Visual Inertial SLAM

Lulua Rakla
PID A59012895
lrakla@ucsd.edu

***Abstract*—This report presents approaches to Simultaneous Localization and Mapping of an automobile trajectory using odometry and stereo camera data using Extended Kalman Filter. The output trajectory and landmark features are plotted in a 2D map.**

***Index Terms*—SLAM, computer vision, autonomous driving, Kalman Filter**

I. Introduction

Simultaneous Localization and Mapping (SLAM) is an approach to help a robot or autonomous agent navigate in an unknown environment. Using readings from various sensors, the agent is able to simultaneously generate the map of the environment and keep a track of its position in it. It seems like a chicken-egg problem as agent needs a map of the environment to know where it is, and it needs to know where it is to figure out the map. There are many approaches which vary depending on the motion and observation model distributions selected (Rao-Blackwellized Particle Filter, Extended Kalman Filter, Unscented Kalman Filter, Factor Graphs etc) This paper presents an approach to localization and mapping using an Extended Kalman Filter (EKF) which is a non-linear Bayes Filter. EKF uses first order Taylor series approximation to the motion and observation models around the state and noise means. It forces the predicted and updated distributions to be Gaussian by evaluating their first and second moments and approximating them with Gaussian distributions of the same moments.

EKF along with Unscented Kalman Filters have applications in non-linear systems of guidance, navigation and control, computer vision, self driving, mobile robotics, unmanned aerial vehicles etc. In this report, there is EKF based Visual Inertial SLAM.

II. Problem Formulation

Visual Inertial SLAM [2] is a a SLAM technique which uses IMU and camera data to map the visual features (landmarks) using the robot pose.

A. Dataset

There are two datasets used in this report - 03.npz and 10.npz The readings of the following sensors and their parameters are used
- IMU : The linear v_t and angular velocities $\omega_t \in R^3$ are provided for the vehicle in body frame
- Timestamp - Time stamps τ_t in UNIX standard secomds are provided.
- Intrinsic Calibration K_s and baseline b is given - - Extrinsic calibration - transformation from camera to IMU frame is given. - Features - Every pixel corresponding to M landmarks are given as $z_{i,t} \in R^4$

B. IMU Localisation

1) IMU Predict: The problem here is that given IMU measurements $u_{0:t}$ with u_t as $[v_t^T, \omega_t^T] \in R^6$ and feature observations $z_{0:T}$, estimate the pose $T_t :=_W T_I \in SE(3)$ over time. This prediction is simplified by using only kinematic equations instead of dynamic equations. An assumption is that the world frame coordinates of landmarks m are known. Another assumption is that the data association $\Delta_t := (1, ..., M) \rightarrow (1, ..., Nt)$ which stipulates that landmark j corresponds to observation $z_{t,i}$ with $i = \Delta_t(j)$ at time t is known or provided external algorithm. Using discrete time rotation kinematics, position at time $k+1$ is

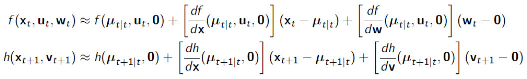
$$T_{k+1} = T_k exp(\tau_k \hat{\zeta}_k) \tag{1}$$

where $\tau_k = t_{k+1} - t_k$ and $\zeta(t) := [v(t)\omega(t)]^T \in R^6$ and $\hat{\zeta} := \begin{bmatrix} \hat{\omega} & v \\ 0 & 0 \end{bmatrix} \in R^{4X4}$ (twist matrix). The exponential map is a mapping from space of twist matrices $se(3)$ to space of pose matrices $SE(3)$ A prior Gaussian pose $T_t \sim N(\mu_{t/t}, \Sigma_{t/t})$ is assumed where $\mu \in SE(3)$ and $\Sigma \in R^{6X6}$ as there are 6 degrees of freedom. The problem is to estimate the mean and covariances which are done using the predict equations for IMU pose given in Fig 2. A Gaussian motion noise $\omega_t \sim N(0, W)$ is added to the covariance.

