Visual-Inertial SLAM

Bolin He
Department of Electrical Computer Engineering
University of California, San Diego
b2he@ucsd.edu

I. INTRODUCTION

Visual-inertial simultaneous localization and mapping (VI-SLAM) that fuses IMU and camera data for localization and environmental perception has significant wide applications in mobile robotics, self-driving car and so forth. In this project, I will implement VI-SLAM using the Extended Kalman Filter with a prediction step based on SE(3) kinematics and an update step based on the stereo camera observation model to perform localization and mapping. More precisely, firstly I find the IMU based localization via EKF Prediction, and then do landmark mapping via EKF Update, and finally combine both the IMU and camera data simultaneously to obtain a complete VI-SLAM result.

II. PROBLEM FORMULATION

The purpose in this project is to do simultaneous localization and mapping based on IMU and camera data. Basically, it includes three parts as:

II-A. IMU-based Localization via EKF Prediction

It is a localization-only problem. To simplify it, some assumptions are made as:

i. Use kinematic rather than dynamic equations for the prediction step
ii. The world-frame landmark coordinates $m \in R^{3*M}$ are known
iii. The data association $\pi_t : \{1,...,M\} \rightarrow \{1,...,N_t\}$ stipulating which landmarks were observed at each time t is known

The objective is to estimate the inverse IMU pose $U_t :=_W T_{1,t}^{-1} \in SE(3)$ over time based on the IMU measurements $\mu_{0:T}$ with $\mu_t := [v_t^T, \omega_t^T]^T$and the visual feature observations $z_{0:T}$.

II-B. Landmark Mapping via EKF Update

It is a mapping-only problem. To simplify it, some assumptions are made as:

i. The calibration matrix M, extrinsic ${}_OT_I \in SE(3)$, inverse IMU pose $U_t \in SE(3)$, projection matrix P, new observation $z_t \in R^{4\times N_t}$

ii. The landmarks are static

The objective is to estimate the homogeneous coordinates $m \in R^{4\times M}$ in the world frame of the landmarks that generated the visual observations based on the visual feature observations $z_{0:T}$.

III-C. Visual-Inertial SLAM

Assume I know calibration matrix M, extrinsic ${}_OT_l \in SE(3)$, landmark positions $m \in R^{3*M}$, new observation $z_{t+1} \in R^{4*N_t}$. By combining the IMU prediction step with the landmark update step and an IMU update step based on the stereo camera observation model jointly, I can obtain a complete VI SLAM result.

III. TECHNICAL APPROACH

III-A. Rigid Body Motion

Assume the rigid body position is S_B in body frame and S_w in world frame. There exists a rotation matrix $R \in SO(3)$, where

$$R = R_z(\psi)R_y(\theta)R_x(\phi) = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{bmatrix}$$

such that

$$S_W = R \cdot S_B$$

by including the translation matrix p, the rigid body transformation can be written as

$$\begin{bmatrix} S_w \\ 1 \end{bmatrix} = \begin{bmatrix} R & p \\ 0 & 1 \end{bmatrix} \begin{bmatrix} S_B \\ 1 \end{bmatrix}$$

III-B. Lie Group Algebra

For $\theta \in so(3)$:

$$\hat{\theta} = \widehat{\begin{bmatrix}\theta_1 \\ \theta_2 \\ \theta_3\end{bmatrix}} = \begin{bmatrix} 0 & -\theta_3 & \theta_2 \\ \theta_3 & 0 & -\theta_1 \\ -\theta_2 & \theta_1 & 0 \end{bmatrix}$$

For $u_t := \begin{bmatrix}\mathbf{v}_t \\ \omega_t\end{bmatrix} \in R^6$

$$\hat{u}_t := \begin{bmatrix}\hat{\omega}_t & \mathbf{v}_t \\ \mathbf{0}^T & 0\end{bmatrix} \in R^{4*4}$$

$$u_t^{\curlywedge} := \begin{bmatrix}\hat{\omega}_t & \hat{\mathbf{v}}_t \\ 0 & \hat{\omega}_t\end{bmatrix} \in R^{6*6}$$

