, and effectiveness of our pipleline to handle many sequential or out-of-order urban scenes.**

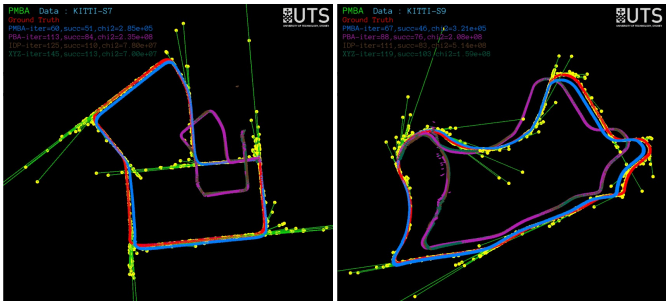
***Index Terms*—Mapping, SLAM**

I. Introduction

URBAN scenes can expose many challenges to algorithms in monocular Simultaneous Localisation and Mapping (SLAM) [1]. Lack of sufficient information for localization on urban roads, collinear vehicle motion, 3D reconstruction from images of diverse proximity scenes, all of these situations can cause SLAM algorithms to fail due to the existence of low parallax angle features.

The low parallax features come in the form of far away features or collinear features that lie along the direction of motion. Such feature points represented in the Euclidean XYZ form show highly non-Gaussian position uncertainty [2]. Being far or very far do not differ much in their image imprints although they contain strong orientation information. Feature positions thus triangulated become very unreliable. SLAM algorithms involving diverse proximity scenes are prone to numerical instability. In BA, the backend for full state estimation, the Gauss Newton process is applied requiring successive linearization and solving of normal equations for state update. Existence of low parallax features lead to singularity in the equation's information matrix [3]. The result is compromised estimation accuracy and prolonged convergence time.

Many visual systems mitigate this issue by applying ad-hoc data handling. In ORB-SLAM2 [4] far points are either discarded or receive a delayed triangulation. Other methods such as [5] process low and high parallax features selectively according to their stability at different stages of initialization. The separation strategy is generally regarded sub-optimal [2].



(*a*) Converged poses in S-7 (*b*) Converged poses in S-9

Fig. 1. Compare BAs on "KITTI" datasets S-7 and S-9: existence of collinear features (yellow dots) cause IDP (brown), XYZ (green) and PBA to be trapped in high cost region and unable to close the loop after long iterations. PMBA (blue) handles these features well and converges to a minimum very close to ground truth (red) with a fast pace .

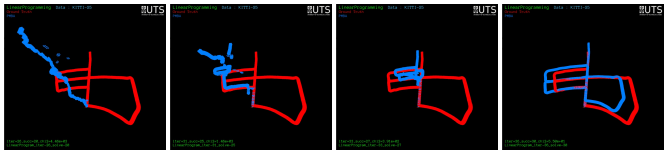


Fig. 2. Convergence sequence of proposed PGILP camera position initialization method, on KITTI dataset S5 (full of collinear motion and EG outliers): random to close-to-optimal, PGILP (blue), ground truth (red).

Feature Parametrization. Exploiting a different line of thinking, the low parallax issue can be fixed fundamentally using feature parametrization. Rather than representing features as XYZ points, Civera *et al* [2] proposed a paramtrization using the inverse depth of features (IDP) relative to their first observing cameras. IDP gives a single Gaussian distribution to cover features of all depth range and is effective in handing far features. However, collinear features remain to be a source of singularity in observation

Manuscript received: September, 10, 2018; Revised December, 26, 2018; Accepted February, 1, 2019.

This paper was recommended for publication by Editor Cyrill Stachniss upon evaluation of the Associate Editor and Reviewers' comments.

[1]All authors are with the Centre for Autonomous Systems (CAS), University of Technology Sydney, Australia. liyang.liu@uts.edu.au

Digital Object Identifier (DOI): see top of this page.

Jacobians [3]. To provide a unified solution to nearby, far and collinear features, Zhao *et al* [3] proposed the parallax angle representation which defines a feature with three highly observable angles (elevation, azimuth and parallax) without involving the unobservable depth. The BA thus formulated (PBA) contains sufficient information for state estimation hence shows low chance of degeneracy. The PMBA stage in our SfM, first introduced in [6] is a further extension of PBA onto the manifold domain. Using a formulation faithfully mimicing the image formation process, PMBA exhibits a well-bounded information matrix hence state update is always observable. See Fig. 1 for illustration of comparison of BA's convergence performance.

Position Initialization. Being highly non-convex, BA requires good initial estimates for convergence to global minimum [7]. The common approaches are categorized as incremental and global. The incremental system, such as Bundler [8] and ORB-SLAM [4], consists of adding camera poses incremently to a pre-existing map and performing intermediate BA's at each insertion. Being slow and easily stuck in picking the right starting map are the known issues of incremental methods. The alternative strategy is global initialization where all camera poses are initialised simultaneously. Structure recovery is left to the BA stage. A global SfM pipeline consisting of pose initialization then a single BA invocation show higher efficiency and accuracy [9]–[15]. This is what this paper addresses.

In global initialization, camera orientation estimation can been robustly computed [16]. Position estimation can be challenging due to its reliance on noisy and outlier-prone pairwise Epipolar Geometry (EG) without scale information.

Related work in position initialization can be viewed along two main perspectives. The first perspective is based on solution stability and robustness. As a first attempt [9] [17], position estimation was solved as a linear least squares problem by minimizing either cross product or orthogonal projection between predicted relative translations and measured directions. The linear methods are intrinsically algebraic methods without geometrical meaning [10], can exhibit extreme bias and must be weighted iteratively [11] before use. Further, they easily lead to the trivial but incorrect solution where the camera locations cluster around a few points [15], and fail easily for slightly noisy or larger data. The instability issue was then tackled in [10] [11] with non-linear optimization methods. [10] formulated position estimation as minimax optimization under the l_∞-norm based on well-conceived geometrical metric. Wilson and Snavely [11] proposed a non-linear cost function (WsNonLin) that minimizes chordal distance of translation directions, and achieves convergence with the LM solver. Other research efforts [13] [14] solved the clustering issue by adding constraints to the linear formulation, retaining efficiency by exploiting linear relations. Vulnerability to outlier translation directions remains to be a big problem in non-linear and constrained linear methods. [11] proposed a pre-filtering heuristic known as 1DSfM to detect and discard outlier EG-pairs, which may result in information loss. The robustness issue is eventually resolved in [15] with its convex Least Unsquared Deviation (LUD) method on translation lengths. This convex problem requires a non-trivial solver by successive quadratic programming approximation with self-adjusted weights. We concern for scalability and simplicity.