2) IMU Update: Here, assuming the position of landmarks are known, update IMU pose mean and covariance. The update equations are as given in Fig. 3. The difference is that the Jacobian $H_t \in R^{4NtX6}$ is wrt IMU pose T_{t+1} evaluated at $\mu_{t+1|t}$

C. Landmark (Visual) Mapping

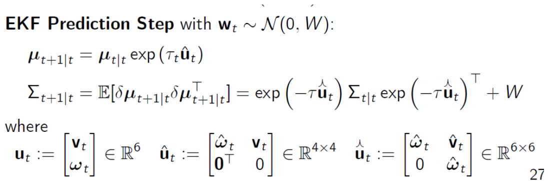
In this step, the objective is given the observations $z_t := [z_{t,1}^T., ..., z_{t,Nt}^T] \in R^{4Nt}$ for $t = 0, ...T$, estimate the 3D landmark world coordinates $m := [m_1^T, ..., m_M^T]^T \in R^{3M}$ of the landmarks that generated the observations. Each $z_{t,i}$ contains pixel values $[u_l, v_l, u_r, v_r]^T$ from the stereo cameras. As the landmarks are static, no motion model is considered in this step. An assumption is that the data association $\Delta_t := (1, ..., M) \rightarrow (1, ..., Nt)$ which stipulates that landmark j corresponds to observation $z_{t,i}$ with $i = \Delta_t(j)$ at time t is known or provided external algorithm. If the i^{th} observation is not observed the pizel coordinates are $[-1, -1, -1, -1]$. Another assumption while formulating this problem is that



$$f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t) \approx f(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, \mathbf{0}) + \left[\frac{df}{d\mathbf{x}}(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, \mathbf{0})\right](\mathbf{x}_t - \boldsymbol{\mu}_{t|t}) + \left[\frac{df}{d\mathbf{w}}(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, \mathbf{0})\right](\mathbf{w}_t - \mathbf{0})$$

$$h(\mathbf{x}_{t+1}, \mathbf{v}_{t+1}) \approx h(\boldsymbol{\mu}_{t+1|t}, \mathbf{0}) + \left[\frac{dh}{d\mathbf{x}}(\boldsymbol{\mu}_{t+1|t}, \mathbf{0})\right](\mathbf{x}_{t+1} - \boldsymbol{\mu}_{t+1|t}) + \left[\frac{dh}{d\mathbf{v}}(\boldsymbol{\mu}_{t+1|t}, \mathbf{0})\right](\mathbf{v}_{t+1} - \mathbf{0})$$

Fig. 1. EKF uses Taylor Series approximation



EKF Prediction Step with $\mathbf{w}_t \sim \mathcal{N}(0, W)$:

$$\boldsymbol{\mu}_{t+1|t} = \boldsymbol{\mu}_{t|t} \exp(\tau_t \hat{\mathbf{u}}_t)$$

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

where

$$\mathbf{u}_t := \begin{bmatrix} \mathbf{v}_t \\ \boldsymbol{\omega}_t \end{bmatrix} \in \mathbb{R}^6 \quad \hat{\mathbf{u}}_t := \begin{bmatrix} \hat{\boldsymbol{\omega}}_t & \mathbf{v}_t \\ \mathbf{0}^\top & 0 \end{bmatrix} \in \mathbb{R}^{4\times4} \quad \overset{\curlywedge}{\mathbf{u}}_t := \begin{bmatrix} \hat{\boldsymbol{\omega}}_t & \hat{\mathbf{v}}_t \\ 0 & \hat{\boldsymbol{\omega}}_t \end{bmatrix} \in \mathbb{R}^{6\times6}$$

27

Fig. 2. IMU Predict mean and covariance (from slides)

IMU position T_t is known and the stereo camera calibration matrix M is known. As a limitation of the sensor, it does not move sufficiently along z-axis, there the estimation of z coordinates are not good. So they are disregarded and only xy coordinates are plotted.

1) Initialize Landmarks: For a given IMU pose T_t and an observation $z_{t,i}$, if landmark is observed for the first time the mean of the landmark is initialized by converting it to world coordinates m using the equation

$$\begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} =_{cam} T_{imu} \cdot_{imu} T_{world} \cdot m \qquad (2)$$

where x_0, y_0, z_0 are coordinates in optical frame.

2) Update Landmarks: If the landmark is encountered for the second time, the landmark mean and covariance is updated using EKF equations given in Fig. 3. K is the Kalman Gain and H is the observation model Jacobian with respect to world frame m_j evaluated at $\mu_{t,j}$ A prior for landmark world positions $\mu_t \in R^{3M}$ and $\Sigma_t \in R^{3Mx3M}$ is initialized.

