Visual-Inertial SLAM

Boyang Zhang
Department of Electrical Engineering
University of California, San Diego
boz083@ucsd.edu

***Abstract*—This project aims to locate the robot in the environment to record its trajectory and generate a 2D mapping of visual features extracted from stereo RGB images. The project first approaches the tracking and mapping problems separately, then combines the two parts using visual-inertial SLAM approach.**

Keywords—Visual-inertial localization, 2D visual features mapping, visual-inertial SLAM, extended Kalman filter

I. INTRODUCTION

Visual inertial slam uses IMU measurements, including linear and rotational velocities, and visual features pixel coordinate from stereo RGB images to locate the robot in the world frame and find the coordinates of the landmarks that generated visual features. This approach to SLAM requires only an IMU and stereo camera. It is a simpler setup compared to SLAM approaches. [1]

The project consists of two parts. In the first part, we locate the robot pose and then generate the mapping of visual features separately. In the localization step, we assume the pose of the robot overtime has Gaussian distribution and use the extended Kalman filter to predict the mean and covariance of the distribution based on the motion model. In the visual mapping step, we assume the visual feature's coordinates in the world frame also has Gaussian distribution and the robot pose is known. We then use the extended Kalman filter to update the mean and covariance of the distribution based on the observation model.

For the second part, we apply both the prediction and update steps together, as well as an additional update step for the robot pose based on the observations. This requires to update both mean and covariance from the two distributions together. We also assume the visual features are static and therefore the prediction step still

II. PROBLEM FORMULATION

A. IMU Localization

In this part of the project, we consider the localization problem only. We are given the IMU measurements of linear and angular velocities at a given time t:

$$u_t := [\mathbf{v}_t^T,\ \omega_t^T]^T$$

and want to predict the inverse IMU pose over time:

$$T_t := {}_W T_{I,t}^{-1} \in SE(3)$$

Using the extend Kalman filter approach, Tt has a Gaussian prior:

$$T_t | z_{0:t},\ u_{0:t-1} \sim \mathcal{N}(\boldsymbol{\mu}_{t|t}, \Sigma_{t|t}) \text{ with } \boldsymbol{\mu}_{t|t} \in SE(3) \text{ and } \Sigma_{t|t} \in \mathbb{R}^{6\times 6}$$

The problem then becomes applying the prediction step of EKF with motion model:

$$T_{t+1} = \exp\left((\tau(-u_t + w_t))^\wedge\right) T_t \qquad u_t := \begin{bmatrix} v_t \\ \omega_t \end{bmatrix}$$

with time discretion τ and zero-mean Gaussian noise w_t.

We need to predict both the mean and covariance for the IMU pose.

B. Landmark Mapping

For the problem of visual landmark mapping, we assume the IMU pose over time Tt is known, and the landmarks are static. We are given the visual feature observations z_t over time, calibration matrix M and oTi pose from IMU to camera.

The objective is to find the landmarks' associated homogeneous coordinates in the world frame: $\mathbf{m} \in \mathbb{R}^{4\times M}$

Again, using EKF, we have the landmark coordinate's prior:

$$\mathbf{m} \mid \mathbf{z}_{0:t} \sim \mathcal{N}(\boldsymbol{\mu}_t, \Sigma_t) \text{ with } \boldsymbol{\mu}_t \in \mathbb{R}^{4\times M} \text{ and } \Sigma_t \in \mathbb{R}^{3M\times 3M}$$

The problem for landmark mapping becomes applying the update step of EKF with observation model:

$$\mathbf{z}_{t,i} = h(T_t, \mathbf{m}_j) + v_t := M\pi\left({}_O T_I\, T_t \mathbf{m}_j\right) + v_t$$
$$\text{with measurement noise } v_t \sim \mathcal{N}(0, V)$$

We need to update both the mean and covariance for the landmark coordinates.

C. Visual-Inertial SLAM

For visual-inertial SLAM, we are given the IMU measurement, visual feature observations, calibration matrix M and oTi pose from IMU to camera.

The objective is to apply both the update and prediction steps at a given time t, to find the IMU pose and landmark coordinates. We also need to combine the mean and covariance of the IMU position and landmark positions for the update step.

III. TECHNICAL APPROACHES

A. IMU Localization via EKF Prediction

We first initialize the IMU pose in SE(3) with 4x4 identity matrix, as well as the covariance with identity. This correspond to setting the initial robot position as the origin in the world frame.