The second perspective of position estimation is the choice of graphical representation of the underlying problem. This relates to information embedded in the graph and affects its ability to solve for collinear poses. The EG-pair formation describes an epipolar graph $G_t = (V_t, E_t)$ with nodes $V_t = \{1, 2, ..., n\}$ representing camera positions $\mathbf{P}_i$ and edges $(i,k) \in E_t$ for translation directions $\mathbf{p}_{ik} = \frac{\mathbf{R}_i^T(\mathbf{P}_i - \mathbf{P}_k)}{\|\mathbf{P}_i - \mathbf{P}_k\|}$. The node set $\{\mathbf{P}_i\}$ are to be solved. Epipolar graph problem with collinear poses present is known to show degeneracy due to lack of sufficient constraint [12]. Even in absence of collinear motion, the problem may still be ill-posed when the number of nodes and edges fail to satisfy the minimum connectivity condition known as parallel rigidity [15]. This imposes limitations for methods such as LUD, WsNonLin (in pose-only form) and linear least squares. The alternative graph domain is the triplet graph and has been successfully explored in [14]. This graph consists of many strongly connected camera-triplets where two triplets share a common edge, scale can be recovered due to the strong (camera-to-camera) connectivity. However, as observed in [12], such a scheme may produce distorted reconstruction when the strong image association does not exist. In an attempt to fix collinearity, [11] suggested an extended WsNonLin mode by including judiciously selected features into its graph node set. Unfortunately this "increases problem size, with diminishing returns" [11]. Further, WsNonLin requires abundant image association [12] and is more suited for internet images than sequential scenario. A hybrid pose graph form was proposed in [12] where a set of judiciously selected feature observations are included to the graph as indirect links between pose nodes, yet does not require solving the feature nodes. With boosted connectivity, the hybrid graph is able to uncover scale in collinear motion.

We recently explored the hybrid graph idea in [6], where positions are estimated with a constrained least square and non-linear optimizer two-stage initializer (CLS-NonLin). However, the linear stage involving cross-product showed clustering effect and the overall estimator is highly sensitive to outliers. We propose yet another hybrid graph based position estimator, addressing all above issues: robustness, simplicity and friendliness to collinear motion.

Contributions and Paper Structure. This paper builds on previous work [6] and presents a complete global SfM pipeline robust to low-parallax scenes. First, we give a review of PMBA (in Section II) on how its formulation and measurement model lead to complete stability in presence of low parallax features with superior convergence properties. As a new contribution in this topic, we propose (In Section III) a PMBA compatible position estimation scheme using linear programming (referred to as PGILP) with outlier robustness and implementation easiness taken into consideration. We achieve collinear motion friendliness by using all feature observations to improve graph connectivity without explicitly

solving for features, thus keeping a reasonable problem size. We show that under noise-free condition the intialization method can recover positions exactly. All this properties are achieved without discarding information, specifically useful to urban SLAM. Using KITTI dataset containing largely street-view scenes, we demonstrate effectiveness of our SfM pipeline at every processing stage (In Section IV).

Notations. Throughout this paper, we use the term $\mathbf{T}_i = (\mathbf{R}_i, \mathbf{P}_i) \in \mathbb{SE}(3)$ to represent the i'th camera pose. $\mathbb{F}$ denotes the set of indices for all features. $\mathbb{T}_j$ denotes to the set of indices for camera poses at which feature j is observed.

II. REVIEW OF PARALLAX BUNDLE ADJUSTMENT ON MANIFOLD THEORY

To address BA instability in presence of low parallax features, a fundamental solution of formulating the BA problem by exploiting observations rays in the manifold domain has been recognized in [6]. Here we present a brief summary of the PMBA theory and its fast convergence capability.

A. Feature parameterization, measurement model and optimization formulation

With a set of images indexed $\{1,..,M\}$ and a set of feature tracks $\{1,..,N\}$ collected therein, a feature point (indexed j) is observed from the camera set $\mathbb{T}_j$, a camera (indexed i) observes the feature at pixel imprint $\mathbf{u}_{j,i}$, the PMBA problem estimates the all-on-manifold state vector:

- $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{F}})$
- $\boldsymbol{\mathcal{T}} = \{(\mathbf{R}_i, \mathbf{P}_i)\}_{i=1,\cdots,M}$, the full set of camera poses
- $\boldsymbol{\mathcal{F}} = \{\mathcal{F}_j \in \mathbb{M}^3\}_{j=1,\cdots,N}$, parallax feature parameters,

using measurements $\mathbf{V} = \{\mathbf{v}_{j,i}\}$ of locally observed ray directions:

- $\mathbf{v}_{j,i} = \frac{\mathbf{K}^{-1}\mathbf{u}_{j,i}}{\|\mathbf{K}^{-1}\mathbf{u}_{j,i}\|}$, an observation ray direction
- K, the camera calibration matrix

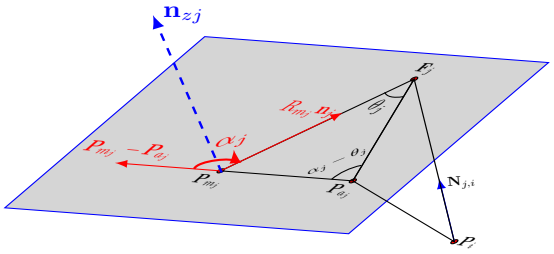


Fig. 3. Feature $\mathbf{F}_j$ is anchored by cameras at $\mathbf{P}_{m_j}$ and $\mathbf{P}_{a_j}$ with parallax angle θ_j. A third camera $\mathbf{P}_i$ sees $\mathbf{F}_j$ along direction of ray $\mathbf{N}_{j,i}$. The plane formed by $\mathbf{F}_j$, $\mathbf{P}_{m_j}$, $\mathbf{P}_{a_j}$ has a normal at $\mathbf{n}_{zj}$

A feature point is uniquely positioned by light rays from two of its observing cameras and the parallax angle in-between. The two positioning camera poses are referred to as the main anchor $\mathbf{T}_{m_j}$ and the associate anchor $\mathbf{T}_{a_j}$ as shown in Fig. 3. In the context of PMBA, the ray direction vector in the main anchor's frame $\mathbf{n}_j \in \mathbb{R}^3$, and the parallax angle θ_j between the anchoring rays constitute the feature parameters $\mathcal{F}_j$. In manifold form, to avoid singularity in angular representation, $\mathcal{F}_j$ is over-parameterized as

$$\mathcal{F}_j = [\,\cos\theta_j,\ \sin\theta_j,\ \mathbf{n}_j\,] \tag{1}$$

With a small perturbations in parallax angle $\delta\theta_j \in \mathbb{R}^1$ and ray direction $\delta\mathbf{n}_j \in \mathbb{R}^2$, the on-manifold feature takes following retraction operation:

$$\mathcal{F}_j \boxplus \delta\mathcal{F}_j = \begin{bmatrix} \cos(\theta_j + \delta\theta_j) \\ \sin(\theta_j + \delta\theta_j) \\ \mathrm{Exp}(\mathbf{A}_{\mathbf{n}_j}\delta\mathbf{n}_j)\mathbf{n}_j \end{bmatrix}, \quad \begin{array}{l} \delta\mathcal{F}_j = \left[\delta\theta_j, \delta\mathbf{n}_j\right] \in \mathbb{R}^3, \\ \mathbf{A}_{\mathbf{n}_j} \in \mathbb{R}^{3x2}, \\ \left[\mathbf{A}_{\mathbf{n}_j}\ \mathbf{n}_j\right] \in \mathbb{SO}(3) \end{array} \tag{2}$$

where the columns of matrix $\mathbf{A}_{\mathbf{n}_j}$ spans the left null-space of $\mathbf{n}_j$ (see [6]), Exp() is the exponential map for $\mathbb{SO}(3)$.

