Visual Inertial SLAM

Huafeng mai
University of Sandiego
La Jolla, CA 92122
Email: humai@ucsd.edu

***Abstract*—This paper come up with a solution to solve the problem of simultaneous localization and maximization (SLAM) through visual inertial sensor configuration and the Extended Kalman Filter algorithm (EKF). EKF is to localize a vehicle based on IMU odometry and visual key-points of camera's data and there is IMU on the robot to collect the movement information, And after implementing EKF, I present the result and dicussion.**

I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) is an crucial topic in Robot motion planning. It is about updating the mapping and its own positioning parameters in an unknown environment [1]. There are many ways to deal with slam problems such as use Particle Filter or combine the neural network method [2]. In localization, we estimate the vehicle's pose given IMU data and so the cycle continues, In this paper, I implement EKF SLAM to create a map of a robot around environment during it drives and track it's position call Visual Inertial SLAM. for Visual Inertial SLAM, need to gain a 2 dimensional understanding of the world from a stereo camera and IMU data. The method use Kalman filters is to promote the performance, about estimate the landmark positions and contain two steps, prediction and update steps to keep track a robot over time, use this method can effectively model a robot's location in the world and since based on the assumption of a Gaussian distribution on current observations and previous inputs. Variance gradually decreases over loop, so that the predictions have a higher tendency to match the real situation. more accurate and faster to track a robot's position than Particle Filter SLAM.

This project is divided into three parts. The first part is IMU based Localization via EKF Prediction which predict where future locations vehicle will be. The second part is Landmark Mapping via EKF Update which updates and estimates where landmarks are on the path in the world frame. The last part is Visual Inertial SLAM which combines the previous two part and add an IMU update step based on the stereo camera observation model to obtain a complete visual inertial SLAM, finally show the estimated landmark positions and vehicle trajectory on a specific map.

The paper structure is as follows. Give the detailed mathematics formulations of SLAM problem in Section II. Technical approaches are introduced in Section III. And at last setup the algorithm, results are presented in Section IV.

II. PROBLEM FORMULATION

The problem is how to localize the vehicle pose over time and mapping a map with its surrounding landmarks and trajectory.

For Visual Inertial Localization problem we have IMU data, which contain three inputs IMU angular acceleration, IMU linear acceleration $a_t \in R^3$, angular velocity $\omega_t \in R^3$ and detected features from stereo camera which include pixel feature coordinates in both left and right stereo images across time. The IMU to camera optical frame transformation${}_O T_I \in \mathrm{SE}(3)$, stereo baseline b and camera calibration matrix K is:

$$K = \begin{bmatrix} fs_u & 0 & c_u \\ 0 & fs_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

if we are not have observable feature at time t have an associated measurement of $\mathbf{z}_{t,i} = \begin{bmatrix} -1 & -1 & -1 & -1 \end{bmatrix}^\top$, So we can obtain the stereo camera calibration matrix M , Its matrix representation is as follows.

$$M := \begin{bmatrix} fs_u & 0 & c_u & 0 \\ 0 & f_{s_v} & c_v & 0 \\ f_{s_u} & 0 & c_u & -f_{s_u}b \\ 0 & f_{s_v} & c_v & 0 \end{bmatrix} \tag{2}$$

f is focal length, s_u, s_v is pixel scaling, cu, cv is principal point.

for the world frame IMU pose, sovle for the ${}_W T_I \in \mathrm{SE}(3)$ over time and for each world-frame coordinates of the landmarks point W_i, generated the visual features $f_{t,i} \in \mathcal{R}^4$.