D. Visual Inertial SLAM

In this step the problem is to combine the update of the m landmarks and IMU pose simultaneously. This is done by using the predicted IMU pose (the problem is same as IMU localisation) from localisation to update the landmarks and IMU parameters $(\mu_{t+1|t}, \Sigma_{t+1|t})$ together. This is done by equations given in Fig. 4. The covariance Σ is defined as

$$\Sigma = \begin{bmatrix} \Sigma_{landmark} & C \\ C & \Sigma_{imu} \end{bmatrix} \in R^{3M+6X3M+6} \qquad (3)$$

Here C is the covariance between the landmarks and IMU pose. This means that IMU pose and landmarks position is correlated. To get this joint covariance, a single Jacobian H_t should be calculated.

$$H_{t+1|t} = [H_{landmarks,t+1|t} H_{imu,t+1|t}] \qquad (4)$$

III. Technical Approach

This section outlines the technical details of the project

A. Initializing IMU and landmarks

To begin the process, it was essential to figure out the intended dimensions for the matrices and initialize them correctly. The table illustrated the dimensions and initializations of the matrices

TABLE I
Initializations and Matrix Dimensions

Matrix	Value	Dimension
Σ	Joint Covariance (IMU and landmark cov)	$3M + 6X3M + 6$
$H_{t+1\|t}$	Joint Jacobian (zeros)	$4NtX3M + 6$
$\mu_{landmarks}$	Landmark mean (zeros)	$3MX1$
μ_{IMU}	Landmark mean (zeros)	$4X4$

B. IMU Localisation via EKF Prediction

To localise the IMU, predict equations from Fig 2 were used. The pose T_t of the IMU was updated using these equations. It was essential to calculate the twist matrix $u^{\curlywedge}$ using angular and linear velocity $\hat{\omega_t}$ and $\hat{v_t}$ respectively. The logic behind this comes from the discrete time motion models which are split into nominal and perturbation kinematics.

$$\mu_{t+1|1} = \mu_{t|t} exp(\tau_t \hat{u_t}) \qquad (5)$$

$$\delta\mu_{t+1|1} = exp(-\tau_t u^{\curlywedge})\delta u_t + w_t \qquad (6)$$

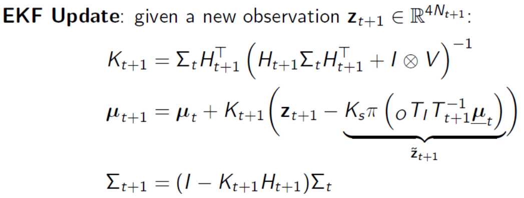
The motion covariance was computed as a multivariate Gaussian random variable with mean 0 and covariance as the standard deviation σ of angular velocity and linear velocity.

1) Dead Reckoning: After appling only prediction equations, dead reckoning trajectories were plotted.

C. Landmark mapping via EKF Update

When a landmark is first observed, it is initialized. To initialize the 3D landmarks in world frame, the stereo camera projections equations were used to map the pixels to world frame.

- Disparity was computed as $d = u_l - u_r$
- z_o is calculated as $\frac{fsub}{d}$
- x_0 and y_o was calculated using camera parameters c_u, c_v
- landmark prior mean u_j was initialized as $[m_x \; m_y \; m_z]$ and landmark Σ_j = $I_{3\times3}$



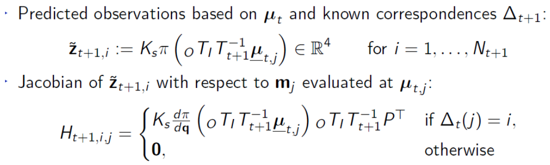
EKF Update: given a new observation $\mathbf{z}_{t+1} \in \mathbb{R}^{4N_{t+1}}$:

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

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + K_{t+1}\Big(\mathbf{z}_{t+1} - \underbrace{K_s\pi\left({}_O T_I T_{t+1}^{-1}\underline{\boldsymbol{\mu}}_t\right)}_{\tilde{\mathbf{z}}_{t+1}}\Big)$$

$$\Sigma_{t+1} = (I - K_{t+1}H_{t+1})\Sigma_t$$

Fig. 3. Landmarks update equations



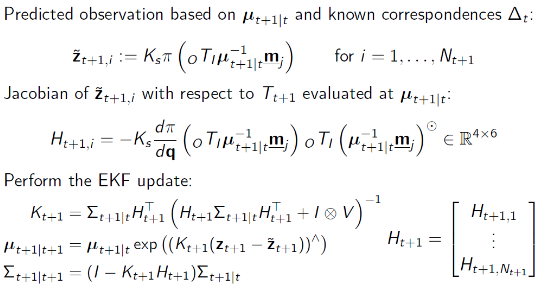
- Predicted observations based on $\boldsymbol{\mu}_t$ and known correspondences Δ_{t+1}:

$$\tilde{\mathbf{z}}_{t+1,i} := K_s\pi\left({}_O T_I T_{t+1}^{-1}\underline{\boldsymbol{\mu}}_{t,j}\right) \in \mathbb{R}^4 \qquad \text{for } i = 1, \ldots, N_{t+1}$$