The PMBA parametrization describes the feature's geometrical relationship to its anchors, with this knowledge its position in Euclidean space can be computed using sine rule:

$$\mathbf{F}_j(\mathcal{F}_j) = \frac{\sin(\alpha_j - \theta_j)}{\sin(\theta_j)}\|\mathbf{P}_{m_j} - \mathbf{P}_{a_j}\|\mathbf{R}_{m_j}\mathbf{n}_j + \mathbf{P}_{m_j} \tag{3}$$

where $\mathbf{R}_{m_j}\mathbf{n}_j$ is the ray direction expressed in global frame, α_j is the angle between $\mathbf{R}_{m_j}\mathbf{n}_j$ and vector $\overrightarrow{P_{a_j}P_{m_j}} = \mathbf{P}_{m_j} - \mathbf{P}_{a_j}$.

To estimate the direction of observed light ray, [6] uses the concept of a length-scaled ray vector $\mathbf{N}_{j,i}$ that sees $\mathbf{F}_j$ from the i'th camera position $\mathbf{P}_i$. $\mathbf{N}_{j,i}$ is a function of state vector set $\mathcal{X}_{j,i} = \{\mathbf{P}_{m_j}, \mathbf{P}_{a_j}, \mathbf{P}_i, \mathbf{R}_{m_j}, \mathcal{F}_j\}$:

$$\begin{aligned} &\mathbf{N}_{j,i}(\mathcal{X}_{j,i}) = \sin(\theta_j)(\mathbf{F}_j - \mathbf{P}_i) \\ &= \sin(\alpha_j - \theta_j)\|\mathbf{P}_{m_j} - \mathbf{P}_{a_j}\|\mathbf{R}_{m_j}\mathbf{n}_j + \sin(\theta_j)(\mathbf{P}_{m_j} - \mathbf{P}_i) \end{aligned} \tag{4}$$

The scale factor $\sin(\theta_j)$ helps to avoid numerical instability in ray length calculation as $\theta_j \to 0$.

This gives rise to the following PMBA formulation:

$$\min_{\boldsymbol{\mathcal{X}}} \sum_{j,i\in\mathbb{T}_j} \| \frac{\mathbf{N}_{j,i}(\mathcal{X}_{j,i})}{\|\mathbf{N}_{j,i}(\mathcal{X}_{j,i})\|} - \mathbf{R}_i\mathbf{v}_{j,i}\|^2, \qquad \boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{F}}) \tag{5}$$

where $\mathbf{R}_i\mathbf{v}_{j,i}$ brings the ray direction to global frame.

As explained in [6], conventional BA's directly use 2D pixels as measurement. They do not differentiate frontal or behind-camera scenario and exhibit "many local minima and saddle points". Further, pixel prediction is done via homogeneous normalization on local coodinates, causing discontinuity in BA as the Z-ordinate cannot be zero. The PMBA's measurement model takes a 3D form that naturally addresses the behind-camera cases and continuity issue. Thus PMBA has the ability to correct many errorneous estimates.

B. Convergence properties of PMBA

[6] provided proof that the PMBA formulation comes with a consistently invertible Hessian at every stage of optimization thus such a system is locally observable. This is a direct consequence of the on-manifold parallax parametrization and its compatible ray-direction measurement model.

In conventional BA, existence of low-parallax features results in degeneracy in the Hessian, causing Gauss Newton (GN) solvers to fail or slow convergence in the Levenberg Macquardt (LM) method due to self-adjustable damping to suppress singularity.

PMBA's high observability implies that fast optimization method Dog-Leg (DL) can be used to solve the problem [6]. In DL, each state increment is a linear summation of GN increment and Steepest Descent (SD), it is therefore faster than GN or LM. Further, within an iteration's fine-tuning steps, the same inverted Hessian is used for all DL re-tries, wheras LM requires inversion on its augment Hessian after each damping adjustment.Hence in presence of low-parallax features, PMBA exhibits superior convergence behaviour than conventional BA's.

III. PMBA COMPATIBLE GLOBAL INITIALIZATION

In this section, we derive a PMBA-compatible initialization strategy for our SfM pipleline.

We show how parallax features can be reliably initialized, and the PMBA problem can be easily converted into a convex pose-graph problem which robustly estimates camera positions by minimizing l_1-norm of observation rays. We prove that in noise-free condition camera positions can be recovered exactly by this strategy. This pipeline of global initialization and final BA are illustrated in Fig. 4.

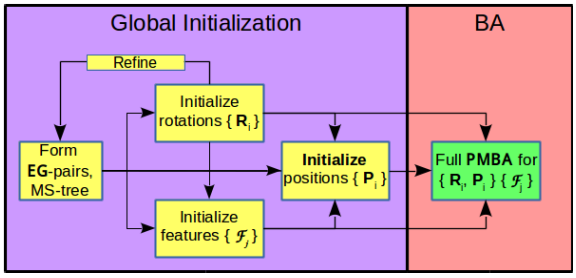


Fig. 4. Global SfM pipeline: initialization (purple) + BA (red).

A. Orientation and feature initialization

As shown in [16], camera orientations can be initialized reliably. We start by forming a maximal set of two-view matches from input data, which can be either sequential or out-of-order, then compute the associated Epipolar Geometry. Using match size as score we build a maximum spanning (MS) tree of EG-pairs. From the connected edges we form rotation priors and apply the rotation averaging method from [16]. The result rotations are highly accurate and are robust against outlier EG-pairs.

Initializing PMBA features boils down to selecting good parallax angles and associated anchors. We list the detailed procedure in Alg. 1. This algorithm is simple and only requires knowledge of rotations, avoiding expensive/unreliable triangulation method in conventional methods. Hence parallax feature values are also highly accurate.

B. Position initialization

With rather accurate rotations and parallax features, the remaining problem in PMBA becomes a position registration problem. Substituting rotations $\{\mathbf{R}_i\}$ and features $\{\mathcal{F}_j\}$ into PMBA formulation (5) gives a non-linear optimization problem for positions $\{\mathbf{P}_i\}$. We now turn this into a convex problem by exploiting the fact that the afore-mentioned ray

Algorithm 1: Feature j initialization.