A. Visual Mapping

The Objective of Visual Mapping is given the observations $\mathbf{z}_t := \begin{bmatrix} \mathbf{z}_{t,1}^\top & \cdots & \mathbf{z}_{t,N_t}^\top \end{bmatrix}^\top \in \mathbb{R}^{4N_t}$ for t = 0, . . . , T, estimate the coordinates m $\mathbf{m} := \begin{bmatrix} \mathbf{m}_1^\top & \cdots & \mathbf{m}_M^\top \end{bmatrix}^\top \in \mathbb{R}^{3M}$ of the landmarks that generated them. we want to represent the environment by where the vehicle is within the sensor observations have add noise. We want to first locate the vehicle and then we can obtain observations at a specific moment t and transalted into landmark $W_{i:t}$ in the world frame. Since the sensor does not move sufficiently along the z-axis, the estimation of the z coordinate of the landmarks will not be very good. For the visual extracted feature $f_{i:t}$ by the stereo camera, each feature contain element pixel cordinates (o_L, o_l, r_R, r_R). we want to estimate the coordinates of landmarks $W_{i:t}$ in world frame. For t is from 1 to n. Suppose l_t are landmark indices at o_t, $c_i$is the vehicle pose and k is landmark.

$$p\left(o_t \mid c_t, k, l_t\right) \tag{3}$$

B. Localization

provide the position of landmark in the environment, locate and estimate the trajectory of the vehicle, which is given a series of control inputs and sensor measurements to estimate the vehicle trajectory, we keep a pmf of the vehicle state and prediction step is performed using the motion model to obtain the predicted pmf and update step will perfrom by using the observation model and the weight need to be scaled, however, the particle poses do not change.

$$\begin{gathered} p_{t|t-1}\left(x_t\right) \\ p_{t|t}\left(x_t\right) \end{gathered} \tag{4}$$

In order to apply the Extended Kalman Filter (EKF) to this, A Bayes filter with the following assumptions:
1. The prior pdf $p_{t|t}$ is Gaussian
2. The motion model is linear in the state with Gaussian noise
3.The observation model add with with Gaussian noise
4. The motion noise wt and observation noise vt are independent of each other, of the state xt and across time

III. TECHNICAL APPROACH

The three main problem we have to solve in this paper are follow:
1. IMU Localization using EKF Prediction
2. Landmark Mapping via EKF Update
3. Visual-Inertial SLAM

The EKF prediction step based on the linear and angular velocity measurements and SE(3) kinematics to estimate the pose $T_t \in SE(3)$ over time and suppose we know the homogenous coordinates of the landmarks $m \in \mathbb{R}^{4\times \mathrm{M}}$ in the world frame. and the IMU measurement is w_t, and given the visual feature observations γ_t, to estimate the inverse IMU pose $U_t = {}_wT_i^{-1}$, IMU's time step pose is $\mu_{t|t} \epsilon R^{4X4}$ and $\Sigma_{t|t} \epsilon R^{6X6}$ since only have six degree of freedom which we have

$$T_t = \mu_{t|t} \exp\left(\hat{\delta}\mu_{t|t}\right)$$

A. IMU Localization using EKF Prediction

Motion Model: nominal kinematics of $\mu_{t|t}$ and perturbation kinematics of $\delta\mu_{t|t}$ with time discretization t :

$$\begin{aligned} \boldsymbol{\mu}_{t+1|t} &= \boldsymbol{\mu}_{t|t} \exp\left(\tau_t \hat{\mathbf{u}}_t\right) \\ \delta\boldsymbol{\mu}_{t+1|t} &= \exp\left(-\tau_t \mathbf{u}_t\right) \delta\boldsymbol{\mu}_{t|t} + \mathbf{w}_t \end{aligned} \tag{5}$$

The EKF prediction step is

$$\begin{aligned} \boldsymbol{\mu}_{t+1|t} &= \exp\left(-\eta \hat{\mathbf{u}}_t\right) \boldsymbol{\mu}_{t|t} \\ \Sigma_{t+1|t} &= \mathbb{E}\left[\delta\mu_{t+1|t}\delta\mu_{t+1|t}^{\top}\right] \\ &= w + \exp\left(-\eta \hat{u}_t\right) \Sigma_{t|t} \exp\left(-\eta \hat{u}_t\right)^{\top} \end{aligned} \tag{6}$$