For each IMU reading at time t, we predict the robot pose following the EKF prediction step:

$$\mu_{t+1|t} = \exp\left(-\tau \hat{u}_t\right) \mu_{t|t} \qquad u_t := \begin{bmatrix} v_t \\ \omega_t \end{bmatrix}$$

$$\Sigma_{t+1|t} = \mathbb{E}[\xi_{t+1|t}\xi_{t+1|t}^T] = \exp\left(-\tau \hat{\hat{u}}_t\right) \Sigma_{t|t} \exp\left(-\tau \hat{\hat{u}}_t\right)^T + \tau^2 W$$

where we use Rodrigues formula to get the exponential map from se(3) to SE(3), and adjoints of SE(3).

We store the mean at each time t as inverse IMU pose. To observe the predicted IMU trajectory, we plot out all inverse of the means.

B. Landmark Mapping via EKF Update

To initialize the coordinates for each landmark, we use the first observation and find the coordinate in the camera frame by solving the inverse of the stereo camera model, which gives:

$$z = f_{su} \text{ x } b \text{ / } (U_L\text{-}U_R)$$
$$x = Z \text{ x } (U_L\text{-}c_u) \text{ / } f_{su}$$
$$y = Z \text{ x } (V_L\text{-}c_v) \text{ / } f_{sv}$$

since we have

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

To initialize the covariance for the first observed coordinate, once the mean is initialized, we set the covariance to be the identity matrix times the norm of the covariance.

For each coordinate, we can obtain the predicted observation based on the mean:

$$\hat{\mathbf{z}}_{t,i} := M\pi\left({}_O T_I \, T_t \mu_{t,j}\right) \in \mathbb{R}^4 \qquad \text{for } i = 1, \ldots, N_t$$

Once the coordinate has been initialized, in order to perform the EKF update for both the mean and covariance, we need to compute the Jacobian of the predicted observation with respect to the current feature m_j observed evaluated at the corresponding landmark mean following:

$$H_{i,j,t} = \begin{cases} M\frac{d\pi}{d\mathbf{q}}\left({}_O T_I \, T_t \mu_{t,j}\right) {}_O T_I \, T_t D & \text{if observation } i \text{ corresponds to} \\ & \text{landmark } j \text{ at time } t \\ \mathbf{0} \in \mathbb{R}^{4\times 3} & \text{otherwise} \end{cases}$$

Then we can perform the EKF update following:

$$K_t = \Sigma_t H_t^T \left(H_t \Sigma_t H_t^T + I \otimes V\right)^{-1} \qquad I \otimes V := \begin{bmatrix} V & & \\ & \ddots & \\ & & V \end{bmatrix}$$
$$\mu_{t+1} = \mu_t + DK_t\left(\mathbf{z}_t - \hat{\mathbf{z}}_t\right)$$
$$\Sigma_{t+1} = (I - K_t H_t)\Sigma_t$$

We use the final mean after all updates as the landmark coordinates in the world frame.

C. Visual-Inertial SLAM

For applying SLAM, similar to localization only above, we initialize the IMU pose's mean and covariance using identity matrices.

Then, at each time t, we perform the update step.

For localization update, we need to compute another Jacobian of the predicted observation following:

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

where $\mathbf{m}^{\odot} = \begin{bmatrix} s \\ \lambda \end{bmatrix}^{\odot} = \begin{bmatrix} \lambda I & -\hat{s} \\ \mathbf{0}^T & \mathbf{0}^T \end{bmatrix} \in \mathbb{R}^{4\times 6}$

Then, since the IMU pose and landmark positions are correlated, we need to stack the two Jacobians together. For each Ht has dimension 4Nt x (3M+6)

Then we follow the same step to perform EKF update:

$$K_{t+1|t} = \Sigma_{t+1|t} H_{t+1|t}^T \left(H_{t+1|t}\Sigma_{t+1|t}H_{t+1|t}^T + I \otimes V\right)^{-1}$$
$$\mu_{t+1|t+1} = \exp\left(\left(K_{t+1|t}(\mathbf{z}_{t+1} - \hat{\mathbf{z}}_{t+1})\right)^{\wedge}\right) \mu_{t+1|t} \qquad H_{t+1|t} = \begin{bmatrix} H_{1,t+1|t} \\ \vdots \\ H_{N_{t+1},t+1|t} \end{bmatrix}$$
$$\Sigma_{t+1|t+1} = (I - K_{t+1|t}H_{t+1|t})\Sigma_{t+1|t}$$

where the mean for IMU and landmarks are updated separately using the corresponding rows of K.

