Visual-Odometric Localization and Mapping for Ground Vehicles Using SE(2)-XYZ Constraints

Fan Zheng and Yun-Hui Liu

***Abstract*— This paper focuses on the localization and mapping problem on ground vehicles using odometric and monocular visual sensors. To improve the accuracy of vision based estimation on ground vehicles, researchers have exploited the constraint of approximately planar motion, and usually implemented it as a stochastic constraint on an SE(3) pose. In this paper, we propose a simpler algorithm that directly parameterizes the ground vehicle poses on SE(2). The out-of-SE(2) motion perturbations are not neglected, but incorporated into an integrated noise term of a novel SE(2)-XYZ constraint, which associates an SE(2) pose and a 3D landmark via the image feature measurement. For odometric measurement processing, we also propose an efficient preintegration algorithm on SE(2). Utilizing these constraints, a complete visual-odometric localization and mapping system is developed, in a commonly used graph optimization structure. Its superior performance in accuracy and robustness is validated by real-world experiments in industrial indoor environments.**

I. Introduction

Localization and mapping of mobile vehicles in unknown environments is a fundamental task for autonomous robot navigation. Over the last twenty years, extensive efforts have been made on exploiting vision for localizing mobile robots. However, high-precision and robust visual localization and mapping remains challenging. Also, visual state estimation for robot platforms with special geometric structures could be further investigated.

Mainstream approaches of visual Simultaneously Localization and Mapping (SLAM) [1] can be classified into *filtering* and *optimization* approaches. Filtering methods like *MonoSLAM* [2] iteratively predict the robot state by motion model and update it using sensor measurements. Optimization approaches minimize a cost function formulated from all motion and measurement constraints, such as PTAM [3] and ORB-SLAM [4] [5], which rely on image *features* and minimize the reprojection errors. *Direct* methods were also proposed to manipulate on the pixel intensity and minimize the photometric errors [6] [7].

Monocular vision is often fused with inertial measurement units (IMU) or odometric sensors to achieve robust and true-scale estimation. MSCKF is a classic filtering based visual inertial navigation system (VINS), which maintains the last several frame poses in a sliding window to increase accuracy. For optimization based VINS, Forster *et al.* [8] introduced a preintegration theory to iteratively generate the inertial constraints between two keyframes. State-of-the-art VINS methods include [9] [10] [11], among which [9] uses nonlinear optimization to performing visual inertial estimation, and [10] [11] added the functionality of loop closure and map reuse. Odometric sensors like wheel encoders are usually available in wheeled robots and can be similarly fused with vision in optimization frameworks [12] [13]. Some works fuse both inertial sensors and wheel encoders with vision [14] [15].

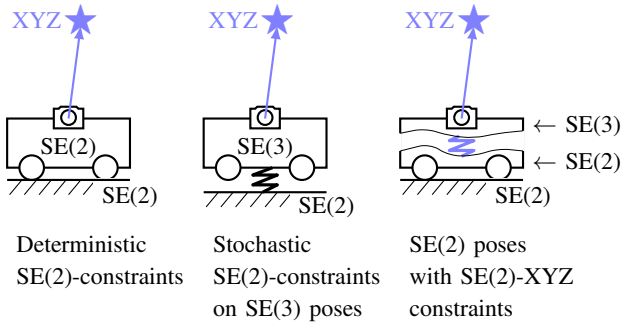

Fig. 1. Illustration of SE(2)-XYZ constraints compared to previous models. Deterministic SE(2)-constraints are not accurate on rough terrains or with motion shaking of the ground vehicles; stochastic SE(2)-constraints on SE(3) poses are accurate enough, but redundant; SE(2)-XYZ constraints help to reserve both the SE(2) poses and the out-of-SE(2) perturbation information.

The approaches mentioned above are for navigation in general 3D space, while this paper mainly focuses on the localization and mapping on ground vehicles. For navigation on ground, usually the SE(2) pose is of the main interest, and the constraint that the vehicle moves on a planar (or nearly planar) surface can be leveraged to aid the state estimation. Lategahn *et al.* [16] presented a filtering based solution for ground vehicles, with the SE(2) poses represented as 3-DoF vectors. Scaramuzza [17] further considered the *Instantaneous Center of Rotation* (ICR) constraint of wheeled robots, showing that visual estimation can be performed with one single point. The main problem with the deterministic SE(2)-constraints is that vehicle motion in real-world environments is often out of the constrained model due to rough terrains or motion shaking. This may degrade the performance of visual estimation, since all the 6 DoF of SE(3) are highly coupled with the visual observations of the landmarks. Experiments in [18] showed that while the constrained model may be good

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.

F. Zheng 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. F. Zheng is also with VisionNav Robotics Limited, China. {fzheng@link.cuhk.edu.hk, yhliu@mae.cuhk.edu.hk}

in outlier removal, a general model provides better estimation results.

To leverage the planar motion constraint more effectively, stochastic SE(2)-constraints have been proposed in some recent works [15] [13] [19]. The stochastic constraints are usually implemented as unary constraints on the SE(3) vehicle poses, allowing small perturbations out of SE(2). Unlike these methods, we present in this paper that ground vehicle poses can be directly parameterized on SE(2) without neglecting the out-of-SE(2) motion perturbations. This is inspired by the consideration that the out-of-SE(2) perturbations usually considerably affect only the visual measurements, hence could be incorporated into the visual constraints. We achieved this by proposing a novel *SE(2)-XYZ* constraint.