Input: $\{\mathbf{v}_{j,i}\}_{i\in\mathbb{T}_j}$, $\{\bar{\mathbf{R}}_i\}_{i\in\mathbb{T}_j}$;
Output: m_j, a_j, $\bar{\mathcal{F}}_j$;
$m_j = \mathbb{T}_j[1]$; $\bar{\mathbf{n}}_j \leftarrow \mathbf{v}_{m_j,j}$;
$a_j \leftarrow m_j$; $\bar{\theta}_j \leftarrow 0$;
$l \leftarrow 2$; $k = \mathbb{T}_j[l]$
while $\sin\bar{\theta}_j < 0.45$ **do**
  $\Theta = atan2(\|\bar{\mathbf{R}}_{m_j}\bar{\mathbf{n}}_j \times \bar{\mathbf{R}}_k\mathbf{v}_{j,k}\|, \bar{\mathbf{R}}_{m_j}\bar{\mathbf{n}}_j \cdot \bar{\mathbf{R}}_k\mathbf{v}_{j,k})$;
  if $\sin\Theta > \sin\bar{\theta}_j$ **then**
    $\bar{\theta}_j \leftarrow \Theta$;
    $a_j \leftarrow k$;
  $l \leftarrow l+1$;
  $k \leftarrow \mathbb{T}_j[l]$
$\bar{\mathcal{F}}_j \leftarrow (\cos\bar{\theta}_j, \sin\bar{\theta}_j, \bar{\mathbf{n}}_j)$;

vector $\mathbf{N}_{j,i}$ can be expressed in a linear function of positions. As illustrated in Fig. 3, the non-linear term $\|\mathbf{P}_{m_j} - \mathbf{P}_{a_j}\|\mathbf{R}_{m_j}\mathbf{n}_j$ in (4) is equivalent to rotating $\overrightarrow{\mathbf{P}_{m_j}\mathbf{P}_{a_j}}$ about axis $\mathbf{n}_{zj}$ towards $\mathbf{R}_{m_j}n_j$ with angle $(\pi - \alpha_j)$. The linearized expression is:

$$\begin{aligned}\bar{\mathbf{N}}_{j,i} =& \sin(\bar{\alpha}_j - \bar{\theta}_j)\mathrm{Exp}(\bar{\mathbf{n}}_{zj}(\pi - \bar{\alpha}_j))(\mathbf{P}_{a_j} - \mathbf{P}_{m_j}) \\ &+ \sin(\bar{\theta}_j)(\mathbf{P}_{m_j} - \mathbf{P}_i),\end{aligned} \tag{6}$$

Here, axis $\mathbf{n}_{zj}$ is normal to the plane constituting vector $\overrightarrow{\mathbf{P}_{m_j}\mathbf{P}_{a_j}}$ and global ray $\mathbf{R}_{m_j}\mathbf{n}_j$, can be computed as $\frac{(\mathbf{P}_{a_j} - \mathbf{P}_{m_j})}{\|\mathbf{P}_{a_j} - \mathbf{P}_{m_j}\|} \times (\mathbf{R}_{m_j}\mathbf{n}_j)$.

Now, expressing $\bar{\mathbf{N}}_{j,i}$ in matrix form, we get

$$\begin{aligned}\bar{\mathbf{N}}_{j,i} &= \mathbf{A}_{N_{j,i}}\mathbf{x}_{j,i} \\ \mathbf{x}_{j,i} = \begin{bmatrix}\mathbf{P}_{m_j} \\ \mathbf{P}_{a_j} \\ \mathbf{P}_i\end{bmatrix}, &\quad \mathbf{A}_{N_{j,i}} = \begin{bmatrix}\mathbf{A}^{(1)}_{N_{j,i}} \\ \mathbf{A}^{(2)}_{N_{j,i}} \\ \mathbf{A}^{(3)}_{N_{j,i}}\end{bmatrix}\end{aligned} \tag{7}$$

where the rows in matrix $\mathbf{A}_{N_{j,i}}$ are

$$\begin{aligned}\mathbf{A}^{(1)}_{N_{j,i}} &= -\sin(\alpha_j - \theta_j)\mathrm{Exp}(\mathbf{n}_{zj}(\pi - \alpha_j)) + \sin(\theta_j)\mathbf{I}_3 \\ \mathbf{A}^{(2)}_{N_{j,i}} &= \sin(\alpha_j - \theta_j)\mathrm{Exp}(\mathbf{n}_{zj}(\pi - \alpha_j)) \\ \mathbf{A}^{(3)}_{N_{j,i}} &= -\sin(\theta_j)\mathbf{I}_3\end{aligned} \tag{8}$$

Note that the ray scaling factor $\sin(\theta_j)$ remains in the linear relation and serves as a weight factor to enhance stability of matrix $\mathbf{A}_{N_{j,i}}$ (more on this in Section III-C).

In previous work [6] we proposed the CLS-NonLin method which minimizes the cross product between $\bar{\mathbf{N}}_{j,i}$ and $\mathbf{v}_{j,i}$ with a linear constraint. This linear relation shows skewed errors and is sensitive to outlier. To ensure balanced error distribution, we introduce an extra ray scale variable $\lambda_{j,i} := \|\mathbf{F}_j - \mathbf{P}_i\|$ into (7) to reconcile the variable ray length.

$$\underset{\{\mathbf{x}_{j,i}\},\{\lambda_{j,i}\}}{\text{minimize}} \quad \sum_{i\in\mathbb{T}_j,j} \|\mathbf{A}_{N_{j,i}}\mathbf{x}_{j,i} - \lambda_{j,i}\mathbf{R}_i\mathbf{v}_{j,i}\|^2 \tag{9}$$

Solving (9) directly does not guarantee cheirality condition as $\lambda_{j,i}$ can be negative. Further, when $\lambda_{j,i} = 0$, we reach the trivial solution $\mathbf{P}_{m_j} = \mathbf{P}_{a_j} = \mathbf{P}_i$, clustering occurs regardless of feature angles or rotations. A constraint of positive $\lambda_{j,i}$

is therefore necessary and should have a repulsion effect on clustering cameras.