η is the difference between previous timestamps and current timestamps. ω is add gaussian noise

$$\begin{aligned} \mathbf{u}_t &:= \left[\begin{array}{c} \mathbf{v}_t \\ \boldsymbol{\omega}_t \end{array}\right] \in \mathbb{R}^6 \\ \hat{\mathbf{u}}_t &:= \left[\begin{array}{cc} \hat{\boldsymbol{\omega}}_t & \mathbf{v}_t \\ \mathbf{0}^{\top} & 0 \end{array}\right] \in \mathbb{R}^{4\times 4} \\ \check{\mathbf{u}}_t &:= \left[\begin{array}{cc} \hat{\boldsymbol{\omega}}_t & \hat{\mathbf{v}}_t \\ 0 & \hat{\boldsymbol{\omega}}_t \end{array}\right] \in \mathbb{R}^{6\times 6} \end{aligned} \tag{7}$$

ω_t is angular velocities, vt is the linear velocities, $\hat{u}$ is the pose space transforming by 6D vector into a 4 by 4 matrix. $\check{u}$ is a 6x6 matrix and comes from the adjoint space because it's needed to update the covariance matrix. The transforms between the world frames and camera is.

$$\begin{gathered} {}_ct_w = \mu_{t+1|t c}T_i \\ {}_wt_c = {}_cT_w^{-1} \end{gathered} \tag{8}$$

B. Landmark Mapping via EKF Update

Suppose the IMU pose over time and the feature observation γ_t is known and we want to estimate the homogeneous coordinates in the word frame of landmarks, and update the pose mean and covariance. Although the observation model is same in visual mapping problem, the variable of interest is not landmark positions $m \in \mathbb{R}^3$, show the prior is

$$T_t \mid \gamma_{0:t}, \mathbf{u}_{0:t-1} \sim \mathcal{N}\left(\boldsymbol{\mu}_{t|t}, \Sigma_{t|t}\right) \tag{9}$$

$$\begin{gathered} p_{t|t-1}\left(x_t\right) \\ p_{t|t}\left(x_t\right) \end{gathered} \tag{10}$$

The EKF update step is [3]:

$$\begin{aligned} K_{t+1} &= \Sigma_{t+1|t} H_{t+1}^{\top}\left(H_{t+1}\Sigma_{t+1|t}H_{t+1}^{\top} + I \otimes V\right)^{-1} \\ \mu_{t+1|t+1} &= \mu_{t+1|t} \exp\left(\left(K_{t+1}\left(\mathbf{z}_{t+1} - \tilde{\mathbf{z}}_{t+1}\right)\right)^{\wedge}\right) \\ \Sigma_{t+1|t+1} &= \left(I - K_{t+1}H_{t+1}\right)\Sigma_{t+1|t} \end{aligned} \tag{11}$$

$$I \otimes V := \left[\begin{array}{ccc} V & & \\ & \ddots & \\ & & V \end{array}\right] \tag{12}$$

H_{t+1} is the Jacobian of z_t evaluated at $\mu_{t+1|t}$ corresponding to $T_{t+1} \in SE(3)$. μ is the updated mean and Σ is the update covariances and the Kalman gain is

$$K_{t+1|t} = \left(H_{t+1}\Sigma_{t+1|t}H_{t+1}^T + R_{t+1}VR_{t+1}^T\right)^{-1}\Sigma_{t+1|t}H_{t+1}^T$$

C. Visual Inertial SLAM

The purpose for this part is estimating pose of the vehicle and position of landmarks at the same time, the landmark positions m_t or the pose T_t is known the method is to combine the predict steps and update steps of visual inertial odometry and Extended Kalman Filter visual mapping. Merge the mean and covariance of IMU Pose and landmarks position. The estimated state and covariance under the Gaussian assumption is:

$$\Sigma = \begin{bmatrix} \Sigma_k & C \\ C^T & \Sigma_l \end{bmatrix} \in \mathbb{R}^{3M+6\times 3M+6}$$
$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_k \\ \boldsymbol{\mu}_l \end{bmatrix} \in \mathbb{R}^{(3m+6)} \tag{13}$$

for Σ_k and c is the mean of landmark and cross covariance we estimate the μ_k landmark position and estimate μ_l six degree of freedom of inverse IMU pose. For we know the covariance of landmarks and pose of IMU are combine to joint covariance, So they will perform correlated during the update process.

