Visual-Inertial Simultaneous Localization and Mapping (SLAM) Using the Extended Kalman Filter (EKF)

Jingpei Lu
Jacobs School of Engineering
University of California, San Diego
jil360@ucsd.edu

***Abstract*—This paper presented my work of implement the visual-inertial SLAM using extended Kalman filter to implement. Our goal is to use the given IMU measurements and the features extracted from the stereo cameras to localize our robot and update the feature map simultaneously. In our implementation, we will perform IMU localization via EKF prediction and landmark mapping via EKF update. Then we will combine those and IMU update to obtain the complete visual-inertial SLAM algorithm.**

Keywords—Visual-Inertial SLAM, Extended Kalman Filter, IMU localization, Landmark mapping

I. Introduction

Simultaneous localization and mapping (SLAM) is a fundamental and important problem in robotic mapping and navigation. It is the computational problem of constructing or updating a map of an unknown environment while simultaneously keeping track of an agent's location within it. Its applications including self-driving cars, unmanned aerial vehicles, and autonomous underwater vehicles.

Nowadays, there are several popular algorithms known for solving or providing approximate solution for the SLAM problem. For example, Particle filter, Kalman filter, and GraphSLAM. The Kalman filter is introduced by R. E. Kalman between 1959 and 1961 [1][2][3]. The EKF is the nonlinear version of the Kalman filter which linearizes about an estimate of the current mean and covariance. Using the EKF, we are able to solve the localization and mapping problem in the case of well-defined transition model. In this paper, we are going to focus on implementing the visual-inertial SLAM using EKF prediction and EFK update. Specifically, we are going to use the IMU measurements to predict the agent's pose and use the landmarks to localize the agent.

Our approach of solving the SLAM problem can be divided into two major parts, prediction and update. In prediction step, we are going to implement the EKF prediction step based on the $SE(3)$ kinematics to estimate the pose $T_t \in SE(3)$ of the IMU over time t. In the update step, we are going to implement EKF with the unknown landmark positions $m \in \mathbb{R}^{3*M}$ as a state and perform EKF update after every visual observation z_t in order to keep track of the mean and covariance of m. Finally, we implement complete visual-inertial SLAM by performing the IMU prediction step and IMU update step with the landmark update step to obtain the more accurate estimation of robot's pose over time.

II. Problem Formulation

A. Locolization Problem

Our data set consists synchronized measurements from an IMU $D_u = \{u_1, u_2, \dots, u_N\}$. Here, N is the number of different timestamps, $u_t \in \mathbb{R}^6$ which includes the measurements of linear velocity $v_t \in \mathbb{R}^3$ and rotational velocity $\omega_t \in \mathbb{R}^3$. We also have the features extracted from a stereo camera $D_z = \{z_{1,1}, z_{2,1}, \dots, z_{M,N}\}$. Here, N is the number of different timestamps, M is the number of different features, $z_{i,t} \in \mathbb{R}^4$ which consists the image coordinates of the stereo camera.

We also have some assumptions that are made: the transformation $T_{ci} \in SE(3)$ from the IMU to the camera optical frame (extrinsic parameters) and the stereo camera calibration matrix M (intrinsic parameters) are known; the homogeneous coordinates $m_i \in \mathbb{R}^4$ in the world frame of the landmarks are known; the data association $\pi_i: \{1, \dots, M\} \to \{1, \dots, N\}$ stipulating which landmarks were observed at each time t is known or provided by an external algorithm.

Giving all these data and assumptions, we want to estimate the agent's position $T_t \in SE(3)$ with respect to the world frame over time.

B. Landmarks Mapping Problem

The data we have here is the same as we described above and the data association assumption also hold. We have other assumptions are made: the agent's pose $T_t \in SE(3)$ over time is known; the landmarks are static.

Giving all these data and assumptions, we want to estimate the world-frame coordinates of the landmarks $m_i \in \mathbb{R}^4, i = 1, 2, \dots, M$.