The covariance is updated together since they are not independent of each other.

Once the mean and covariance are updated, we use EKF prediction as before to predict the next IMU pose.

Iterate over all samples then we use the IMU mean as pose and landmark mean as homogeneous coordinates in the world frame.

IV. RESULTS

A. IMU Localization and Landmark Mapping (Separately)

Figures below show the IMU trajectory and landmark locations on xy-plane using IMU localization and landmark mapping separately.

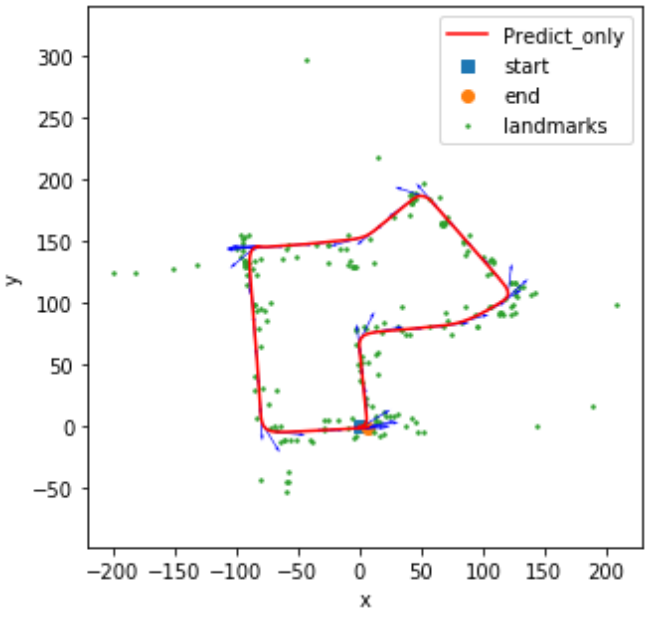

Fig. 1. IMU Trajectory and Landmarks in Dataset 27 (Separate)

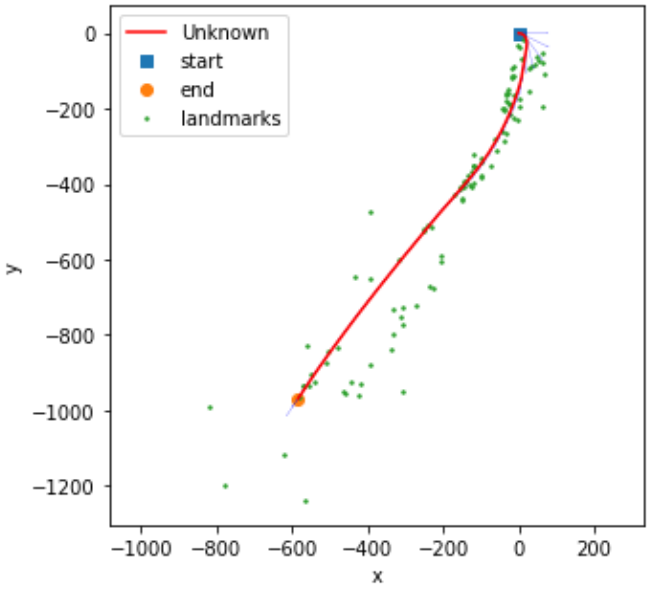

Fig. 2. IMU Trajectory and Landmarks in Dataset 42 (Separate)

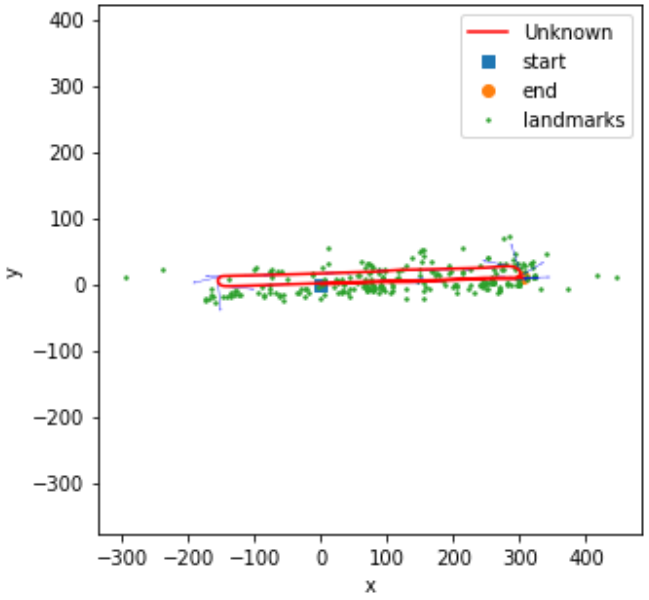

Fig. 3. IMU Trajectory and Landmarks in Dataset 20 (Separate)

The algorithm performs relatively well. The trajectories and landmark positions seem to agree with the video that is used to generate the dataset. Since we do not have ground true for trajectory or landmark position, it is hard to evaluate numerically.