To enforce robustness to outlier EG pairs or incorrect matches, we minimize the sum of l_1-norm of every ray length error in the final cost function. Putting all this requirements into (7) and (9) we have

$$\begin{aligned} &\underset{\{\mathbf{P}_i\},\{\lambda_{j,i}\}}{\text{minimize}} \sum_{i\in\mathbb{T}_j,j} \| \mathbf{A}_{N_{j,i}}[\mathbf{P}_{m_j}^T \mathbf{P}_{a_j}^T \mathbf{P}_i^T]^T - \lambda_{j,i}\mathbf{R}_i\mathbf{v}_{j,i} \|_1 \\ &\text{subject to} \quad \lambda_{j,i} \geq 1, \qquad j \in \mathbb{F}, i \in \mathbb{T}_j, \\ &\qquad\qquad \mathbf{P}_1 = (0,0,0). \end{aligned} \tag{10}$$

The last two conditions removes scale and translation ambiguity of the solution. Now, for ease of solving, we introduce the slack variable $\boldsymbol{\gamma}_{j,i} \in \mathbb{R}^3_+$ to represent l_1 residual and transform (10) into the following equivalent linear program:

$$\begin{aligned} &\underset{\{\mathbf{P}_i\},\{\lambda_{j,i}\},\{\boldsymbol{\gamma}_{j,i}\}}{\text{minimize}} \sum_{i\in\mathbb{T}_j,j} \mathbf{1}^T\boldsymbol{\gamma}_{j,i} \\ &\text{subject to:} \\ &| \mathbf{A}_{N_{j,i}}^{(k)}[\mathbf{P}_{m_j}^T \mathbf{P}_{a_j}^T \mathbf{P}_i^T]^T - \lambda_{j,i}\mathbf{R}_i^{(k)}\mathbf{v}_{j,i} | \leq \boldsymbol{\gamma}_{j,i}^{(k)}, \quad k = 1,2,3, \\ &\lambda_{j,i} \geq 1, \qquad j \in \mathbb{F}, i \in \mathbb{T}_j, \\ &\mathbf{P}_1 = (0,0,0). \end{aligned} \tag{11}$$

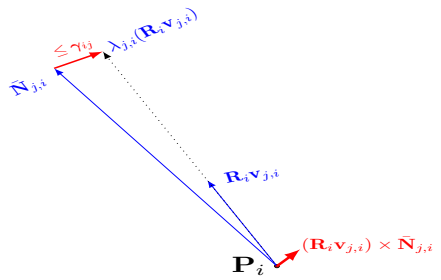


Fig. 5. $\mathbf{R}_i\mathbf{v}_{j,i}$ is the measured ray direction, $\lambda_{j,i}$ is the ray length, $\boldsymbol{\gamma}_{j,i}$ is the l_1 residual. PGILP simplifies PMBA into a position estimation problem: $\sum \mathbf{1}^T\boldsymbol{\gamma}_{j,i}$. This error function is geometrically meaningful compared to the cross-product based CLS-NonLin error: $(\mathbf{R}_i\mathbf{v}_{j,i}) \times \bar{\mathbf{N}}_{j,i}$.

Remark 1: The objective in (10), being an l_1-norm of affine functions, is therefore a convex function. The inequality constraints define a convex set $S_\lambda \subset \mathbb{R}^{|\{j\in\mathbb{F}, i\in\mathbb{T}_j\}|}$, an intersection of half spaces. The optimization problem (10) is therefore convex, convergence to a global minimum is guaranteed from any initial guess.

For each feature observed in more than two poses, every extra observation of the feature effectively adds a new edge to the graph, significantly improving graph connectivity. With feature observations far exceeding that of pose nodes, the graph can be assumed parallel rigid in general (unique topology).

Remark 2: Since the trivial solution of $\lambda_{j,i} = 0, j \in \mathbb{F}, i \in \mathbb{T}_j$ is not in the feasible set, in absence of scale ambiguity, at least one of the constraints needs to be active at the optimum solution, the optimal $\lambda_{j,i}$'s must satisfy $\min_{j,i} \lambda_{j,i} = 1$.

Remark 3: The objective of (10) is based on feature observation rays. A feature point observed in the main, associate and any third pose encodes the position ratio between the three cameras. Collinear poses can therefore be solved from problem (10) without degeneracy.

C. Theoretical analysis

We now give an analysis on the initialization scheme.

Proposition 1: Given the noiseless EG-pairs $\{\mathbf{p}_{ik}\}$, camera rotations $\{\mathbf{R}_i\}$ and observation directions $\{\mathbf{v}_{j,i}\}$ of sufficient size that render the associated hybrid graph parallel rigid, our PGILP solver recovers the locations $\{\mathbf{P}_i\}$ exactly in the sense that any solution is congruent to $\{\mathbf{P}_i\}$.

Proof: From Alg. 1 noise-free $\{\mathbf{R}_i\}$ and $\{\mathbf{p}_{ik}\}$ produce noise-free feature parameters $\{\mathcal{F}_j\}$, hence noise-free $\{\theta_j\}$, $\{\alpha_j\}$ and $\{\mathbf{n}_{zj}\}$. From (8), each $\mathbf{A}_{N_{j,i}}$ is also noise-free. Since $\{\mathbf{P}_i\}$ are the ground truth location of cameras, from Eq. (7) we obtain noise-free ray length $\lambda_{j,i} = \|\mathbf{A}_{N_{j,i}}\begin{bmatrix}\mathbf{P}_{m_j}^T & \mathbf{P}_{a_j}^T & \mathbf{P}_i^T\end{bmatrix}^T\|, i \in \mathbb{T}_j$. Now substitute noise-free $\{\mathbf{P}_i\}$, $\{\lambda_{j,i}\}$, $\{\mathbf{R}_i\}$ and $\{\mathbf{v}_{j,i}\}$ into the objective function in (10), we obtain a cost value of zero. We then scale $\lambda_{j,i}$ to ${\lambda_{j,i}}^s$ such that $\min_{j,i} {\lambda_{j,i}}^s = 1$ to satisfy the first constraint, hence obtain the scaled positions $\{\mathbf{P}_i^s\}$. $\{\mathbf{P}_i^s\}$ is therefore an optimal solution to PGILP as well as to the original PMBA problem. Since the graph is parallel rigid, $\{\mathbf{P}_i^s\}$ has to be congruent to any other solution of PGILP. ■