Our contributions can be summarized as below. First, for on-SE(2) pose estimation, a novel *SE(2)-XYZ* constraint is proposed, which incorporates both the image feature measurements and the out-of-SE(2) motion perturbations. It is simpler and more robust than the stochastic constraints, and conforms with the real world better than the previous deterministic SE(2) constraints. Second, a preintegration algorithm is proposed for odometric measurements, which is directly on SE(2), unlike the previous on-SE(3) methods [13]. Finally, a complete localization and mapping system is presented based on these constraints, demonstrating superior performance in accuracy and robustness. The implementation is open-source at *https://github.com/izhengfan/se2lam*.

In what follows, Section II, III and IV introduce the proposed system in the preliminary theory, the novel proposed constraints, and the system implementation. Experimental analysis is presented in Section V. Section VI concludes this work.

II. Preliminaries

A. Frames and Notation

The multiple frames are defined as in Fig. 2.

We use ${}^W\mathbf{R}_B, {}^W\mathbf{p}_B$ to denote the 3D rotation matrix and position of frame B w.r.t. frame W. A vector ${}^W\boldsymbol{\nu}_B = {}^W[\mathbf{r}' \ \phi]'_B$ denotes the corresponding SE(2) pose:

$$
\begin{aligned}
{}^W\mathbf{r}_B := {}^W\begin{bmatrix} r_x \\ r_y \end{bmatrix}_B &= {}^W\begin{bmatrix} p_x \\ p_y \end{bmatrix}_B \\
{}^W\phi_B &= \mathrm{Log}({}^W\mathbf{R}_B)_z
\end{aligned} \tag{1}
$$

where we use $()_x, ()_y, ()_z$ to represent the x, y, z component of a 3D vector. $\mathrm{Log}(\mathbf{R}) := \boldsymbol{\theta}$ is the rotation vector corresponding to rotation matrix; inversely there is $\mathbf{R} = \mathrm{Exp}(\boldsymbol{\theta})$ [8] [20]. Since the states of frame B w.r.t. frame W is of the main concern here, they are simply written as $\mathbf{R}, \mathbf{p}, \boldsymbol{\nu}$ in what follows.

The position of a landmark is denoted by $\mathbf{l}$. It is written as ${}_C\mathbf{l}$ when transformed to frame C: ${}_C\mathbf{l} = {}^C\mathbf{R}_W\,\mathbf{l} + {}^C\mathbf{p}_W$.

B. Nonlinear Graph Optimization

Graph optimization [21] formulates a state estimation problem as a graph $\mathcal{G}$, and solves it by finding the state values that minimize a cost function like:

$$
F(\mathcal{X}) = \sum_{k\in\mathcal{G}} \mathbf{e}_k(\mathcal{X})^T \boldsymbol{\Omega}_k \, \mathbf{e}_k(\mathcal{X}) \tag{2}
$$

$$
\mathcal{X}^* = \arg\min_{\mathcal{X}} F(\mathcal{X}) \tag{3}
$$

in which a state parameter block $\mathcal{X}_i$ in $\mathcal{X}$ is represented by a *node/vertex*, and a constraint on the state is an *edge*, whose error function is $\mathbf{e}_k(\mathcal{X})$. The information matrix $\boldsymbol{\Omega}_k$ serves as an weighting factor for $\mathbf{e}_k$, usually determined by the inverse of the covariance matrix of the constraint.

By linearizing the generalized incremental model (denoted by $\boxplus$) of the state $\mathcal{X}$, the cost function can be approximated as

$$
\begin{aligned}
F(\mathcal{X} \boxplus \boldsymbol{\delta\mathcal{X}}) &\approx \sum_{k\in\mathcal{G}} \|\mathbf{e}_k + \mathbf{J}_k \boldsymbol{\delta\mathcal{X}}\|_{\boldsymbol{\Omega}_k} \\
&= \sum_{k\in\mathcal{G}} \underbrace{\|\mathbf{e}_k\|_{\boldsymbol{\Omega}_k}}_{c_k} + 2\underbrace{\mathbf{e}_k^T\boldsymbol{\Omega}_k\mathbf{J}_k}_{\mathbf{b}_k^T}\boldsymbol{\delta\mathcal{X}} + \|\boldsymbol{\delta\mathcal{X}}\|_{\underbrace{\mathbf{J}_k^T\boldsymbol{\Omega}_k\mathbf{J}_k}_{\mathbf{H}_k}}
\end{aligned} \tag{4}
$$

where the constraint Jacobian of $\mathbf{e}_k$ is

$$
\mathbf{J}_k = \frac{\partial \mathbf{e}_k(\mathcal{X} \boxplus \boldsymbol{\delta\mathcal{X}})}{\partial \boldsymbol{\delta\mathcal{X}}}. \tag{5}
$$

Taking the information matrices, Jacobians and errors of all the constraints together as $\mathbf{b} = -\sum_{k\in\mathcal{G}} \mathbf{b}_k$, $\mathbf{H} = \sum_{k\in\mathcal{G}} \mathbf{H}_k$, a Gauss-Newton system can be constructed to solve for the incremental state $\boldsymbol{\delta\mathcal{X}}$:

$$
\mathbf{H}\,\boldsymbol{\delta\mathcal{X}} = \boldsymbol{b}. \tag{6}
$$

The state can then be iteratively updated like $\mathcal{X}^* \leftarrow \mathcal{X} \boxplus \boldsymbol{\delta\mathcal{X}}$ until convergence.

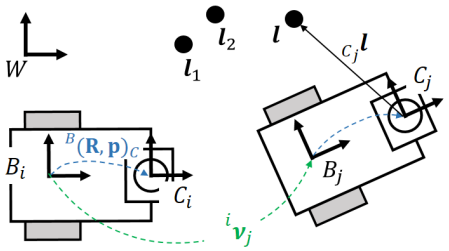

Fig. 2. Frame definition: the world frame is W; the vehicle base frame (odometry frame) is B; the camera frame is C.