B. Visual-Inertial SLAM

Figures below show the IMU trajectory and landmark locations on xy-plane using visual-inertial SLAM approach.

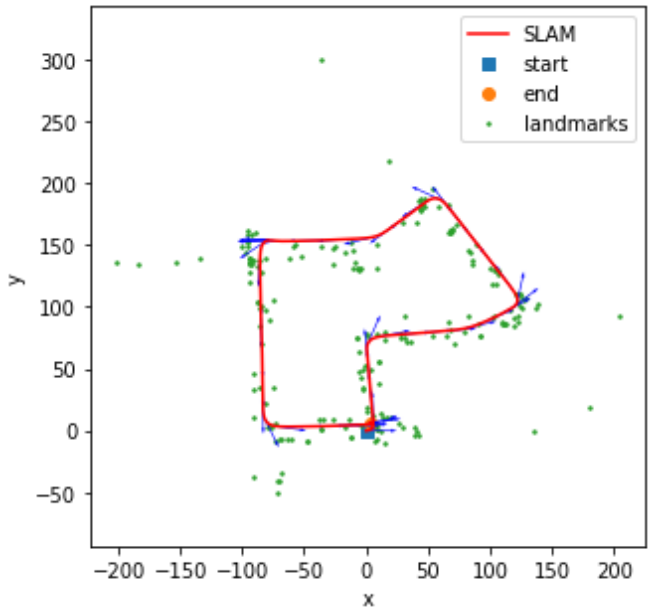

Fig. 4. IMU Trajectory and Landmarks in Dataset 27 (SLAM)

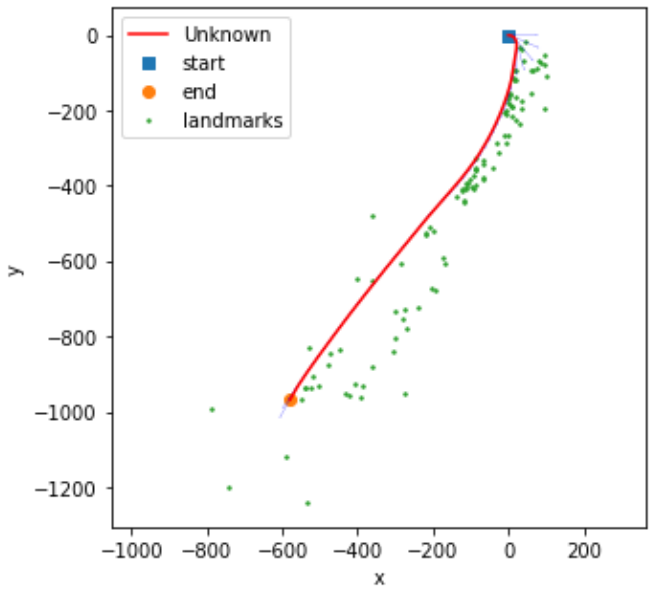

Fig. 5. IMU Trajectory and Landmarks in Dataset 42 (SLAM)

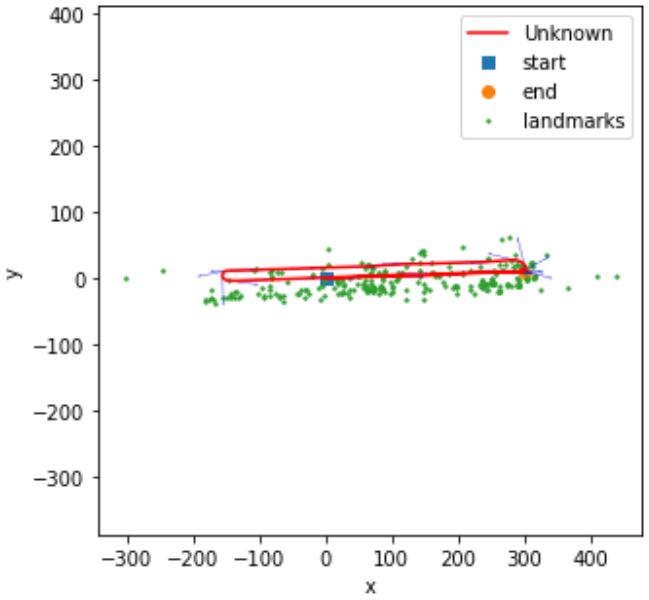

Fig. 6. IMU Trajectory and Landmarks in Dataset 20 (SLAM)