III-C. IMU-based Localization via EKF Prediction

Based on our assumption, the prior is

$$U_t \,|\, z_{0:t}, u_{0:t1} \sim \mathcal{N}(\mu_{t|t}, \Sigma_{t|t})$$

with $\mu_{t|t} \in SE(3)$, $\Sigma_{t|t} \in R^{6*6}$

With time discretization τ and noise $w_t \sim \mathcal{N}(0, W)$, the motion model can be define as

$$U_{t+1} = exp(-\tau((u_t + w_t))^{\hat{}})U_t$$

with $u_t := \begin{bmatrix}\mathbf{v}_t \\ \omega_t\end{bmatrix} \in R^6$

Now consider the pose kinematics with perturbation, which can separate the effect of the noise w_t from the motion of deterministic part of T_t as

$$\delta\mu_{t+1} = exp(-\tau u_t{}^{\curlywedge})\delta\mu_t + w_t$$

I can derive the motion model in terms of nominal kinematics of the mean of Tt and zero-mean perturbation kinematics as

$$\mu_{t+1|t} = exp(-\tau\hat{u}_t)\mu_{t|t}$$

$$\Sigma_{t+1|t} = E[\delta\mu_{t+1|t}\delta\mu_{t+1|t}^T] = exp(-\tau u_t{}^{\curlywedge})\Sigma_{t|t}exp(-\tau u_t{}^{\curlywedge})^T + W$$

In this part, I first load the data into IDE, and then downsample the features with step size 10. In this way, the new features are only one tenth of the original features. I obtain the IMU measurements by stacking the linear velocity and angular velocity together. I set the Gaussian noise covariance with value 10. Utilize the motion model above, I can derive the EKF prediction result.

III-D. Landmark Mapping via EKF Update

Based on our assumption, the prior are $\mu_t \in R^{3*M}$ and $\Sigma_t \in R^{3M*3M}$.

So I can predict observations $\tilde{z}_{t,i}$ based on μ_t and known correspondences π_t, where

$$\pi(q) := \frac{q}{q3}\mathbf{q} \in R^4$$

$$\frac{d\pi}{d\mathbf{q}}(\mathbf{q}) = \frac{1}{q3}\begin{bmatrix} 1 & 0 & -\frac{q1}{q3} & 0 \\ 0 & 1 & -\frac{q2}{q3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{q4}{q3} & 1 \end{bmatrix} \in R^{4*4}$$

$$\tilde{z}_{t,i} := M\pi({}_oT_lU_t\underline{\mu}_{t,j}) \in R^4$$

And the the Jacobin of of $\tilde{z}_{t,i}$ with respect to m_j evaluated at $\mu_{t,j}$ becomes

$$H_{t,i,j} = \begin{cases} M\frac{d\pi}{d\mathbf{q}}({}_oT_IU_t\underline{\mu}_{i,j})_oT_IU_tP^T & \text{if observation i corresponds to landmark j at time t} \\ 0 \in R^{4*3} & \text{otherwise} \end{cases}$$

Then EKF update can be

$$K_t = \Sigma_tH_t^T(H_t\Sigma_tH_t^T + I \otimes V)^{-1}$$
$$\mu_{t+1} = \mu_t + K_t(z_t - \tilde{z}_t)$$
$$\Sigma_{t+1} = (I - K_tH_t)\Sigma_t$$

In details, firstly I extract all the available features at every single time stamp. And then I convert their coordinates from optical frame into world frame for landmarks. If the landmark is first seen, initialize it, otherwise update the its position. After iteration, I can obtain the landmark mapping result.