The prediction step of motion model is same as before, the difference is we need Jacobian Matrix to update IMU pose and landmark positions at the same time.

$$H_{t+1|t} = \left[H_{L,t+1|t} H_{imu,t+1|t}\right] \in \mathbb{R}^{4N_t\times(3M+6)} \tag{14}$$

For the IMU measurement u_t the equations are listed as follows.

$$\mathbf{w} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_p \end{bmatrix}$$
$$\boldsymbol{\mu}_{t+1|t} = \begin{bmatrix} \boldsymbol{\mu}_{m,t+1|t} \\ \exp\left(-\tau \mathbf{u}_t^{\wedge}\right) \boldsymbol{\mu}_{p,t|t} \end{bmatrix} \tag{15}$$
$$\mathbf{D}_t = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \exp\left(-\tau \mathbf{u}_t^{\curlywedge}\right) \end{bmatrix}$$
$$\boldsymbol{\Sigma}_{t+1|t} = \mathbf{w} + \mathbf{D}_t \boldsymbol{\Sigma}_{t|t} \mathbf{D}_t^T$$

The update step conatin visual inertial odometry and visual mapping update step

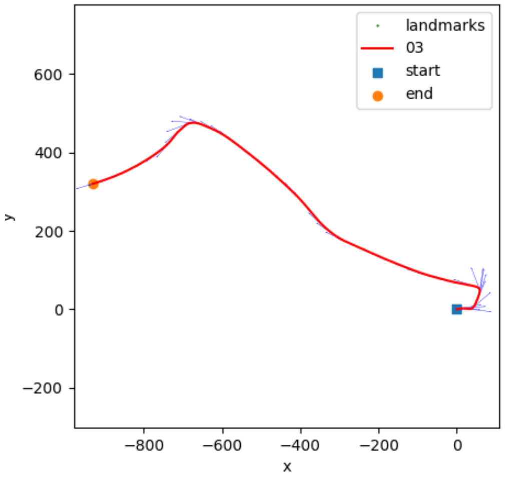
$$\bar{z}_{t,i} = C\pi(\mu_{t,i}^{Landmark} * T_{IC}) \tag{16}$$

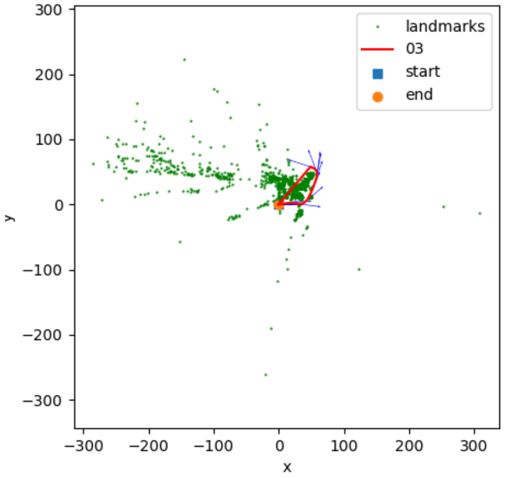
T_{IC} is the IMU to camera frame transformation matrix, μ is the mean of the i landmark pose, C is stereo camera calibration matrix.
The EKF Update step is

