Visual-Inertial SLAM

Muqing Li
Department of Electrical and Computer Engineering
University of California, San Diego
mul003@ucsd.edu

I. INTRODUCTION

It is important for autonomous vehicles or robots to understand their pose and the real environment. With the data collected from stereo cameras and sensors, it is possible to generate mapping during motion. Given the mapping of the environment, vehicles or robots can perform motions without collision in the environment, and as a result, it will be able to plan trajectories more efficiently. A common technique of performing localization and mapping at the same time is called SLAM. In addition, Extended Kalman Filter is a transformed nonlinear type of Bayes filter to estimate and update the landmarks in the world-frame.

II. PROBLEM FORMULATION

Prediction trajectory and the motion model based on the pose $\mu t|t$ at each time step can be derived as:

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t \exp\left(\tau \hat{\mathbf{u}}_t\right)$$

$$\delta\boldsymbol{\mu}_{t+1} = \exp\left(-\tau \hat{\mathbf{u}}_t^{\curlywedge}\right) \delta\boldsymbol{\mu}_t + \mathbf{w}_t$$

where,

$$\mathbf{u}_t := \begin{bmatrix} \mathbf{v}_t \\ \boldsymbol{\omega}_t \end{bmatrix} \in \mathbb{R}^6 \quad \hat{\mathbf{u}}_t := \begin{bmatrix} \hat{\boldsymbol{\omega}}_t & \mathbf{v}_t \\ \mathbf{0}^\top & 0 \end{bmatrix} \in \mathbb{R}^{4\times 4}$$

We can only use the pose kinematics at discretization τ to get the motion model, where the motion model can be derived by using IMU motion date including, linear and angular velocities

The EKF model is a loop closure of prediction and update stages to find the best estimation by:

$$\textbf{Prediction: } p_{t+1|t}(\mathbf{x}) = \int p_f(\mathbf{x} \mid \mathbf{s}, \mathbf{u}_t) p_{t|t}(\mathbf{s}) d\mathbf{s}$$

$$\textbf{Update: } p_{t+1|t+1}(\mathbf{x}) = \frac{p_h(\mathbf{z}_{t+1}|\mathbf{x}) p_{t+1|t}(\mathbf{x})}{p(\mathbf{z}_{t+1}|\mathbf{z}_{0:t}, \mathbf{u}_{0:t})} = \frac{p_h(\mathbf{z}_{t+1}|\mathbf{x}) p_{t+1|t}(\mathbf{x})}{\int p_h(\mathbf{z}_{t+1}|\mathbf{s}) p_{t+1|t}(\mathbf{s}) d\mathbf{s}}$$

After we have the predicted trajectory, landmark positions can be estimated by only applying the EKF update part. Observation is considered static which does not need EKF prediction.

Also, by given the IMU linear acceleration, angular acceleration, and stereo feature coordinates over time t_i, the problem becomes deriving the world-frame IMU pose in the time period t_1- t_n, and the world-frame coordinates over M landmarks, based on the observation model $z_{t,i}$.

Given the observation model:

$$\mathbf{z}_{t,i} = h(T_t, \mathbf{m}_j) + \mathbf{v}_{t,i} := M\pi\left({}_O T_I T_t^{-1} \underline{\mathbf{m}}_j\right) + \mathbf{v}_{t,i} \qquad \mathbf{v}_{t,i} \sim \mathcal{N}(0, V)$$

where,

The probability $\boldsymbol{z_{t,i}}$ is related to the image data from stereo camera's with the IMU pose $\boldsymbol{T_t}$ and landmark $\boldsymbol{m_j}$.

III. TECHNICAL APPROACH

The problem is divided into three main parts: IMU Localization via EKF Prediction; Landmark Mapping via EKF Update; and Visual-Inertial SLAM.

a. Date Initialization

First, we need to initialize all the data from the provided two original image sequence videos. It includes timestamps, features, linear velocity, angular velocity, IMU to camera transformation, baseline, calibration matrix, and baseline.

Then we predict IMU pose μ at *t+1* from time *t* from:

$$\mu_{t+1|t} = \mu_{t|t} \exp(\tau \hat{\mathbf{u}}_t)$$
$$\Sigma_{t+1|t} = \mathbb{E}[\delta\mu_{t+1|t}\delta\mu_{t+1|t}^\top] = \exp\left(-\tau \overset{\wedge}{\mathbf{u}}_t\right)\Sigma_{t|t}\exp\left(-\tau \overset{\wedge}{\mathbf{u}}_t\right)^\top + W$$

where, we use

$$W = \begin{bmatrix} 0.3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.05 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.05 \end{bmatrix}$$

as an initial guess of 0.3's correspond to noise in the linear velocities and 0.05 to noise in the angular velocities.

τ: IMU timestamp and control input difference.

v_t : linear velocity; ω_t : linear velocity.