C. Visual-Inertial SLAM Problem

Provided the data as we described above and the data association assumption still hold, we want to simultaneously localize the agent by estimating the agent's pose $T_t \in SE(3)$ over time and mapping the landmarks $m_i \in \mathbb{R}^4$ to the world-coordinates frame.

III. Technical Approach

A. Extended Kalman Filter

In general, a nonlinear Kalman filter is a Bayes filter that has these characteristics: the prior pdf $p_{0|-1}$ is Gaussian; the motion model is nonlinear in the state and affected by Gaussian noise; the observation model is nonlinear in the state and affected by Gaussian noise; the process noise w_t and measurement noise v_t are independent of each other, of the state x_t and across time; the posterior pdf is forced to be

Gaussian via approximation. Given these, a nonlinear Kalman filter can be represent as the prior

$$x_t \mid z_{0:t}, u_{0:t-1} \sim N(\mu_{t|t}, \Sigma_{t|t}) \quad (1)$$

with the motion model

$$x_{t+1} = f(x_t, u_t, w_t) \quad (2)$$

and the observation model

$$z_t = h(x_t, v_t) \quad (3)$$

where z_t is the observation at time t, u_t is the IMU measurements at time t, $N(\mu_{t|t}, \Sigma_{t|t})$ indicates the Gaussian distribution with mean $\mu_{t|t}$ and covariance $\Sigma_{t|t}$.

The extended Kalman filter is a specific Kalman filter that uses a first-order Taylor series approximation to the motion and observation models. Therefore, the motion model can be written as:

$$f(x_t, u_t, w_t) \approx f(\mu_{t|t}, u_t, 0) + F_t(x_t - \mu_{t|t}) + Q_t w_t \quad (4)$$

where $F_t := \frac{df}{dx}(\mu_{t|t}, u_t, 0)$ and $Q_t := \frac{df}{dw}(\mu_{t|t}, u_t, 0)$. And the observation model can be written as:

$$h(x_{t+1}, v_{t+1}) \approx h(\mu_{t+1|t}, 0) + H_{t+1}(x_{t+1} - \mu_{t+1|t}) + R_{t+1} v_{t+1} \quad (5)$$

where $H_{t+1} := \frac{dh}{dx}(\mu_{t+1|t}, 0)$ and $R_{t+1} := \frac{dh}{dv}(\mu_{t+1|t}, 0)$. And then, we can represent the mean and covariance of the prediction step as:

$$\mu_{t+1|t} = f(\mu_{t|t}, u_t, 0) \quad (6)$$

$$\Sigma_{t+1|t} = F_t \Sigma_{t|t} F_t^T + Q_t W Q_t^T \quad (7)$$

where W is the covariance of the process noise w_t. For the update step, we can update the mean and covariance as:

$$\mu_{t+1|t+1} = \mu_{t+1|t} + K_{t+1|t}\left(z_{t+1} - h(\mu_{t+1|t}, 0)\right) \quad (8)$$

$$\Sigma_{t+1|t+1} = (I - K_{t+1|t} H_{t+1}) \Sigma_{t+1|t} \quad (9)$$

In this equation, K is the Kalman gain, and can be represented as:

$$K_{t+1|t} := \Sigma_{t+1|t} H_{t+1}^T (H_{t+1} \Sigma_{t+1|t} H_{t+1}^T + R_{t+1} V R_{t+1}^T)^{-1} (10)$$

where V is the covariance of the measurement noise v_t.