All 3 datasets yield satisfactory result similar to the ones in previous sections. However, in both cases the noise for the IMU model is small, since we have high confidence in the accuracy of the IMU data. In such cases, the benefit of using SLAM is not as prominent.

C. *Performance Comparison*

Figures below compare the performances of the two approaches when the noise for the IMU and visual feature observations are artificially increased. The IMU trajectories and landmark locations are shown below.

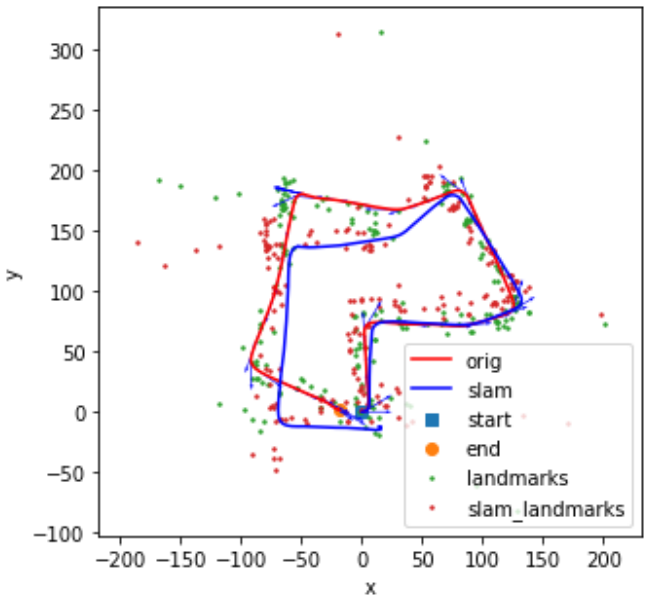

Fig. 7. IMU Trajectories and Landmarks in Dataset 27

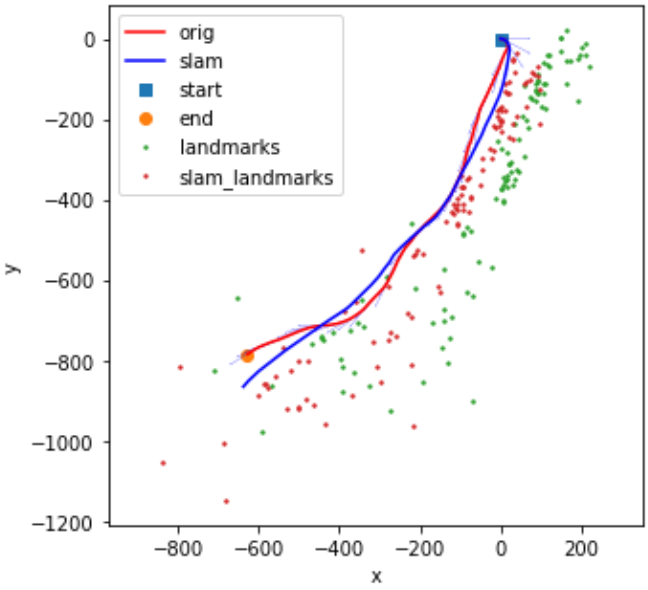

Fig. 8. IMU Trajectories and Landmarks in Dataset 42

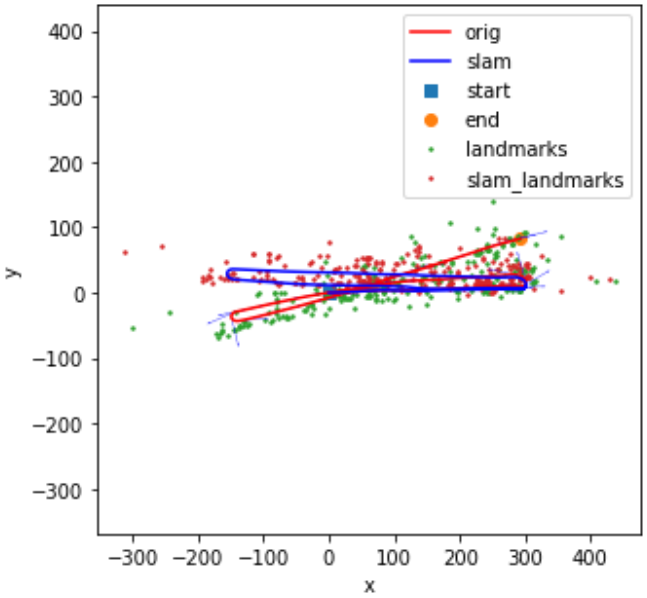

Fig. 9. IMU Trajectories and Landmarks in Dataset 20

When the noise is artificially increased, both approaches performance degrades. However, since the SLAM also updates the IMU pose after each observation, the degradation is not as severe as the approach performing localization and mapping separately. Both trajectories and landmark positions for SLAM have better resemblance to the correct positions as seen in the videos.