III-E. Visual-Inertial SLAM

Now the known prior is

$$U_{t+1} \,|\, z_{0:t}, u_{0:t} \sim \mathcal{N}(\mu_{t+1|t}, \Sigma_{t+1|t})$$

with $\mu_{t+1|t} \in SE(3)$ and $\Sigma_{t+1|t} \in R^{6*6}$

And our observation model $z_{t+1,i}$ with measurement noise $v_t \sim \mathcal{N}(0,V)$ is

$$z_{t+1,i} = h(U_{t+1}, m_j) + v_{t+1,i} := M\pi({}_oT_I U_{t+1}\underline{m}_j) + v_{t+1,i}$$

Now the observation model is the same as in the visual mapping problem. However, instead of the landmark positions $m \in R^{3\times M}$, now the variable of interest is the inverse IMU pose $U_{t+1} \in SE(3)$. So the Jacobin matrix becomes $H_{t+1|t} \in R^{4N_t*6}$ with respect to the inverse IMU pose evaluated at $\mu_{t+1|t}$. Utilize the first-order Taylor series approximation of observation i at time t + 1 using an inverse IMU pose perturbation $\delta\mu_{t+1|t+1}$, I can obtain $z_{t_+1,i}$.

So now the predicted observation $\tilde{z}_{t+1,i}$ and the Jacobin of it are

$$\begin{aligned} &\tilde{z}_{t+1,i} := M\pi({}_oT_l\mu_{t+1|t}m_j) \\ &H_{i,t+1|t} = M\frac{d\pi}{d\mathbf{q}}({}_oT_I\mu_{t+1|t}m_j){}_oT_I(\mu_{t+1|t}m_j)^{\odot} \in R^{4*6} \end{aligned}$$

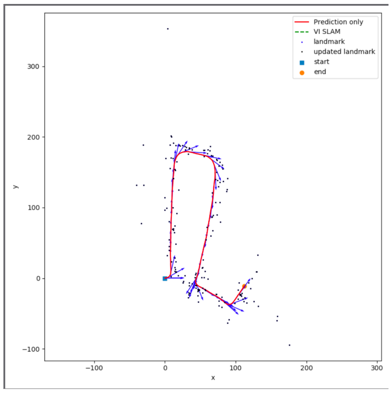
Now the EKF update becomes

$$\begin{aligned} &K_{t+1|t} = \Sigma_{t+1|t}H_{t+1|t}^T(H_{t+1|t}\Sigma_{t+1|t}H_{t+1|t}^T + I \otimes V)^{-1} \\ &\mu_{t+1|t+1} = exp((K_{t+1|t}(z_{t+1} - \tilde{z}_{t+1}))^{\wedge})\mu_{t+1|t} \\ &\Sigma_{t+1|t+1} = (I - K_{t+1|t}H_{t+1|t})\Sigma_{t+1|t} \end{aligned}$$

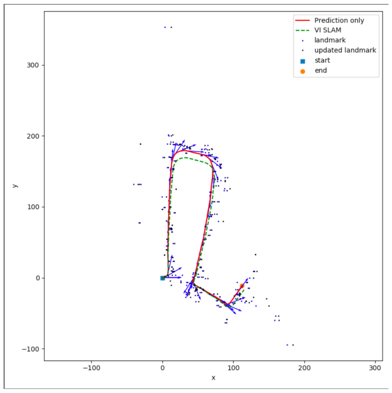
In this part, with the basic idea from the prior parts, I combine the IMU prediction step with the landmark update step and an IMU update step based on the stereo camera observation model jointly. Finally I can obtain a complete VI SLAM result.

IV. RESULT

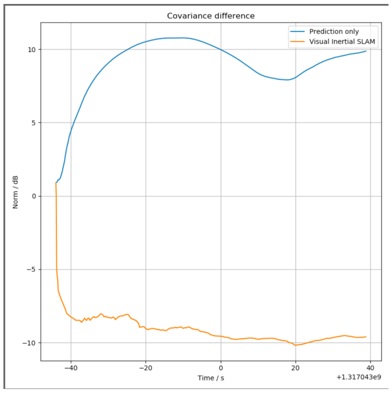
IV-A. Trainset 0022



figure(a.1)



figure(a.2)

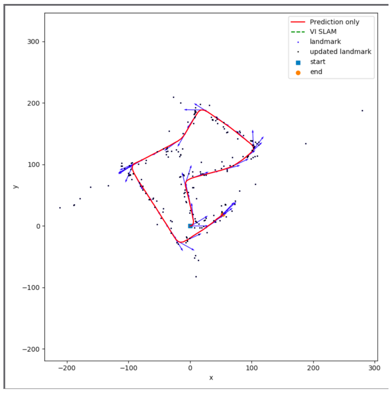


figure(a.3)