Given the homogeneous coordinate of a landmark m. The observation model with the measurement noise $v_t \sim N(0, V)$ is defined as

$$z_t = M\pi(T_i T_t m) + v_t \quad (11)$$

Here, z_t is the observation a time t, M is the stereo camera calibration matrix, T_i is the transformation matrix from IMU frame to optical frame, and π is the operator defined as

$$\pi(q) := \frac{1}{q_3} q \quad (12)$$

Given the IMU pose T_t. The motion model with time discretization τ and noise $w_t \sim N(0, W)$ is represented as

$$T_{t+1} = \exp\left(\tau(-\widehat{u_t + w_t})\right) T_t \quad (13)$$

Here, T_{t+1} is the IMU pose at time t+1, the hat map is defined in Lie Algebra of $SE(3)$ and the IMU measurements u_t is defined as

$$u_t := \begin{bmatrix} v_t \\ \omega_t \end{bmatrix} \quad (14)$$

v_t is the linear velocity and ω_t is the angular velocity that obtain from the IMU.

Given the prediction and update rules of EKF in (6)(7)(8)(9) and the specified motion model and observation model in (11), (13). We are interested how to implement the visual-inertial SLAM by associating the motion and observation model with the EFK.

Our data set consists synchronized measurements from an IMU and a stereo camera as well as the intrinsic camera calibration and the extrinsic calibration between the two sensors, and the transformation from the IMU to the left camera frame. Giving all this data, we want to implement the visual-inertial SLAM to estimate the agent's position, which is equivalent to obtain the pose of the IMU $T_t \in SE(3)$ with respect to the world frame over time, and the world-frame coordinates of the landmarks $m_i \in \mathbb{R}^4$.

We can divide this SLAM problem into 3 parts. First, we perform the dead reckoning to estimate the IMU pose over time. Second, we can use the estimated pose to update and map the observed landmarks on the world-coordinates frame. Finally, we combine the IMU prediction step and IMU update step with landmarks update step to form a complete visual-inertial SLAM algorithm and obtain the estimate IMU pose.

B. IMU Localization via EKF Prediction

The EKF prediction can be considered as a localization-only problem. We want to predict the agent's position in the world frame over time. In our case, this is the same as estimating the inverse IMU pose $T_t \in SE(3)$ over time, given the IMU measurements $\{u_t\}_{t=0}^T$ and the visual feature observations $\{z_t\}_{t=0}^T$.

In EKF implementation, the IMU pose has the prior $T_t \mid z_{0:t}, u_{0:t-1} \sim N(\mu_{t|t}, \Sigma_{t|t})$ where $\mu_{t|t} \in SE(3)$ and $\Sigma_{t|t} \in \mathbb{R}^{6*6}$. The motion model with time discretization τ and noise $w_t \sim N(0, W)$ is specified in (13).

With this motion model, we can re-write the updated mean and covariance, specified in (6) and (7), in terms of nominal kinematics of the mean of T_t and zero-mean perturbation kinematics:

$$\mu_{t+1|t} = \exp(-\tau \widehat{u_t}) \mu_{t|t} \quad (15)$$

In here, $\widehat{u_t} \in se(3)$ is the hat map defined in the Lie Algebra:

$$\widehat{u_t} = \begin{bmatrix} \widehat{\omega_t} & v_t \\ 0 & 0 \end{bmatrix} \quad (16)$$

and the covariance is

$$\Sigma_{t+1|t} = \exp(-\tau(u_t)^\curlywedge) \Sigma_{t|t} \exp(-\tau(u_t)^\curlywedge)^T + \tau^2 W \quad (17)$$

In here, $(u_t)^\curlywedge \in \mathbb{R}^{6*6}$ and it is defined as

$$(u_t)^\curlywedge = \begin{bmatrix} \widehat{\omega_t} & \widehat{v_t} \\ 0 & \widehat{\omega_t} \end{bmatrix} \quad (18)$$

C. *Landmark Mapping via EKF Update*

Visual mapping is a mapping only problem. Assuming we know the agent's pose in the world frame, which is the inverse IMU pose in our case, we want to estimate the homogeneous coordinates $m \in \mathbb{R}^4$ in the world frame of the landmarks, given the visual feature observations $\{z_t\}_{t=0}^{T}$, which is the coordinates in the image plane of the stereo camera.