III. Graph Optimization with SE(2)-XYZ Constraints

The proposed graph system (or *map*) includes nodes of *keyframes* and *landmarks*. A keyframe is a selected one from a consecutive sequence, with other frames discarded to ensure real-time efficiency.

Keyframe poses are directly parameterized on SE(2). In real-world environments, the vehicle may suffer from rough terrains, load changes, and shaking in motion. These perturbations out of SE(2) are formulated into the visual measurement constraints, termed *SE(2)-XYZ* constraints here, as illustrated in Fig. 1. Odometric measurements are processed using a preintegration technique on SE(2). Fig. 3 shows the proposed constraints in a graph.

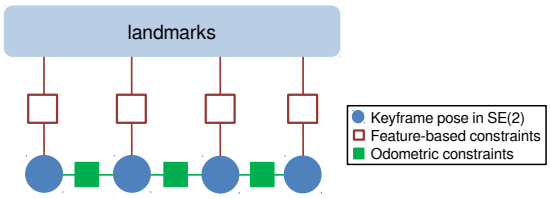

Fig. 3. The proposed constraints implemented for optimization.

A. Feature-Based SE(2)-XYZ Constraints

Feature-based reprojection errors are rather standard terms in common visual SLAM methods. Unlike the conventionally used SE(3)-XYZ constraint, the keyframes here are parameterized with SE(2) poses, and a novel SE(2)-XYZ constraint is introduced. To avoid the possible accuracy reduction caused by the deterministic planar motion constraint, the out-of-SE(2) motion uncertainty are also encapsulated in the SE(2)-XYZ constraint, along with the image feature measurement noises to formulate the constraint covariances.

Consider landmark ℓ observed by keyframe i, given the beforehand calibration $({}^{C}\mathbf{R}_B, {}^{C}\mathbf{p}_B)$ [22], the measurement model for the associated image feature coordinates is

$$\mathbf{u}(\boldsymbol{\nu}_i, \mathbf{l}_\ell) = \pi\big(\underbrace{{}^{C}\mathbf{R}_B\mathbf{R}_i^T(\mathbf{l}_\ell - \mathbf{p}_i) + {}^{C}\mathbf{p}_B}_{{}_{C_i}\mathbf{l}_\ell}\big) + \boldsymbol{\eta}_u \tag{7}$$

in which $\boldsymbol{\eta}_u \sim \mathcal{N}(\mathbf{0}, \sigma_u^2\mathbf{I}_2)$ is the feature measurement noise. $(\mathbf{R}_i, \mathbf{p}_i)$ can be retrieved from the SE(2) variables:

$$\mathbf{R}_i = \mathrm{Exp}([0\ 0\ \phi_i]^T), \quad \mathbf{p}_i = [\mathbf{r}_i^T\ \ 0]^T. \tag{8}$$

$\pi(\cdot)$ denotes the camera projection function:

$$\pi(\mathbf{l}) = \frac{1}{l_z}\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \end{bmatrix}\mathbf{l} \tag{9}$$

where (f_x, f_y) is the camera focal length, and (c_x, c_y) the principal point. It is easy to know that the projection Jacobian matrix is

$$\mathbf{J}^\pi(\mathbf{l}) = \frac{\partial\pi(\mathbf{l})}{\partial\mathbf{l}} = \frac{1}{l_z}\begin{bmatrix} f_x & 0 & -\frac{f_x l_x}{l_z} \\ 0 & f_y & -\frac{f_y l_y}{l_z} \end{bmatrix}. \tag{10}$$

We then incorporate the out-of-SE(2) perturbations as 'noises' into (7). Assume the vertical translational perturbation is $\eta_z \sim \mathcal{N}(0, \sigma_z^2)$, the rotational perturbations are $\boldsymbol{\eta}_{\theta_{xy}} \sim \mathcal{N}(\mathbf{0}_{2\times1}, \boldsymbol{\Sigma}_{\theta_{xy}})$, and the perturbed pose is like

$$\mathbf{R}_i \leftarrow \mathrm{Exp}(\underbrace{[\boldsymbol{\eta}_{\theta_{xy}}^T\ 0]^T}_{\boldsymbol{\eta}_\theta})\,\mathbf{R}_i, \quad \mathbf{p}_i \leftarrow \mathbf{p}_i + \underbrace{[0\ 0\ \eta_z]^T}_{\boldsymbol{\eta}_z}. \tag{11}$$

The measurement model (7) then becomes

$$\begin{aligned}\mathbf{u}(\boldsymbol{\nu}_i, \mathbf{l}_\ell) &= \pi\left({}^{C}\mathbf{R}_B\mathbf{R}_i^T\mathrm{Exp}(-\boldsymbol{\eta}_\theta)(\mathbf{l}_\ell - \mathbf{p}_i - \boldsymbol{\eta}_z) + {}^{C}\mathbf{p}_B\right) \\ &\quad + \boldsymbol{\eta}_u \\ &\approx \pi({}_{C_i}\mathbf{l}_\ell) + \mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_\theta}\boldsymbol{\eta}_\theta + \mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_z}\boldsymbol{\eta}_z + \boldsymbol{\eta}_u\end{aligned} \tag{12}$$

where we use first-order approximation to linearize the noise terms. The Jacobians of $\mathbf{u}$ w.r.t. $(\boldsymbol{\eta}_\theta, \boldsymbol{\eta}_z)$ are computed like

$$\begin{aligned}\mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_\theta} &= \mathbf{J}^\pi({}_{C_i}\mathbf{l}_\ell){}^{C}\mathbf{R}_B\mathbf{R}_i^T(\mathbf{l}_\ell - \mathbf{p}_i)^\wedge \\ \mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_z} &= -\mathbf{J}^\pi({}_{C_i}\mathbf{l}_\ell){}^{C}\mathbf{R}_B\mathbf{R}_i^T.\end{aligned} \tag{13}$$

in which $()^\wedge$ denotes a skew-symmetric mapping.