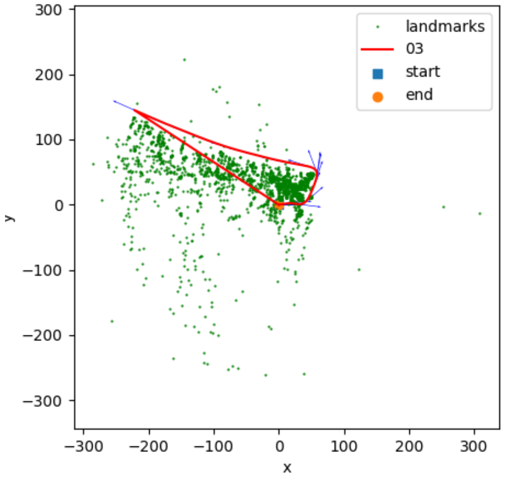
$$\mathbf{K}_{t+1|t} = \begin{bmatrix} \Sigma_L & C \\ C^T & \Sigma_{IMU} \end{bmatrix} \begin{bmatrix} H_L^T \\ H_{IMU}^T \end{bmatrix} S^{-1}$$
$$\boldsymbol{\mu}_{t+1|t+1} = \begin{bmatrix} \boldsymbol{\mu}_{m,t+1|t} + \mathbf{K}_{t+1|t}\left(\mathbf{z}_t - \tilde{\mathbf{z}}_t\right) \\ \exp\left(\left(\mathbf{K}_{t+1|t}\left(\mathbf{z}_{t+1} - \tilde{\mathbf{z}}_{t+1}\right)\right)^{\wedge}\right) \boldsymbol{\mu}_{p,t+1|t} \end{bmatrix}$$
$$\boldsymbol{\Sigma}_{t+1|t+1} = \left(\mathbf{I} - \mathbf{K}_{t+1|t}\mathbf{H}_{t+1|t}\right) \boldsymbol{\Sigma}_{t+1|t}$$
(17)

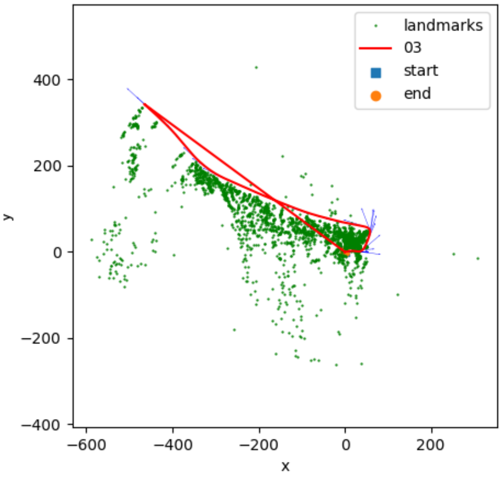
IV. Result and Analysis

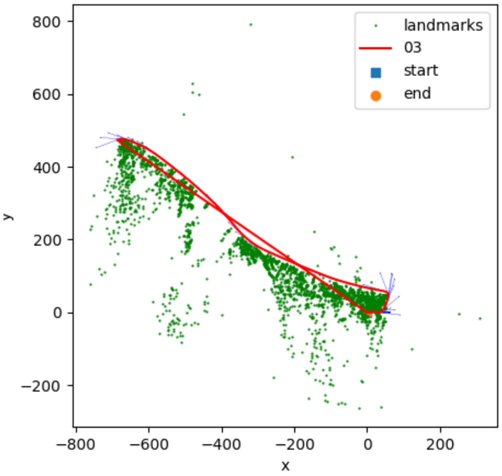
The EKF based SLAM algorithm is tested with 2 data, I start by test our code with dead reckoning to plot the estimated vehicle IMU trajectory, on the other hand, I output the pictures displayed at different iteration times. Set up the process noise covariance ω_p at 0.0003 and observation noise covariance V at 4200*I and when change the observation noise covariance the result show it quite robost for this variable.
Show the 03 data image follow by trajectory without landmark, trajectory with landmark with iteration 200, 400, 600, 800 and final result