The homogeneous coordinate of a landmark has the prior $m \mid z_{0:t} \sim N(\mu_t, \sum_t)$ where $\mu_t \in \mathbb{R}^4$ and $\sum_t \in \mathbb{R}^{3*3}$. The observation model with the measurement noise $v_t \sim N(0, V)$ is specified in (11).

With this observation model, we can rewrite the equation for updating the mean and covariance for EKF update step, as specified in (8) and (9):

$$\mu_{t+1} = \mu_t + DK_t(z_t - \widehat{z_t}) \tag{19}$$

$$\sum_{t+1} = (I - K_t H_t)\sum_t \tag{20}$$

where z_t is the new observation at time t, $\widehat{z_t}$ is the predicted observation based on μ_t and it is computed using the observation model specified in (11). D is a dilation matrix, K_t is the Kalman gain and it is computed as

$$K_t = \sum_t H_t^T (H_t \sum_t H_t^T + I \otimes V)^{-1} \tag{21}$$

and H_t is the Jacobian of $\widehat{z_t}$ with respect to m evaluated at μ_t, which is derived as

$$H_t = M \frac{d\pi}{dq}(T_i T_t \mu_t) T_i T_t D \tag{22}$$

D. *IMU Update via EKF Update*

For this EKF update, the variable of interest is the inverse IMU pose $T_{t+1} \in SE(3)$ instead of the landmark positions $m \in \mathbb{R}^4$.

Again, the IMU pose has the prior similar to the prediction step $T_{t+1} \mid z_{0:t}, u_{0:t} \sim N(\mu_{t+1|t}, \sum_{t+1|t})$ where $\mu_{t+1|t} \in SE(3)$ and $\sum_{t+1|t} \in \mathbb{R}^{6*6}$. The observation model is the same as we used in the visual mapping step, which is defined in (11). The difference is that, at this time we need the observation model Jacobian $H_{t+1|t} \in \mathbb{R}^{4*6}$ with respect to the inverse IMU pose and evaluated at $\mu_{t+1|t}$. This Jacobian is derived as

$$H_{t+1|t} = M \frac{d\pi}{dq}(T_i \mu_{t+1|t} m) T_i (\mu_{t+1|t} m)^{\odot} \tag{23}$$

where m is the landmark positions in world frame and $\odot$ operator is defined as

$$\begin{bmatrix} s \\ \lambda \end{bmatrix}^{\odot} = \begin{bmatrix} \lambda I & -\hat{s} \\ 0 & 0 \end{bmatrix} \tag{24}$$

Giving these, we can derivate the equation of mean and covariance for EKF update step:

$$\mu_{t+1|t+1} = \exp((K_{t+1|t}(z_{t+1} - \widehat{z_{t+1}}))^{\wedge}) \mu_{t+1|t} \tag{25}$$

$$\sum_{t+1|t+1} = (I - K_{t+1|t} H_{t+1|t}) \sum_{t+1|t} \tag{26}$$

In (25), the hat map is defined in (16), z_{t+1} is the new observation, $\widehat{z_{t+1}}$ is the predicted observation based on $\mu_{t+1|t}$ and it is computed using the observation model specified in (11). And the Kalman gain $K_{t+1|t}$ is derived as

$$K_{t+1|t} = \sum_{t+1|t} H_{t+1|t}^T (H_{t+1|t} \sum_{t+1|t} H_{t+1|t}^T + I \otimes V)^{-1} \tag{27}$$

E. *Pipeline of visual-inertial SLAM*

We first want to have a very rough estimation of the IMU pose by performing the dead reckoning using only the IMU measurements, which is the linear velocity v_t and the angular velocity ω_t.

Given the measurements at time t, we can write them as a twist $u_t \in \mathbb{R}^6$, which is described in (14). Then we predict the new pose with the new mean using the equation (15) and the new covariance using the equation (17) and the Gaussian noise $w_t \sim N(0, W)$.

After running the process over the given time period, we should have the IMU pose for each time in the episode. To obtain the trajectory of the agent, we just need to invert the IMU pose as the IMU pose is the same as the transformation from world frame to IMU frame.