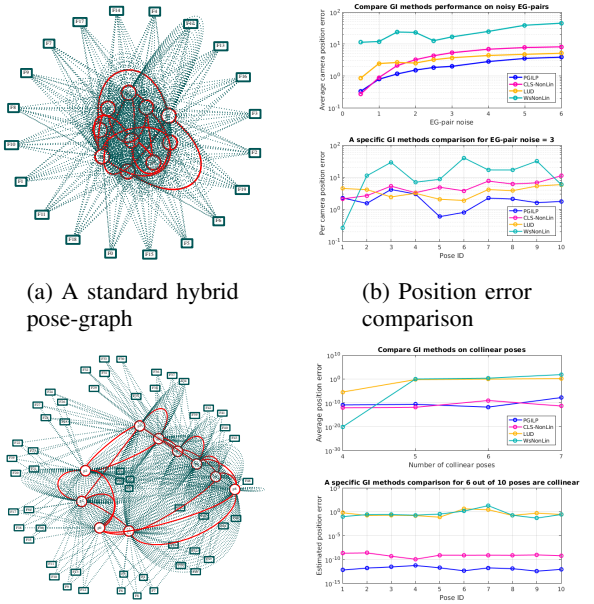
The PGILP formulation shares similarity to the Least Unsquared Deviation (LUD) convex model in [15]. LUD minimizes the sum of pairwise translation error in l_2-norm (without squaring), which is in effect a form of l_1-norm error. In comparison our method minimizes the l_1-norm of observation ray errors. Both methods are convex and robust to outliers as their l_1-norm based formulation promotes sparsity in residual errors. The difference, however comes in two-folds. First LUD takes inputs solely from the epipolar graph, and shows degeneracy with collinear poses and fails completely for graphs that do not satisfy the parallel rigidity criteria. Our graph uses more information: the edge set includes not only the EG-pairs, but also all feature observations (low-occurrence and low-parallax ones) indiscriminately. Since feature points are not included in the graph node set, the problem scale remains reasonable. This hybrid structure readily handles collinear motion and satisfies the parallel rigidity condition. Secondly, in terms of problem simplicity, LUD is a type of convex Second-Order Cone Programming (SOCP) problem which is intricate to solve. The iteratively reweighted least squares (IRLS) solver used in [15] applies successive "smooth regularization of l_2 norm penalty" to guide search towards low-error estimates, this may result in poor conditioning and other convergence issues. Further, LUD requires locations to be i.i.d. Gaussian, a condition difficult to meet in urban SLAM due to diverse camera and feature positions. The observation ray error in our PGILP does not experience the uneven error distribution situation, as the linear operator $\mathbf{A}_{N_{j,i}}$ in its objective is already weighted by feature parallax $\sin(\theta_j)$ that naturally equalizes uncertainty on variable sized observation rays. Thus PGILP can be easily implemented with any linear programming tool and achieves better convergence.

IV. EXPERIMENTS

In this section we evaluate every stage of the proposed SfM pipeline: initialization, PMBA; and give an overall evaluation on the full system. All experiments except simulation are implemented in C++ and tested on an 8-core Intel-i7 computer.

A. Evaluation on robust initialization performance

1) Simulation test on Global Initialization: A group of selected Global Initialization (GI) methods are tested on their robustness to noise and ability to handle collinear motion. The



(a) A standard hybrid pose-graph

(b) Position error comparison

(c) A graph where poses 3-8 are collinear

(d) Position error comparison

Fig. 6. Pose graph showing 10 pose nodes(red), feature nodes (green) and edges. Pose 3-8 are collinear. Red edges are EG-pairs, they are the only source of inputs in Epipolar graph based methods. The feature induced edges (green) are abundant, they constitute constraints in our PGILP initialization method

test is set to contain 10 camera poses with a set of covisible features and corresponding pairwise matches, as illustrated in the hybrid graph of Fig. 6(c). The GI schemes covered in this test include: PGILP, CLS-NonLin [6], LUD [15], WsNonLin [11]. For robustness test, we collected GI's result error rates while progressively increasing noise in the EG-pair directions. For the collinearity test, we kept the EG-pair size constant while increasing number of collinear poses. The GI error rates for both scenarios are plotted in Fig. 6 (b,d). The GI scheme's attributes as shown from test results match our prediction and are summarized in Table I. Our proposed PGILP method has the most positive attributes.

TABLE I
PROS AND CONS IN GLOBAL INITIALIZATION SCHEMES

GI method	PGILP (ours)	CLS-NonLin [6]	LUD [15]	WsNonLin [11]
Robust	yes	no	yes	no
Collinearity friendly	yes	yes	no	no

2) Global Initialization experiments on benchmark datasets (sequential): In this experiment, we compared various GI methods (LUD [15], WsNonLin [11], CLS-NonLin, [6] and PGILP) using selected sequences of benchmark data store KITTI [18] and Malaga [19], all of which are collected from vehicles moving along straight-line roads. Our test procedure follows the flow shown in Fig. 4 – purple region. We performed crude pairwise-matching to produce EG-pairs with many outliers. This data source exposes many challenges for tested GI algorithms. From the EG-pairs we built an MT-tree and compute camera orientations according to [16]. The tree root camera is chosen to be the reference frame for all systems (including GT). For a fair position comparison, we scale every method's result by its median estimate to GT position ratio for best GT alignment. The full accuracy test results are given in Table II and selected trajectories are plotted in Fig. 7. These results clearly show that PGILP is the most accurate position estimator.

TABLE II
GLOBAL INITIALIZATION COMPARISON

		LUD		WsNonLin		CLS-NonLin		PGILP	
Dataset	# poses	e_{men}	e_{med}	e_{men}	e_{med}	e_{men}	e_{med}	e_{men}	e_{med}
KITTI-S0	2349	0.06	0.04	3.40	0.21	19.2	0.26	**0.05**	**0.03**
KITTI-S5	1624	0.16	0.17	0.82	0.24	1.07	0.22	**0.14**	**0.14**
KITTI-S7	1041	0.38	0.12	17.0	1.44	13.4	2.44	**0.18**	**0.10**
KITTI-S9	1591	0.60	0.03	8.99	0.72	13.0	1.23	**0.05**	**0.02**
Malaga	171	0.98	0.72	1.63	1.29	0.26	0.19	**0.21**	**0.14**
Campus2L	1016	0.09	0.06	108	0.26	0.54	0.14	**0.03**	**0.02**

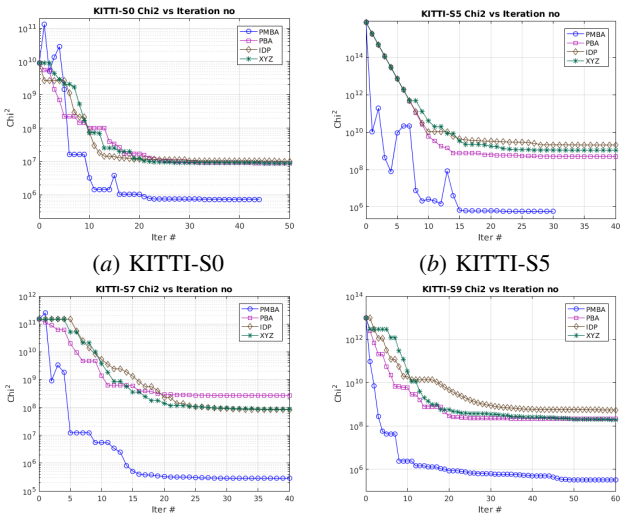
e_{men} denotes average distance between estimated camera position to ground truth (GT), e_{med} denotes median distance

B. Experiments on PMBA stage

To evaluate the BA stage of our propsed SfM pipeline, we performed convergence test comparing PMBA against other BAs including PBA, IDP and XYZ on all KITTI sequences used previously in Section III. These data contain many low-parallax features so present good case for BA evaluation. We run the PGILP method with premature termination to produce erroneous initial values for testing the BA stage, and investigate the convergence performance of PMBA with single thread running mode. We use DL optimization for PMBA and the stable LM for conventional BA's (DL not considered here due to instability). The conventional BA's are programmed to use uv-based Chi2 error. PMBA, on the other hand, requires ray direction error in the optimization implementation. We solve this problem by intercepting the current estimate at each iteartion step then compute and record the corresponding uv errors. All BA's thus are compared on a common error metric. This scheme is not optimal for PMBA, yet is the only convincing way to demonstrate PMBA's advantage over conventional methods. The collected results in Table III. Selected convergence plots are shown in Fig. 8. In terms of robustness, efficiency and accuracy, we found PMBA the best performer in all tests. As illustrated in Fig. 1, the PMBA results and GT are very close, suggesting convergence to global minimum, yet conventional BAs give significant error, a sign of converging to local minimum.