Formulating a synthetic zero-mean noise $\delta\boldsymbol{\eta}_u = \mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_\theta}\boldsymbol{\eta}_\theta + \mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_z}\boldsymbol{\eta}_z + \boldsymbol{\eta}_u$, we can approximate its covariance like:

$$\boldsymbol{\Sigma}_{\delta\boldsymbol{\eta}_u} = \mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_\theta}\Lambda_{12}\boldsymbol{\Sigma}_{\theta_{xy}}\Lambda_{12}^T\mathbf{J}^{\mathbf{u}T}_{\boldsymbol{\eta}_\theta} + \sigma_z^2\mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_z}\mathbf{e}_3\mathbf{e}_3^T\mathbf{J}^{\mathbf{u}T}_{\boldsymbol{\eta}_z} + \sigma_u^2\mathbf{I}_2 \tag{14}$$

due to that $\boldsymbol{\eta}_\theta, \boldsymbol{\eta}_z, \boldsymbol{\eta}_u$ are independent with each other. Here $\mathbf{e}_i$ denotes the ith column in $\mathbf{I}_3$, and $\Lambda_{12} = [\mathbf{e}_1\ \mathbf{e}_2]$.

We can now construct the reprojection residual error term of the SE(2)-XYZ constraint, given the coordinates of the detected feature ${}^{i\ell}\mathbf{u}$:

$${}^{i\ell}\mathbf{e} = \pi({}_{C_i}\mathbf{l}_\ell) - {}^{i\ell}\mathbf{u}. \tag{15}$$

The information matrix for it is $\boldsymbol{\Sigma}^{-1}_{\delta\boldsymbol{\eta}_u}$. The Jacobians of ${}^{i\ell}\mathbf{e}_{\mathrm{img}}$ w.r.t. $\boldsymbol{\nu}_i, \mathbf{l}_\ell$ are

$$\begin{aligned}\mathbf{J}_i^{i\ell} &= \begin{bmatrix}\frac{\partial{}^{i\ell}\mathbf{e}}{\partial\mathbf{r}_i} & \frac{\partial{}^{i\ell}\mathbf{e}}{\partial\phi_i}\end{bmatrix}, \\ \frac{\partial{}^{i\ell}\mathbf{e}}{\partial\mathbf{r}_i} &= -\mathbf{J}^\pi({}_{C_i}\mathbf{l}_\ell){}^{C}\mathbf{R}_B\mathbf{R}_i^T\Lambda_{12} \\ \frac{\partial{}^{i\ell}\mathbf{e}}{\partial\phi_i} &= \mathbf{J}^\pi({}_{C_i}\mathbf{l}_\ell){}^{C}\mathbf{R}_B\mathbf{R}_i^T(\mathbf{l}_\ell - \mathbf{p}_i)^\wedge\mathbf{e}_3. \\ \mathbf{J}_\ell^{i\ell} &= \frac{\partial{}^{i\ell}\mathbf{e}}{\partial\mathbf{l}_\ell} = \mathbf{J}^\pi({}_{C_i}\mathbf{l}_\ell){}^{C}\mathbf{R}_B\mathbf{R}_i^T.\end{aligned} \tag{16}$$

Note that $(\mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_z}, \mathbf{J}^{\mathbf{u}}_{\boldsymbol{\eta}_\theta})$ are identical to $\mathbf{J}_i^{i\ell}$ if omitting $\Lambda_{12}, \mathbf{e}_3$ in $\mathbf{J}_i^{i\ell}$. Therefore, they can be computed along with $\mathbf{J}_i^{i\ell}$, and the extra computational cost caused by formulating $\boldsymbol{\Sigma}_{\delta\boldsymbol{\eta}_u}$ is limited.

B. Preintegrated Odometric Constraints

We here derive an odometric constraint using the preintegration technique analogous to the IMU preintegration in [8]. Similar odometric preintegration was also presented in [13]. While [13] performs preintegration on SE(3), we formulate it on SE(2), leading to a more concise algorithm.

To achieve a more generic formulation, without assigning any specific type of odometric sensors or any kinematic model, we assume that the odometry provides a discrete-time measurement between two consecutive frames $k, k+1$:

$$\tilde{\boldsymbol{\nu}}_k = {}^{k}\boldsymbol{\nu}_{k+1} + \boldsymbol{\eta}_{\nu k}, \ \boldsymbol{\eta}_{\nu k} = \begin{bmatrix}\boldsymbol{\eta}_{rk} \\ \eta_{\phi k}\end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\nu k}) \tag{17}$$

and the state propagation on SE(2) is like

$$\begin{aligned}\boldsymbol{\nu}_{k+1} &= \begin{bmatrix}\mathbf{r}_k + \Phi(\phi_k)(\tilde{\mathbf{r}}_k - \boldsymbol{\eta}_{rk}) \\ \phi_k + \tilde{\phi}_k - \eta_{\phi k}\end{bmatrix}, \\ \Phi(\phi) &= \exp(\phi^\times) = \exp\begin{bmatrix}0 & -\phi \\ \phi & 0\end{bmatrix} = \begin{bmatrix}\cos\phi & -\sin\phi \\ \sin\phi & \cos\phi\end{bmatrix}.\end{aligned} \tag{18}$$