Then, we want to map the observations of the landmarks to the world-coordinates frame and update the landmarks position. For example, given an observation z_t, we can map it to world coordinates frame by inversing the observation model, and update the mean of the landmark position using (19) and the covariance of the landmark position using (20). After running the process through all landmarks over the given time period, we should have a set of updated landmarks' position in the world-coordinates frame.

Finally, we want to combine the IMU prediction step and IMU update step with landmarks update step to obtain a more accurate estimation of the IMU pose over time. Given the measurements at time t, we first predict the IMU pose as what we did in EKF predict. Then we obtain a new observation of a landmark z_t. We map it to the world-coordinates frame and update the landmark position as what we did in visual mapping. And also, we update the mean and covariance of the IMU pose using equations (25) and (26) for the updated landmark position of the new observation.

After running the process over the timestamps of the given dataset, we should obtain a more accurate IMU pose over time and the trajectory of the agent.

IV. Result

From my implementation of the visual-inertial SLAM algorithm, the estimation results can be affected by the additive Gaussian noise v_t and w_t. From the experiment, larger covariance of observation noise produces the better results, which makes sense because the observation is noisy. If we less believe the observations, the trajectory is less affected by the observations.

We have 3 different datasets. Here, I will present the results of dead reckoning, landmarks mapping and visual-inertial SLAM.