TABLE III
COMPARISON OF CONVERGENCE PERFORMANCE FOR PMBA (DL), PBA (LM), XYZ-BA (LM), IDP-BA (LM)

Dataset	Test-type	# Pose / # Feat / # Obsv	# Equation solving / # Iteration	Final Chi2	Time[sec]
KITTI -S0	PMBA	2349	**45 / 21**	**7.3E+5**	**133.2**
	PBA	/275,751	223 / 155	8.1E+6	799.3
	IDP	/1,015,187	201 / 151	9.3E+6	629.0
	XYZ		201 / 190	8.8E+6	554.9
KITTI -S5	PMBA	1624	**31 / 19**	**5.8E+5**	**122.2**
	PBA	/281,012	178 / 135	4.9E+8	673.2
	IDP	/1,024,677	201 / 166	1.5E+8	647.1
	XYZ		201 / 148	4.7E+8	517.9
KITTI -S7	PMBA	1045	**61 / 52**	**2.8E+5**	**381.8**
	PBA	/435,212	201 / 153	2.3E+8	1236.7
	IDP	/1,469,301	201 / 173	7.5E+7	1133.1
	XYZ		201 / 168	1.0E+7	1049.9
KITTI -S9	PMBA	1591	**75 / 54**	**3.2E+5**	**339.0**
	PBA	/561,574	201 / 169	1.9E+8	1342.7
	IDP	/1,648,120	201 / 151	5.0E+8	1156.8
	XYZ		201 / 163	1.3E+8	1095.5



(*a*) KITTI-S0 (*b*) KITTI-S5

(*c*) KITTI-S7 (*d*) KITTI-S9

Fig. 8. Convergence plots for PMBA, PBA, IDP and XYZ

C. Evaluation of SfM pipeplines on large-scale datasets

Finally, we evaluate performance of our complete SfM pipeline and compare with an incremental SfM released in OpenMVG [20]. Our pipeline consists of an initialization stage and a single PMBA call, as illustrated in Fig. 4.

We choose selected datasets from the "Bundle Adjustment in the Large" (BAL) database [21] for out-of-order tests and KITTI sequences for sequential tests. These datasets are selected for showing street scene (KITTI) and diverse proximity scene, all are very challenging for conventional SfM. The GT poses are provided by KITTI and BAL respectively.

We run reconstruction and collect timing information for all datasets. Our pipeline has robustness built into its formulation, whereas the incremental SfM uses a RANSAC-based outlier detection. For this they give similar camera trajectory, we therefore only compare the time consumption (Table IV) and give a visual illustration in Fig. 9 of our reconstruction results showing high density due to inclusion of low-parallax features. These results show that our global SfM pipeline is as robust as the incremental procedure with a much reduced computation time.

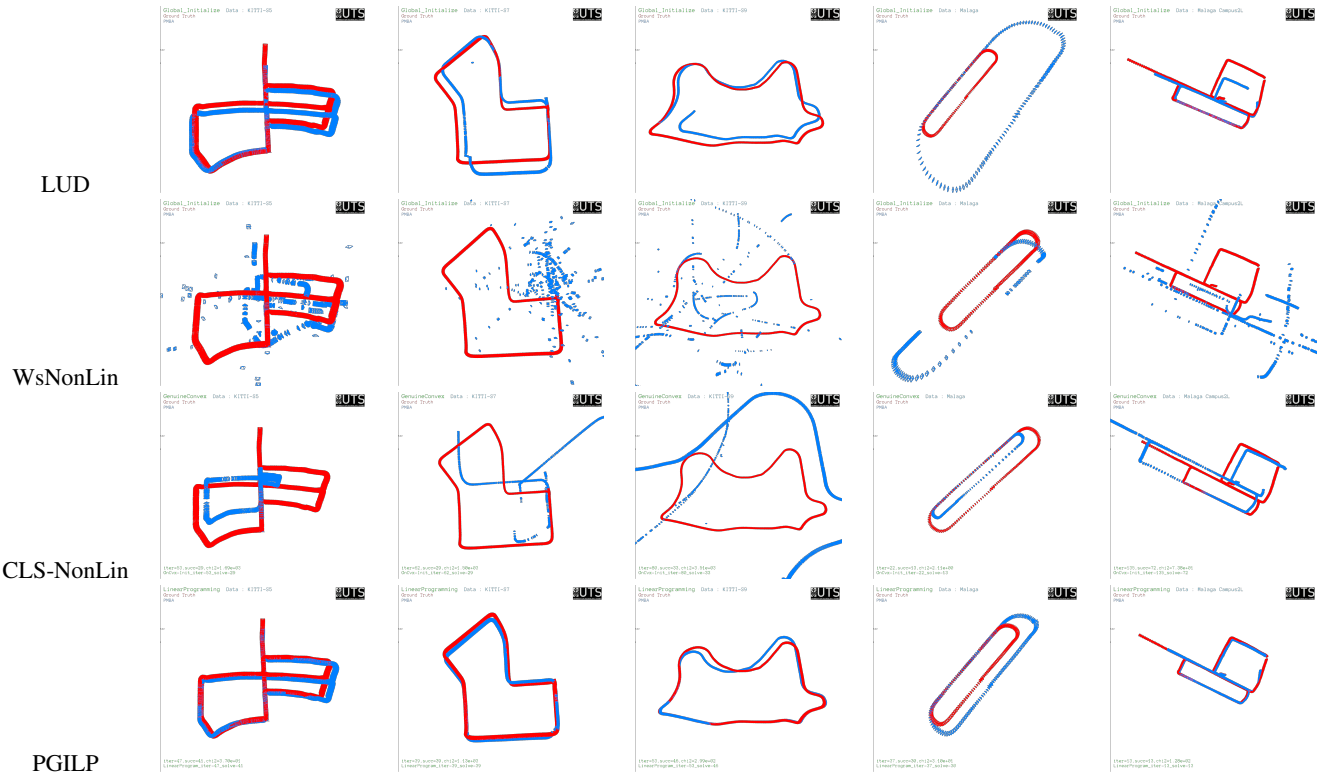


Fig. 7. Compare results of different Global Initialization methods (LUD, WsNonLin, CLS-NonLin and PGILP, all in blue), versus ground truth (red).