Notice that the propagated $\mathbf{r}_{k+1}$ depends on the rotational state ϕ_k. To generate a measurement between two keyframes i and j without depending on $\boldsymbol{\nu}_i$, we can multiple $\Phi(-\phi_i)$

to the left of $\mathbf{r}_{k+1}$:

$$\begin{aligned}\Phi(-\phi_i)\mathbf{r}_{k+1} &= \Phi(-\phi_i)\mathbf{r}_k + \Phi(\underbrace{\phi_k - \phi_i}_{{}^i\phi_k})(\tilde{\mathbf{r}}_k - \boldsymbol{\eta}_{rk}) \\ &= \Phi(-\phi_i)\mathbf{r}_i + \underbrace{\sum_{n=i}^{k} \Phi({}^i\phi_n)(\tilde{\mathbf{r}}_n - \boldsymbol{\eta}_{rn})}_{{}^i_i\mathbf{r}_{k+1}}. \end{aligned} \tag{19}$$

Then we formulate the preintegrated measurements and the corresponding noises between keyframe i and j as:

$$\begin{aligned} {}^i\phi_j &= \sum_{k=i}^{j-1}(\tilde{\phi}_k - \eta_{\phi k}) = \sum_{k=i}^{j-1}\tilde{\phi}_k - \sum_{k=i}^{j-1}\eta_{\phi k} \\ &:= {}^i\tilde{\phi}_j - \delta^i\phi_j \\ {}^i_i\mathbf{r}_j &= \sum_{k=i}^{j-1}\Phi({}^i\tilde{\phi}_k)\Phi(-\delta^i\phi_k)(\tilde{\mathbf{r}}_k - \boldsymbol{\eta}_{rk}) \\ &\approx \sum_{k=i}^{j-1}\Phi({}^i\tilde{\phi}_k)(\mathbf{I}_2 - \delta^i\phi_k^{\times})(\tilde{\mathbf{r}}_k - \boldsymbol{\eta}_{rk}) \\ &\approx \sum_{k=i}^{j-1}\Phi({}^i\tilde{\phi}_k)\tilde{\mathbf{r}}_k - \sum_{k=i}^{j-1}\Phi({}^i\tilde{\phi}_k)(\boldsymbol{\eta}_{rk} + \delta^i\phi_k^{\times}\tilde{\mathbf{r}}_k) \\ &:= {}^i_i\tilde{\mathbf{r}}_j - \delta^i_i\mathbf{r}_j. \end{aligned} \tag{20}$$

For the propagation of the integrated noise terms $\delta^i\phi_j, \delta^i_i\mathbf{r}_j$, consider their iterative propagation forms:

$$\begin{aligned} \delta^i\phi_{k+1} &= \delta^i\phi_k + \eta_{\phi k} \\ \delta^i_i\mathbf{r}_{k+1} &= \delta^i_i\mathbf{r}_k + \Phi({}^i\tilde{\phi}_k)(\boldsymbol{\eta}_{rk} + \delta^i\phi_k^{\times}\tilde{\mathbf{r}}_k) \\ &= \delta^i_i\mathbf{r}_k + \Phi({}^i\tilde{\phi}_k)\boldsymbol{\eta}_{rk} + \Phi({}^i\tilde{\phi}_k)1^{\times}\tilde{\mathbf{r}}_k\delta^i\phi_k \end{aligned} \tag{21}$$

Writing (21) into a compact form, we have

$$\begin{aligned} &\begin{bmatrix}\delta^i_i\mathbf{r}_{k+1} \\ \delta^i\phi_{k+1}\end{bmatrix} := \delta^i\boldsymbol{\nu}_{k+1} = \mathbf{A}_k\delta^i\boldsymbol{\nu}_k + \mathbf{B}_k\boldsymbol{\eta}_{\nu k}, \\ &\mathbf{A}_k = \begin{bmatrix}\mathbf{I}_2 & \Phi({}^i\tilde{\phi}_k)1^{\times}\tilde{\mathbf{r}}_k \\ \mathbf{0} & 1\end{bmatrix}, \ \mathbf{B}_k = \begin{bmatrix}\Phi({}^i\tilde{\phi}_k) & \mathbf{0} \\ \mathbf{0} & 1\end{bmatrix}. \end{aligned} \tag{22}$$

Therefore, given the odometric measurement noise covariance $\boldsymbol{\Sigma}_{\nu k}$ in each step, the covariance of $\delta^i\boldsymbol{\nu}_k$ is propagated like

$$\boldsymbol{\Sigma}_{\delta^i\nu_{k+1}} = \mathbf{A}_k\boldsymbol{\Sigma}_{\delta^i\nu_k}\mathbf{A}_k^T + \mathbf{B}_k\boldsymbol{\Sigma}_{\nu k}\mathbf{B}_k^T \tag{23}$$

which, starting from an initial condition $\boldsymbol{\Sigma}_{\delta^i\nu_i} = \mathbf{0}_3$, can be computed incrementally along with the preintegrated measurement terms ${}^i\tilde{\phi}_j, {}^i_i\tilde{\mathbf{r}}_j$.