A. *Dataset 0042*

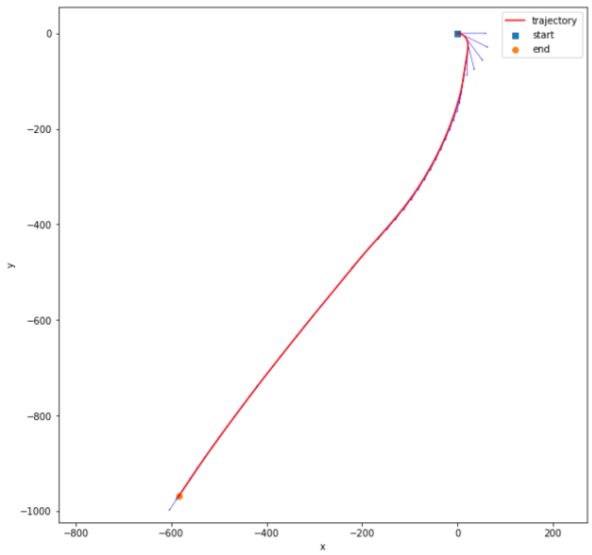

Figure 1

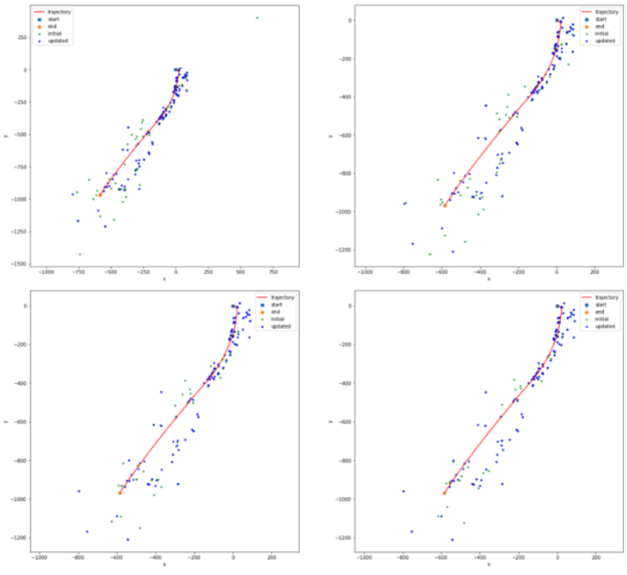

Figure 2

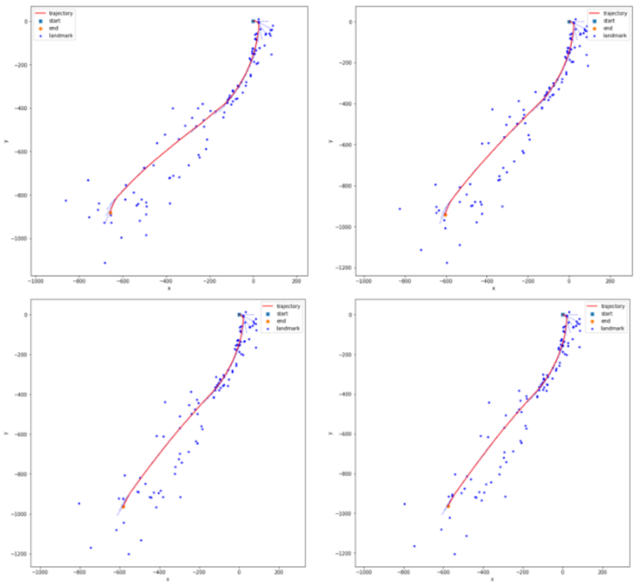

Figure 3

Figure 1 is the result of only running the EKF prediction. Figure 2 is the results of only running the EKF update for landmarks mapping. Green dots represent the initial landmarks positions and blue dots represent the updated landmarks positions. Figure 3 is the results of the visual-inertial SLAM algorithm. Blue dots represent the mapped landmarks position and the red line shows the agent's trajectory. For each plot on the Figure 2 and 3, different Gaussian noise $v_t \sim N(0, V)$ is added, where $V = sI \in \mathbb{R}^{4*4}$. For the top left one s = 0.01, for the top right one s = 0.1, for the bottom left one s = 1, for the bottom right one s = 10.