There are also cases where both approaches fail, as seen in figures below.

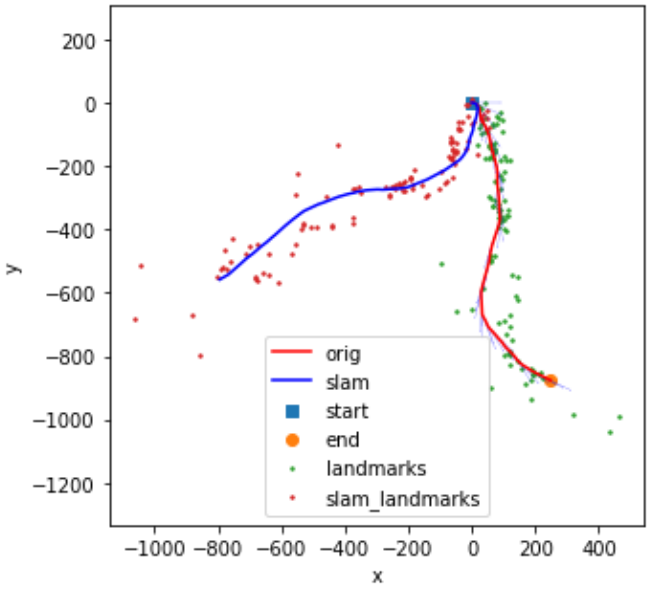

Fig. 10. IMU Trajectories and Landmarks in Dataset 42

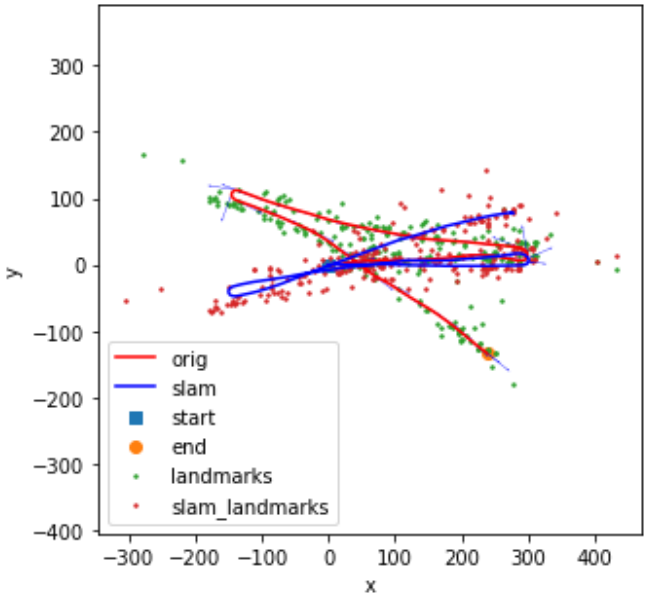

Fig. 11. IMU Trajectories and Landmarks in Dataset 20

This is likely caused by moving landmarks in the dataset. For both approaches, we assume all landmarks are static. However, from observing the video for the dataset, there are multiple landmarks from moving vehicles. Updating using these observations will cause the trajectories in SLAM to divert from the correct direction.

D. Conclusion

Both approaches perform well when the magnitude of noise is chosen appropriately. However, when higher noise is present, SLAM approach is more robust.

One future improvement for the algorithm is adding a prediction steps for the landmarks. As we have seen, the assumption that the landmarks are static is not always correct. Adding a prediction step will likely solve this problem.

REFERENCES

[1] UCSD ECE276A Lec 15