We can now construct an error term for optimization based on the preintegrated odometric measurements. Given two keyframes with SE(2) poses $\boldsymbol{\nu}_i$ and $\boldsymbol{\nu}_j$, the residual error is defined as

$${}^{ij}\mathbf{e} = \begin{bmatrix}\Phi(-\phi_i)(\mathbf{r}_j - \mathbf{r}_i) \\ \phi_j - \phi_i\end{bmatrix} - \begin{bmatrix}{}^i_i\tilde{\mathbf{r}}_j \\ {}^i\tilde{\phi}_j\end{bmatrix}. \tag{24}$$

According to the derivation (19)-(23), the information matrix for ${}^{ij}\mathbf{e}$ should be $\boldsymbol{\Sigma}^{-1}_{\delta^i\nu_j}$. The Jacobians of ${}^{ij}\mathbf{e}$ w.r.t. $\boldsymbol{\nu}_i, \boldsymbol{\nu}_j$ are analytically computed like

$$\begin{aligned} \mathbf{J}_i^{ij} &= \frac{\partial^{ij}\mathbf{e}}{\partial\boldsymbol{\nu}_i} = \begin{bmatrix}-\Phi(-\phi_i) & -\Phi(-\phi_i)1^{\times}(\mathbf{r}_j - \mathbf{r}_i) \\ \mathbf{0} & -1\end{bmatrix} \\ \mathbf{J}_j^{ij} &= \frac{\partial^{ij}\mathbf{e}}{\partial\boldsymbol{\nu}_j} = \begin{bmatrix}\Phi(-\phi_i) & \mathbf{0} \\ \mathbf{0} & 1\end{bmatrix}. \end{aligned} \tag{25}$$

(a) AGV for experiments.

(b) *Dataset Warehouse* environment.

Fig. 4. The experimental platform and environment, provided by *VisionNav Robotics Ltd.* (a) The platform, a forklift AGV. (b) The environment of *Dataset Warehouse*.

IV. Tracking, Mapping, and Loop Closing

Inspired by PTAM [3] and ORB-SLAM [5], the proposed system performs the estimation in three parallel threads: *tracking*, *local mapping* and *loop closing*.

The *tracking* thread performs the odometric preintegration described in III-B, and initially guesses the pose of a new frame using the odometric measurements. ORB features [23] are extracted from the frame, matched with those in the latest keyframe, and further matched with the landmarks in the *local map* for more correspondences. When there are large enough parallax or time difference between the current frame and the latest keyframe, the current frame is selected as a new keyframe and inserted to the map.

Local mapping manages a *local map* $\mathcal{L}$, a collection of the last M keyframes $\{\boldsymbol{\nu}_i \,|\, i \in \mathcal{L}\}$ and their observed landmarks $\{\mathbf{l}_\ell \,|\, \ell \in \mathcal{L}\}$. Once a new keyframe is inserted, $\mathcal{L}$ is optimized using the feature-based SE(2)-XYZ constraints and the odometric constraints. To retain constant-time local optimization, the oldest keyframe here is marginalized out of $\mathcal{L}$ and serves as prior for the optimization. Keyframes that are out of $\mathcal{L}$, but observe landmarks in $\mathcal{L}$, also contribute to the optimization as prior. This strategy is inspired by [10].

The *loop closing* thread searches among all keyframes for one that views the same scene with the current keyframe, based on visual *Bags of Words* [24]. If loop closed, the constraint between the two matched keyframes is generated based on their co-visible landmarks [19]. This constraint is on SE(3), and can be mapped to SE(2) according to (1), after which we can construct a close-loop constraint numerically identical to ${}^{ij}\mathbf{e}$ in III-B. The global pose graph are optimized using this close-loop constraint and the odometric constraints. Landmarks are then adjusted w.r.t. the optimized keyframes.

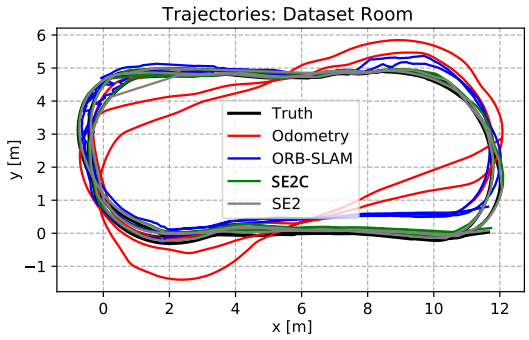

Fig. 5. *Dataset Room* evaluation: estimated keyframe trajectories by the odometry measurements, ORB-SLAM2, SE(2)-constrained algorithm (**SE2C**) [19], our method (**SE2**), as well as the ground truth.

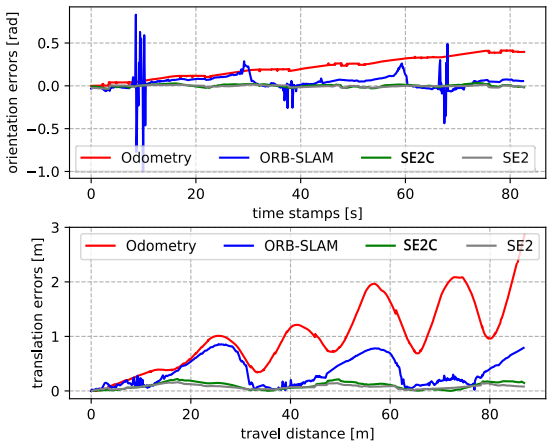

Fig. 6. *Dataset Room* evaluation: estimation errors computed w.r.t. the ground truth. The first graph visualizes the errors in the orientation angle. The second graph shows the translation errors (distances between estimated and true locations) with the traveling distance growing larger.

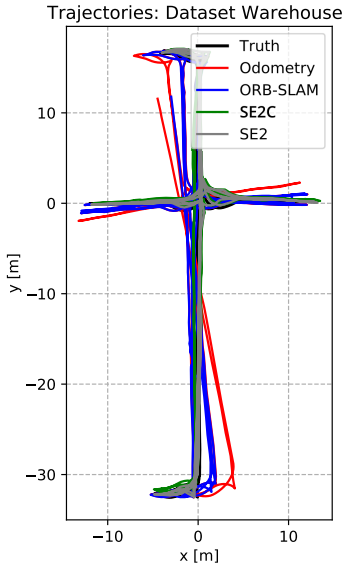

Fig. 7. *Dataset Warehouse* evaluation: estimated keyframe trajectories, by the odometry measurements, ORB-SLAM2, SE(2)-constrained algorithm (**SE2C**) [19], our method (**SE2**), as well as the ground truth.