B. Dataset 0027

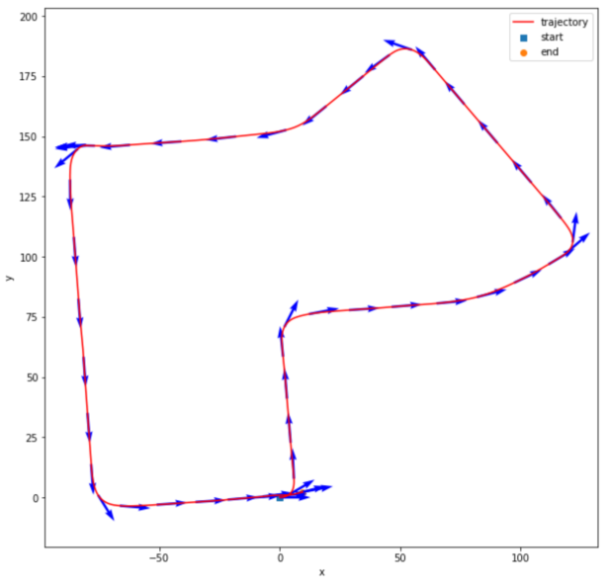

Figure 4

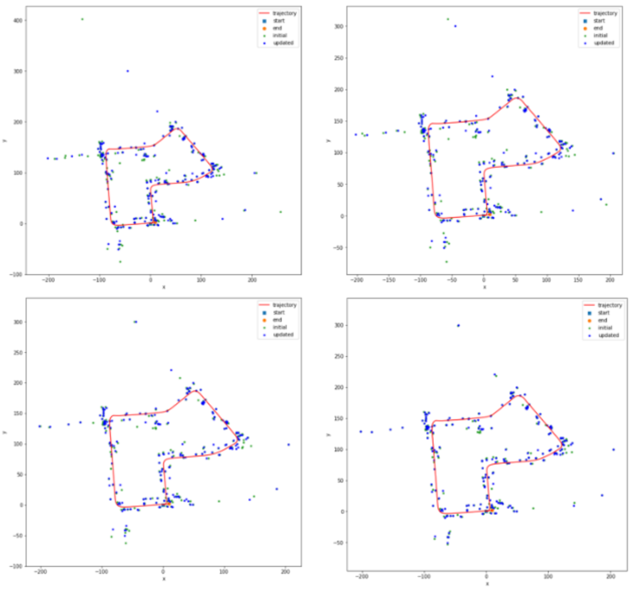

Figure 5

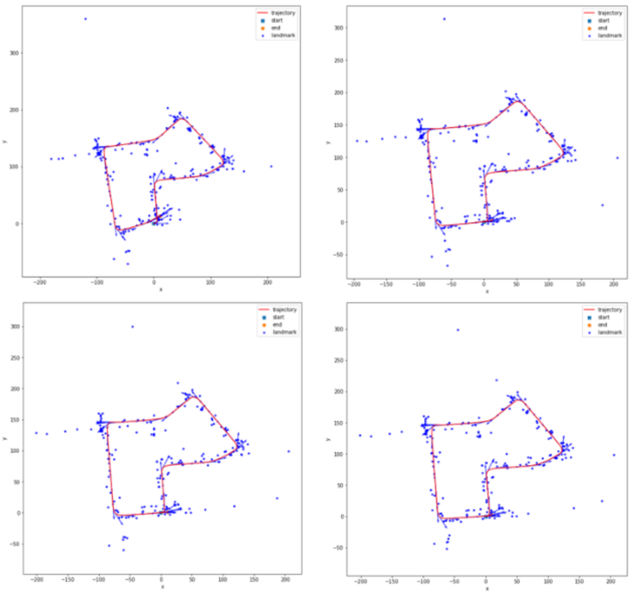

Figure 6

Figure 4 is the result of only running the EKF prediction. Figure 5 is the results of only running the EKF update for landmarks mapping. Green dots represent the initial landmarks positions and blue dots represent the updated landmarks positions. Figure 6 is the results of the visual-inertial SLAM algorithm. Blue dots represent the mapped landmarks position and the red line shows the agent's trajectory. For each plot on the Figure 5 and 6, different Gaussian noise $v_t \sim N(0, V)$ is added, where $V = sI \in \mathbb{R}^{4*4}$. For the top left one s = 0.01, for the top right one s = 0.1, for the bottom left one s = 1, for the bottom right one s = 10.