Other control inputs are from the equation below:

$$\mathbf{u}_t := \begin{bmatrix} \mathbf{v}_t \\ \omega_t \end{bmatrix} \in \mathbb{R}^6 \quad \hat{\mathbf{u}}_t := \begin{bmatrix} \hat{\omega}_t & \mathbf{v}_t \\ \mathbf{0}^\top & 0 \end{bmatrix} \in \mathbb{R}^{4\times 4} \quad \overset{\wedge}{\mathbf{u}}_t := \begin{bmatrix} \hat{\omega}_t & \hat{\mathbf{v}}_t \\ 0 & \hat{\omega}_t \end{bmatrix} \in \mathbb{R}^{6\times 6}$$

b. Landmark Location Update

The next step is to apply EKF to update landmark positions. From prior landmark estimate μt and the new observation Zt+1, we can generate the new landmark estimate μt+1, w.r.t the world-frame. The update follows the equation below:

$$\mu_{t+1} = \mu_t + K_{t+1}(\mathbf{z}_{t+1} - \tilde{\mathbf{z}}_{t+1})$$
$$\Sigma_{t+1} = (I - K_{t+1}H_{t+1})\Sigma_t$$

Observation z_{t+1} and $K_{t+1|t}$ follow:

$$\tilde{\mathbf{z}}_{t+1,i} := M\pi\left(_O T_I \mu_{t+1|t}^{-1} \underline{\mathbf{m}}_j\right) \qquad \text{for } i = 1, \ldots, N_{t+1}$$

$$K_{t+1} = \Sigma_t H_{t+1}^\top \left(H_{t+1}\Sigma_t H_{t+1}^\top + I \otimes V\right)^{-1}$$

where, H_{t+1} is the Jacobian of the new observation z_{t+} . Below is the Jacobian equation:

$$H_{t+1,i} = -M\frac{d\pi}{d\mathbf{q}}\left(_O T_I \mu_{t+1|t}^{-1}\underline{\mathbf{m}}_j\right)\, _O T_I \left(\mu_{t+1|t}^{-1}\underline{\mathbf{m}}_j\right)^\odot \in \mathbb{R}^{4\times 6}$$

Stereo Camera Calibration Matrix M:

$$M := \begin{bmatrix} fs_u & 0 & c_u & 0 \\ 0 & fs_v & c_v & 0 \\ fs_u & 0 & c_u & -fs_u b \\ 0 & fs_v & c_v & 0 \end{bmatrix}$$

f = focal length $[m]$
s_u, s_v = pixel scaling $[pixels/m]$
c_u, c_v = principal point $[pixels]$
b = stereo baseline $[m]$

The derivative of projection function $d\pi/dq$:

$$\pi(\mathbf{q}) := \frac{1}{q_3}\mathbf{q} \in \mathbb{R}^4 \qquad \frac{d\pi}{d\mathbf{q}}(\mathbf{q}) = \frac{1}{q_3}\begin{bmatrix} 1 & 0 & -\frac{q_1}{q_3} & 0 \\ 0 & 1 & -\frac{q_2}{q_3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{q_4}{q_3} & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4}$$

Also, in the Jacobian function, TI represents the transformation from IMU to camera frame, while mj represents homogeneous landmark coordinate.

c. Visual-Inertial Slam:

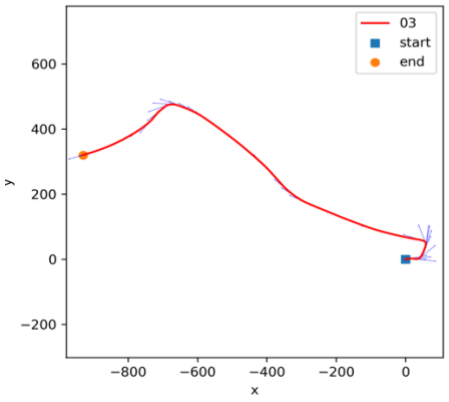
Lastly, we combine the previous two parts, which are the IMU prediction step with the EKF with the landmark update step. SLAM decides between the mapping and SLAM to localize the position at each timestamp to find new landmark locations, which is represented by mean mu and variance σ. Landmarks are uncorrelated in mapping-only stage but correlated in SLAM stage. Since vehicle sensor has minimal movement in z direction, it is not needed to apply prediction to landmarks in Z-axis, which means we only estimate x and y values in landmark coordinates.

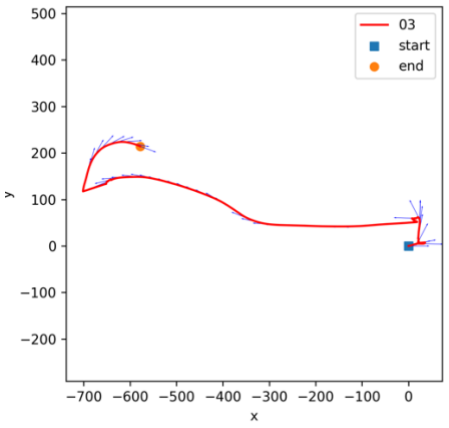
IV. RESULTS

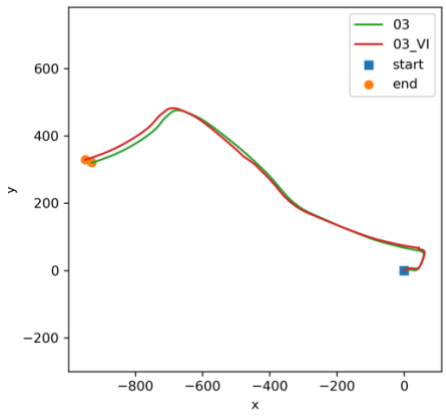
Generally, the implemented Visual-Inertial SLAM worked well in generating a smoother trajectory. Other than previous frame transformation and apply feature data, we have a better understanding of combining the prediction and update together in EKF. The limitation is mainly the processing speed. The processing speed is not fast enough to let the algorithm be applied in real world, like the autonomous driving scenarios.

V. VISUALIZATION

03.NPZ Dataset RESULT







10.NPZ Dataset Result