V. Experimental Results

A. Evaluation Setup

We use the similar experimental setup to [19]. A forklift AGV as shown in Fig. 4-(a) is equipped with two encoders for odometric measurement. One encoder is a SICK DSK40 incremental encoder with a resolution of 1000 threads per round at the driving wheel, and the other a SICK ATM60 absolute encoder with a resolution of 2048 threads per round to measure the steering angle. The fiber optic gyro in [19] is not used here. A PointGrey Chameleon 2.0 camera with a fish-eye lens is mounted on the vehicle top, looking upwards and collecting 640×480 images at 30 Hz. Calibration and synchronization between odometric and visual sensors are performed offline [22]. The system runs in real time with an Intel i7-6500U CPU.

To collect experimental data, the AGV was manually driven in two industrial indoor scenarios, a small test room of around $8\times20\,\mathrm{m}^2\times3\,\mathrm{m}$, named *Dataset Room*, and a warehouse of around $40\times70\,\mathrm{m}^2\times6\,\mathrm{m}$, named *Dataset Warehouse* (see Fig. 4-(b)). Ground-truth trajectories are approximated by using a secondary downward-viewing camera mounted on the vehicle bottom to detect some markers with pre-measured location on the ground [19].

We compare our work (termed **SE2** here) with 1) the SE(2)-constrained system (termed **SE2C**) in [19] and 2) an odometric-aided version of the state-of-the-art visual estimation system ORB-SLAM2 [5] on the two datasets. Odometric measurements aid ORB-SLAM2 in its initialization and 'tracking with motion model' steps [1].

B. Results and Discussion

Fig. 5 demonstrates the trajectory of *Dataset Room* estimated by our system **SE2**, in comparison with the propagated odometry, the odometric-aided ORB-SLAM2, the SE(2)-constrained system **SE2C**, and the ground truth. The corresponding orientation and translation errors w.r.t. the ground truth are shown in Fig. 6. Similarly, Fig. 7 and Fig. 8 show the trajectories, orientation and translation errors on *Dataset Warehouse*.

[1]The source code of our adaptation on ORB-SLAM2 is available at *https://github.com/izhengfan/ORB_SLAM2*.

As can be seen from the graphs, odometric-aided visual SLAM (ORB-SLAM2) provides less accurate estimation results than both the SE(2)-constrained system and our on-SE(2) method. This indicates the importance of introducing the planar motion constraint into the state estimation problem when the vehicle moves on a planar surface. Such importance can also be inferred by observing the estimated z coordinates shown in Fig. 9. ORB-SLAM2 gives deviating and unstable z values on *Dataset Room*; similar results can also be seen on *Dataset Warehouse*. Since all the 6 DoFs of an SE(3) pose are highly coupled when involving visual measurements, inaccurate z values would naturally reduce the estimation accuracy of the other DoFs.

We further examine the improvement of our system compared to **SE2C**. Table I presents the numerical results of

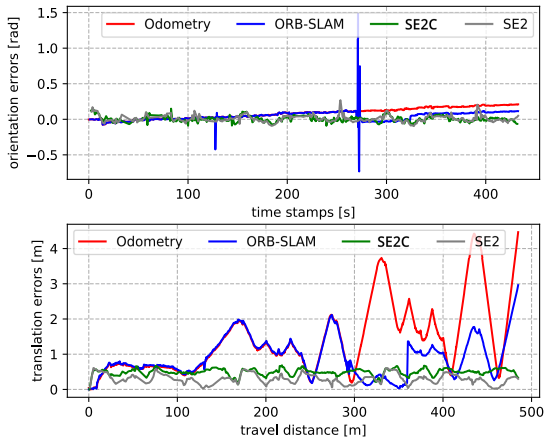

Fig. 8. *Dataset Warehouse* evaluation: estimation errors computed w.r.t. the ground truth. Graphs are arranged in the same order to Fig 6.

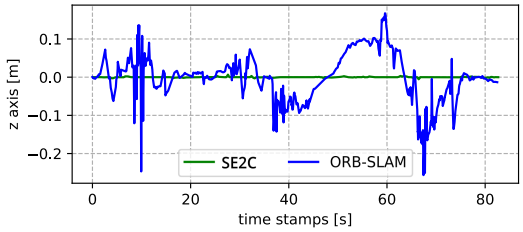

Fig. 9. Estimated z coordinates of the vehicle in *Dataset Room*. The true values should be near zero, just like the results by **SE2C**.

the estimation errors on the two datasets. We use the Root-Mean-Square of the errors (RMSE) here, including those of the errors in x, y, ϕ (yaw) coordinates, and the translation errors. In terms of RMSE, the results of our **SE2** yield accuracy of around 0.288% (0.085m/29.5m) in *Dataset Room* and 0.223% (0.328m/147m) in *Dataset Warehouse*, both of which are better than those by **SE2C** (0.381% and 0.339%).

A more important improvement of our system compared to **SE2C** is in robustness. In the experiments of **SE2C**, we found that the estimation is rather susceptible to the parameter setting of the stochastic SE(2)-constraints applied on the SE(3) poses. It sometimes requires quite some manual parameter tuning to make the estimation system work normally, which is undesirable in practices. As illustrated by the example in Fig. 10, even some parameter setting that looks normal may cause the estimated keyframe poses to deviate from the ground. This is probably due to the inherent instability of numerical optimization for complicated systems. Our system never suffers from such estimation failure, due to that the optimization graph is formulated directly on SE(2), leading to a much simpler linear system to solve.