C. *Dataset 0020*

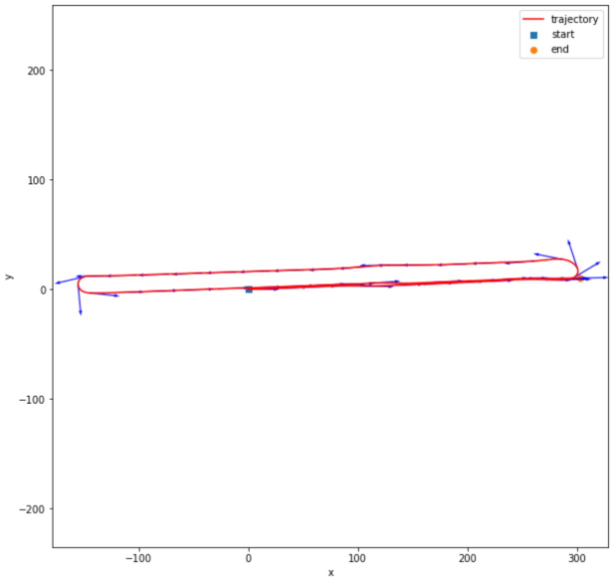

Figure 7

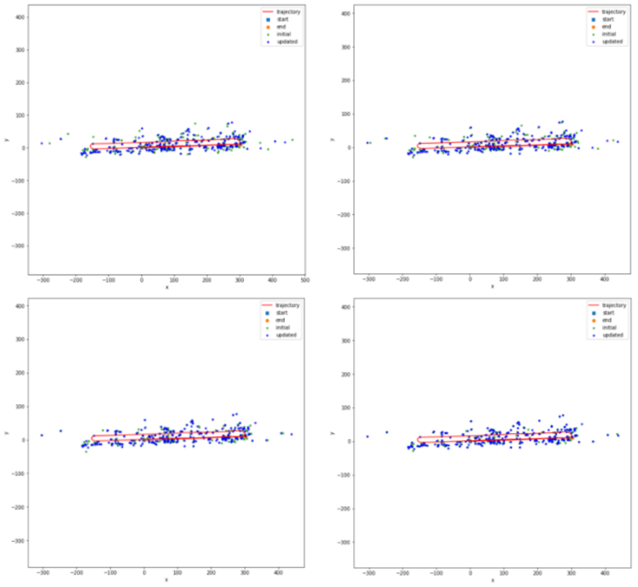

Figure 8

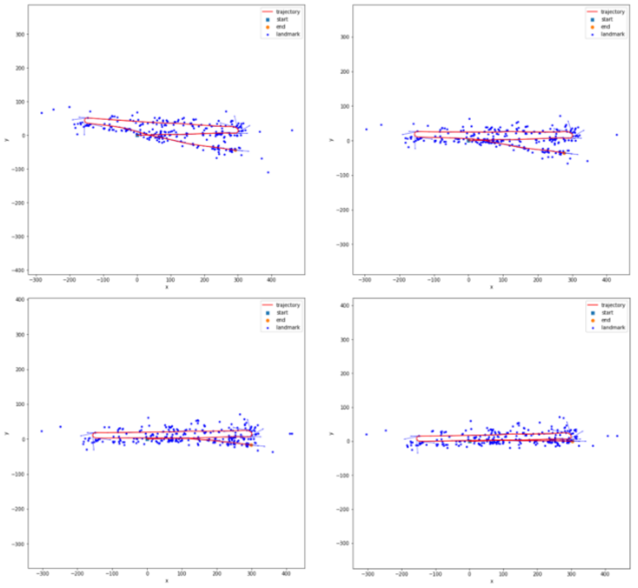

Figure 9

Figure 7 is the result of only running the EKF prediction. Figure 8 is the results of only running the EKF update for landmarks mapping. Green dots represent the initial landmarks positions and blue dots represent the updated landmarks positions. Figure 9 is the results of the visual-inertial SLAM algorithm. Blue dots represent the mapped landmarks position and the red line shows the agent's trajectory. For each plot on the Figure 8 and 9, different Gaussian noise $v_t \sim N(0, V)$ is added, where $V = sI \in \mathbb{R}^{4*4}$. For the top left one s = 0.01, for the top right one s = 0.1, for the bottom left one s = 1, for the bottom right one s = 10.

Conclusion

In this experiment, we are trying to localize the agent and mapping the landmarks in the world-coordinates frame using the visual-inertial SLAM algorithm. We solve the predict only problem using EFK prediction, and we solve the mapping only problem using EKF update. Finally, we solve the visual-inertial SLAM problem using both EFK prediction step and update step. From the experiment, we can see that only using the IMU measurements can give us a pretty good estimation, and the performance of the SLAM is pretty good but can be affected by the additive noise. The EKF is a first order linearization of the Kalman filter. It linearizes about an estimate of the current mean and covariance, and it works well in the case of well-defined transition models. However, in practice, most systems are nonlinear. Therefore, more complicated filter might be considered when dealing with the nonlinear systems.

References

[1] *R.E. Kalman (1960). "Contributions to the theory of optimal control". Bol. Soc. Mat. Mexicana: 102–119.* CiteSeerX 10.1.1.26.4070.

[2] *R.E. Kalman (1960).* "A New Approach to Linear Filtering and Prediction Problems" (PDF). *Journal of Basic Engineering. 82: 35–45.* doi*:*10.1115/1.3662552.

[3] *R.E. Kalman; R.S. Bucy (1961).* "New results in linear filtering and prediction theory" (PDF). *Journal of Basic Engineering. 83: 95–108.* doi*:*10.1115/1.3658902.