TABLE IV
COMPLETE PIPELINE COMPARISON: KITTI AND BAL

	Dataset	Order	# pose	Ours		Inctl [20]	
				# BA	Time[*min*]	# BA	Time[*min*]
K	-S0	SEQ	2349	1	6.01	1300	184.5
T	-S5	SEQ	1624	1	3.24	682	64.7
T	-S7	SEQ	1045	1	5.8	623	47.5
I	-S9	SEQ	1591	1	7.9	933	59.8
B	Venice-744	OOF	744	1	16.6	71	66.3
A	Trafalgar-257	OOF	257	1	1.11	25	2.56
L	Dubrovnik-356	OOF	356	1	5.03	60	14.0

"Ours" refers to our proposed pipeline, "Incrtl" referes to the incremental pipeline; "SEQ denotes sequential data", "OOF" refers to out-of-order.

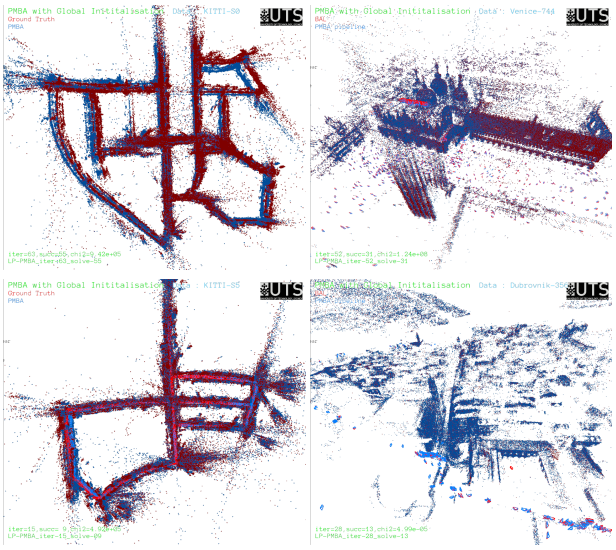


Fig. 9. Comparing our pipeline reconstruction results (blue) with reference data (red) on KITTI and RAL datasets (clockwise): KITTI-S0, Venice-744, Dubrovnik-356, KITTI-S5

V. CONCLUSION

We presented a novel global SfM pipeline that robustly form reconstruction for urban environment that is challenging to conventional methods. This pipeline exploits the recently explored bundle adjustment method using parallax angle between observation rays to accommondate low-parallax feature while achieving superior convergence performance. For realistic application, we provide in the SfM a robust initialization method that is compatible to the BA stage. We showed that camera positions can be estimated as a hybrid pose-graph problem that is friendly to collinear motion, robust to mismatched image-pairs and simple in formulation. Experimental results show that the proposed pipeline outperforms conventional method at every stage of SfM in terms of friendliness to collinear motion, tolerance to low-parallax features, accuracy, robustness and convergence speed. The consistent test results on benchmark datasets demonstrates that the proposed pipeline is a complete working solution in addressing urban SLAM issues, for sequential or out-of-order scenario. As future work, we plan to transform the offline SfM pipeline into a real-time visual SLAM system with a guaranteed zero-frame loss initialization.

REFERENCES

[1] G. Bresson, Z. Alsayed, and L. Yu, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, pp. 1309–1332, 2017.

[2] J. Civera, A. J. Davison, and J. M. M. Montiel, "Inverse depth parametrization for monocular slam," *IEEE Transactions on Robotics*, vol. 24, pp. 932–945, 2008.

[3] L. Zhao, S. Huang, Y. Sun, L. Yan, and G. Dissanayake, "Parallaxba: bundle adjustment using parallax angle feature parametrization," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 493–516, 2015.

[4] R. Mur-Artal and J. Tardos, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and rgb-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, Oct 2017.

[5] H. Zhang, K. Hasith, and H. Wang, "A hybrid feature parametrization for improving stereo-slam consistency," in *13th IEEE International Conference on Control Automation (ICCA)*, 2017, pp. 1021–1026.

[6] L. Liu, T. Zhang, Y. Liu, B. Leighton, L. Zhao, S. Huang, and G. Dissanayake, "Parallax bundle adjustment on manifold with improved global initialization," in *accepted by 2018 The 13th International Workshop on the Algorithmic Foundations of Robotics (WAFR), https://github.com/liyang-liu/PMBA_WAFR*.

[7] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, pp. 1309–1332, 2016.

[8] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," *ACM Trans. Graph.*, vol. 25, pp. 835–846, 2006.

[9] V. M. Govindu, "Combining two-view constraints for motion estimation," in *CVPR*, 2001.

[10] K. Sim and R. I. Hartley, "Recovering camera motion using $l\infty$ minimization," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*.

[11] K. Wilson and N. Snavely, "Robust global translations with 1dsfm," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[12] Z. Cui, N. Jiang, and P. Tan, "Linear global translation estimation from feature tracks," in *Proceedings of the British Machine Vision Conference (BMVC) 2015*, pp. 46.1–46.13.

[13] R. Tron and R. Vidal, "Distributed 3-d localization of camera sensor networks from 2-d image measurements," *IEEE Transactions on Automatic Control*, vol. 59, pp. 3325–3340, 2014.

[14] P. Moulon, P. Monasse, and R. Marlet, "Global fusion of relative motions for robust, accurate and scalable structure from motion," *2013 IEEE International Conference on Computer Vision*.

[15] O. Ozyesil and A. Singer, "Robust camera location estimation by convex programming," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] A. Chatterjee and V. M. Govindu, "Efficient and robust large-scale rotation averaging," *2013 IEEE International Conference on Computer Vision*.

[17] M. Brand, M. E. Antone, and S. J. Teller, "Spectral solution of large-scale extrinsic camera calibration as a graph embedding problem," in *Computer Vision - ECCV 2004*.

[18] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *I. J. Robotics Res.*, vol. 32, pp. 1231–1237, 2013.

[19] J.-L. Blanco, F.-A. Moreno, and J. Gonzalez, "A collection of outdoor robotic datasets with centimeter-accuracy ground truth," *Autonomous Robots*, vol. 27, no. 4, p. 327, Aug 2009.

[20] P. Moulon, P. Monasse, R. Marlet, and Others, "Openmvg. an open multiple view geometry library." https://github.com/openMVG/openMVG.

[21] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, "Bundle adjustment in the large," in *11th European Conference on Computer Vision (ECCV)*, 2010.