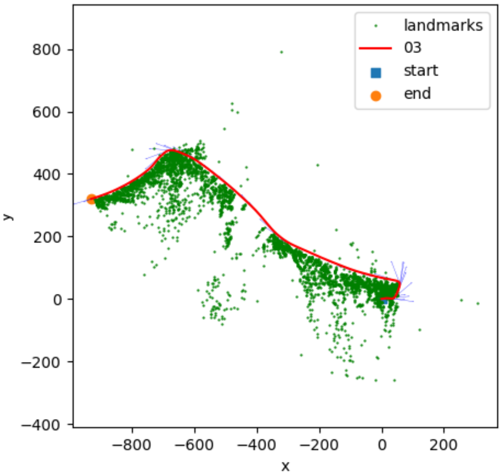




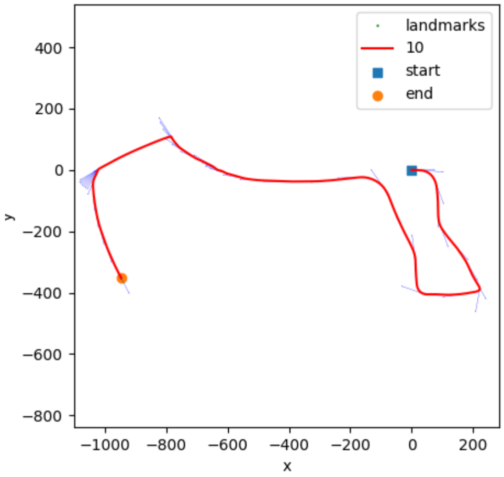


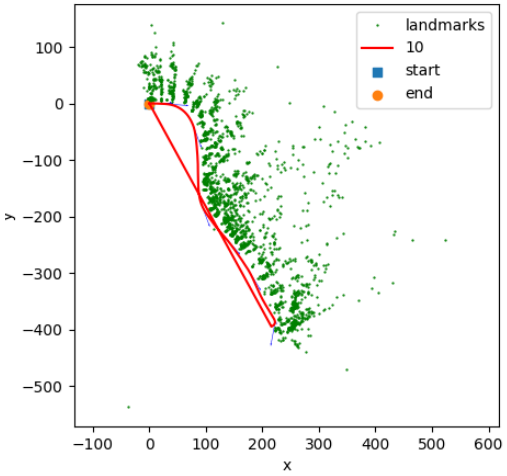


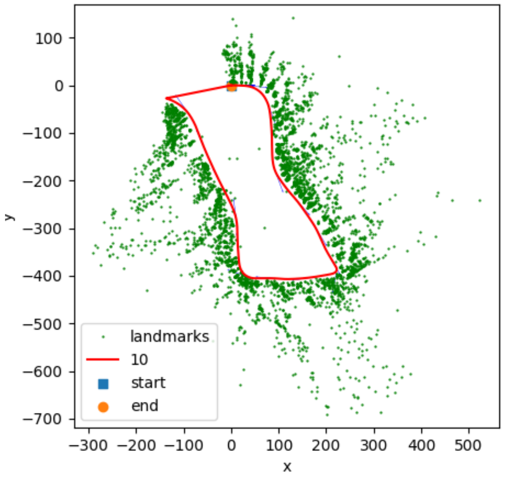


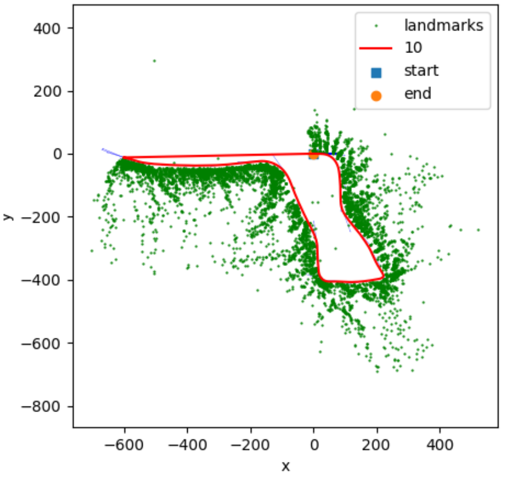


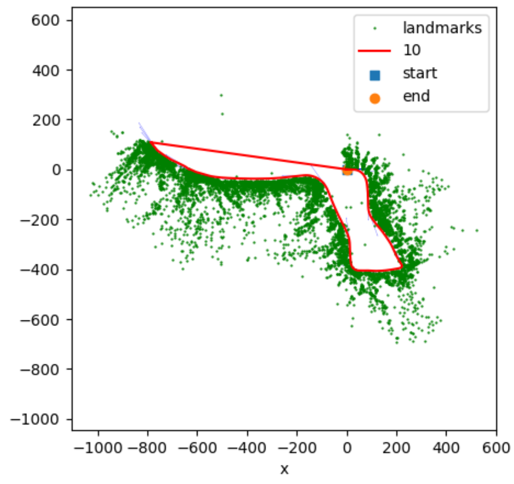
Show the 10 data image follow by trajectory without landmark, trajectory with landmark with iteration 500, 1000, 1500, 2000, 2500, 3000 and final result

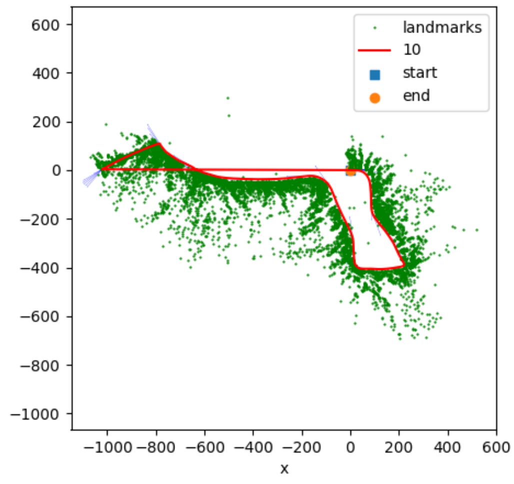


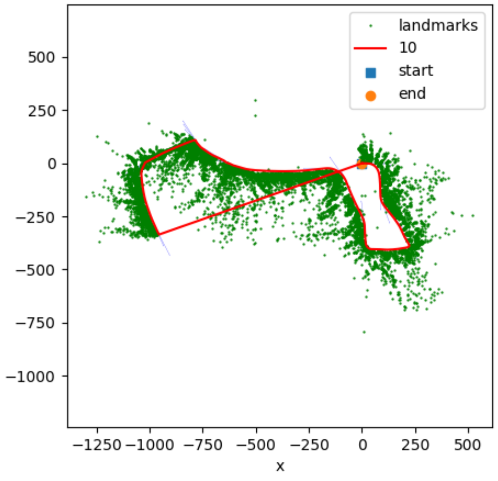


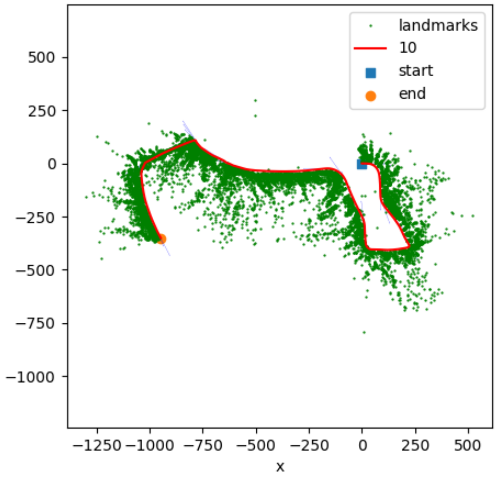












In conclustion, the result looks reasonable with the vehicle trajectory being consistent with the vehicle show in video. Stereo camera make us can predict the trajectory of the vehicle and IMU pose prediction step is a crucial part of Visual inertial SLAM, which helpful for localization by get the map and then finally update the pose of both the IMU and the landmarks.

References

[1] Beril Sirmaccek, Nicolò Botteghi, and Mustafa Khaled, “Reinforcement learning and slam based approach for mobile robot navigation in unknown environments,” in ISPRS Workshop Indoor 3D 2019, 2019.

[2] Tateno, K., Tombari, F., Laina, I. and Navab, N., 2017. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6243-6252).

[3] Nikolay Atanasov. Ece276a: Sensing estimation in robotics.