VI. Conclusion

In this paper, we develop a localization and mapping framework for ground vehicles based on odometric and monocular visual sensors in a graph optimization formulation. Compared to common visual SLAM systems that parameterize the vehicle state with general SE(3) poses, or some

TABLE I

Estimation Errors Statistics (RMSE)

	Odom.	**ORB-SLAM**	**SE2C**	**SE2**
Dataset Room				
x err. (mm)	541.24	135.33	62.44	61.106
y err. (mm)	1028.83	371.01	93.15	59.021
ϕ err. (rad)	0.24835	0.12809	0.01567	0.01181
trans. err.(mm)	1162.51	394.93	112.15	84.956
accuracy*	4.469%	1.343%	0.381%	0.288%
Dataset Warehouse				
x err. (mm)	1615.19	1038.66	304.22	171.129
y err. (mm)	460.90	507.13	393.87	279.766
ϕ err. (rad)	0.10062	0.17149	0.03924	0.04921
trans. err.(mm)	1679.67	1155.85	497.68	327.955
accuracy	1.142%	0.787%	0.339%	0.223%

*Accuracy is calculated by the translation error over the travel distance in *one loop*.

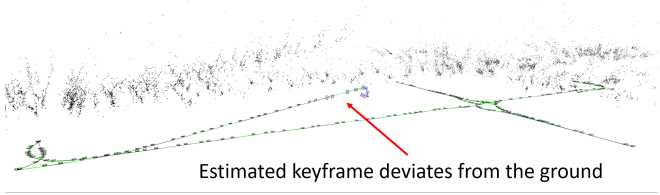

Fig. 10. Estimation using the SE(2)-constrained model [19] may 'collapse' under some perturbation parameter settings. Here the perturbation covariance for each out-of-SE(2) rotational DoF is set as 10^{-4}rad^2, in *Dataset Warehouse*. If setting it as 10^{-6}rad^2, no deviation happens and the estimation works well. The might be caused by the inherent instability of complicated numerical optimization systems.

recent works that apply the SE(2)-constraint on SE(3) poses stochastically, we directly formulate the vehicle poses on SE(2), without omitting the out-of-SE(2) perturbations in the real-world environments, leading to a simple and robust estimation algorithm. This is achieved by using a novel SE(2)-XYZ constraint, which not only involves information from the visual feature-based measurements, but also incorporates the out-of-SE(2) perturbations into its integrated noise term. We also develop an efficient on-SE(2) preintegration algorithm to generate odometric constraints between keyframes. We implement these constraints into a multi-thread software system, which runs in real-time to simultaneously estimate the vehicle pose and the landmark locations. Experiments in indoor industrial environments validate the superior performance of our system compared to state-of-the-art visual SLAM systems and the recently proposed SE(2)-constrained estimation system, in both accuracy and robustness.

In the future, more tests will be conducted on publicly available datasets to validate the performance of the proposed system. Furthermore, the way is paved for using the SE(2)-XYZ constraints in more estimation problems with different sensors. We consider building a VINS for ground vehicles by combining the SE(2)-XYZ constraints and the IMU preintegration techniques. Monocular or stereo visual SLAM for ground vehicles is also a good direction.

REFERENCES

[1] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] A. Davison *et al.*, "Monoslam: Real-time single camera slam," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, 2007.

[3] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *IEEE/ACM Int. Symp. Mix. Aug. Real.*, Nov 2007, pp. 225–234.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.

[5] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Trans. Robot.*, 2017.

[6] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[7] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[8] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, 2017.

[9] S. Leutenegger *et al.*, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.

[10] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular slam with map reuse," *IEEE Robot. Autom. Lett.*, vol. 2, no. 2, pp. 796–803, 2017.

[11] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 1004–1020, Aug 2018.

[12] A. Eudes *et al.*, "Fast odometry integration in local bundle adjustment-based visual slam," in *Int. Conf. Pattern Recog.*, Aug 2010, pp. 290–293.

[13] M. Quan, S. Piao, M. Tan, and S.-S. Huang, "Tightly-coupled Monocular Visual-odometric SLAM using Wheels and a MEMS Gyroscope," *ArXiv e-prints*, Apr. 2018.

[14] L. Li *et al.*, "Estimating position of mobile robots from omnidirectional vision using an adaptive algorithm," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1633–1646, 2015.

[15] K. J. Wu, C. X. Guo, G. Georgiou, and S. I. Roumeliotis, "Vins on wheels," in *Proc. IEEE Int. Conf. Robot. Autom*, May 2017, pp. 5155–5162.

[16] H. Lategahn, A. Geiger, and B. Kitt, "Visual slam for autonomous ground vehicles," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 1732–1737.

[17] D. Scaramuzza, "1-point-ransac structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints," *Int. J. Comput. Vis.*, vol. 95, no. 1, p. 74, 2011.

[18] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point ransac," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 4293–4299.

[19] F. Zheng, H. Tang, and Y. H. Liu, "Odometry-vision-based ground vehicle motion estimation with se(2)-constrained se(3) poses," *IEEE Trans. Cybern.*, pp. 1–12, 2018.

[20] T. Barfoot, *State Estimation for Robotics*. Cambridge Univ. Press, 2017.

[21] R. Kümmerle *et al*, "g2o: A general framework for graph optimization," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 3607–3613.

[22] H. Tang and Y. H. Liu, "A fully automatic calibration algorithm for a camera odometry system," *IEEE Sensors J.*, vol. 17, no. 13, pp. 4208–4216, July 2017.

[23] E. Rublee *et al.*, "Orb: An efficient alternative to sift or surf," in *Proc. Int. Conf. Comput. Vis.*, Nov 2011, pp. 2564–2571.

[24] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, October 2012.