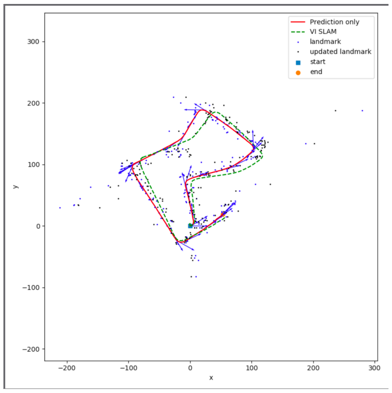
figure(a.1). EKF Prediction only and Landmark Mapping
figure(a.2). VI SLAM compared with figure(a.1)
figure(a.3). Covariance difference between Prediction only and VI SLAM

IV-B. Trainset 0027

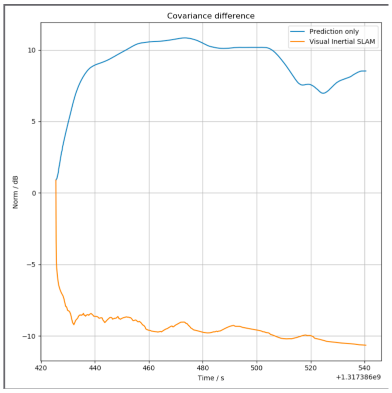
figure(b). Map and trajectory of transit 1



figure(b.1)



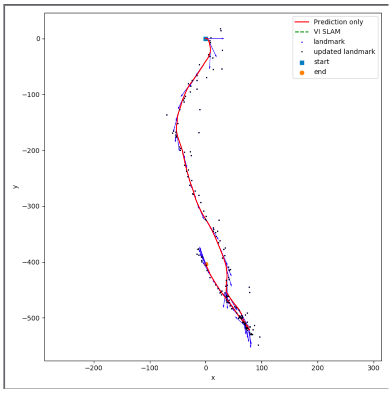
figure(b.2)



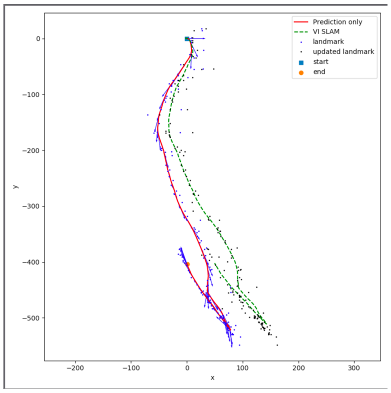
figure(b.3)

figure(b.1). EKF Prediction only and Landmark Mapping
figure(b.2). VI SLAM compared with figure(b.1)
figure(b.3). Covariance difference between Prediction only and VI SLAM

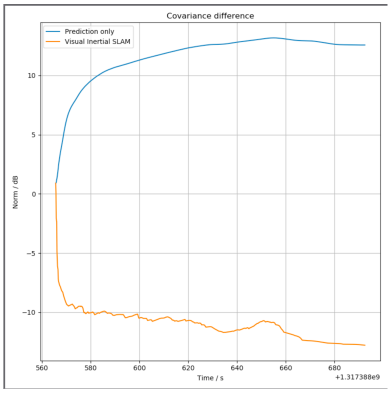
IV-C. Trainset 0034



figure(c.1)



figure(c.2)



figure(c.3)

figure(c.1). EKF Prediction only and Landmark Mapping
figure(c.2). VI SLAM compared with figure(c.1)
figure(c.3). Covariance difference between Prediction only and VI SLAM

IV-D. Discussion

In this project, I find that different step sizes or different methods to downsample our data may influence the final results in different ways. Generally, the more data we have, the more information we can extract, and thus the more comprehensive the result we are supposed to achieve. But in reality, some information are redundant and sometimes too much information exceed our computation capacity. For example, choose step size as 10 and step 15 may not have a huge impact on the result, but choose step size as 100 will certainly make a huge difference. However, If I choose to utilize the original data, my laptop will run a long time and then return error, which means the data exceed its computation capacity. The balance of the accuracy and the configuration varies from algorithms and hardwares. It requires us to have more tests to keep a balance.