- Jacobian of $\tilde{\mathbf{z}}_{t+1,i}$ with respect to $\mathbf{m}_j$ evaluated at $\boldsymbol{\mu}_{t,j}$:

$$H_{t+1,i,j} = \begin{cases} K_s\frac{d\pi}{d\mathbf{q}}\left({}_O T_I T_{t+1}^{-1}\underline{\boldsymbol{\mu}}_{t,j}\right){}_O T_I T_{t+1}^{-1}P^\top & \text{if } \Delta_t(j) = i, \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

Fig. 4. Computing z_{pred} and Observation Jacobian



Predicted observation based on $\boldsymbol{\mu}_{t+1|t}$ and known correspondences Δ_t:

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

Jacobian of $\tilde{\mathbf{z}}_{t+1,i}$ with respect to T_{t+1} evaluated at $\boldsymbol{\mu}_{t+1|t}$:

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

Perform the EKF update:

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

$$\boldsymbol{\mu}_{t+1|t+1} = \boldsymbol{\mu}_{t+1|t}\exp\left(\left(K_{t+1}(\mathbf{z}_{t+1} - \tilde{\mathbf{z}}_{t+1})\right)^{\wedge}\right)$$

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

$$H_{t+1} = \begin{bmatrix} H_{t+1,1} \\ \vdots \\ H_{t+1,N_{t+1}} \end{bmatrix}$$

Fig. 5. Simultaneously updating parameters of IMU and landmarks during update (from slides)

Calibration matrix M was used in the stereo camera equations $\begin{bmatrix} fsu & 0 & c_u & 0 \\ u & fsv & c_v & 0 \\ 0 & 0 & 0 & fsub \end{bmatrix}$ where f is focal length, su,sv are pixel scaling, cu cv are principle points and b is baseline. The observation model of landmarks with measurement noise $v_{t,i} \sim N(0, V)$ was defined as

$$z_{t,i} = h(T_t, m_j) + v_{t,i} := K_s\pi({}_oT_I T_t^{-1} m_j) + v_{t,i} \tag{7}$$

Here m_j = $[m_j \ 1]^T$ and $\pi(q)$ is the projection function whose derivative is $\frac{1}{q_3}\begin{bmatrix} 1 & 0 & \frac{-q_1}{q_3} & 0 \\ 0 & 1 & \frac{-q_2}{q_3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{-q_4}{q_3} & 1 \end{bmatrix}$

To map the landmarks, when a landmark was observed again,it's z_{pred} ($\tilde{z_{t+1,i}}$) was computed using current mean of the landmarks and ${}_{cam}T_{world}$. (Fig 4.)The innovation $z_{t,i} - z_{pred}$ is used further in computing mean in EKF Update. The landmark's mean and covariance was updated using landmark update equations in Fig. 3. The observation Jacobian $H_{t+1,i,j}$ was computed as shown in Fig 4. Here each $H_{t+1,i,j}$ is $\in R^{4\times 3}$ where i is the observation at t corresponding to the j landmark. In the $H_{t+1,i,j}$ has a term P^t which is just $[I \ 0]^T$ and K_s is the camera intrinsic matrix. As landmarks are static, no prediction is performed.

D. Visual Inertial SLAM

In this section, simultaneous location and mapping was performed by predicting the IMU pose T_t and then updating the landmarks and IMU together using Extended Kalman Filter equations. The approach was to update the covariance together to account for correlation between IMU pose and landmarks. The covariance Σ is defined as

$$\Sigma = \begin{bmatrix} \Sigma_{landmark} & C \\ C & \Sigma_{imu} \end{bmatrix} \in R^{3M+6X3M+6} \tag{8}$$

Here C is the covariance between the landmarks and IMU pose. It is important to note that this was assumed to be 0 in the separate prediction and mapping steps. Having C means that IMU pose and landmarks position is correlated. To get this joint covariance, a single Jacobian H_t should be calculated.

$$H_{t+1|t} = [H_{landmarks,t+1|t}H_{imu,t+1|t}] \tag{9}$$

The equations for simultaneous update are given in Fig 5. Observation noise V is added so the matrix in Kalman Gain is invertible. V is taken to be $100I$. For mean, the equations of update were used as defined in Fig 5 for IMU pose and as defined in Fig. 3 for landmarks. $\mu_{t+1|t}^{-1}$ is the inverse of IMU pose, hence it is ${}_{imu}T_{world}$. $\odot$ operator is defined as follows

$$\begin{bmatrix} s \\ 1 \end{bmatrix}^{\odot} := \begin{bmatrix} I & -\hat{s} \\ 0 & 0 \end{bmatrix} \in R^{4\times 6} \tag{10}$$

IV. Results

The algorithm was run in different scenarios on 2 datasets 03.npz and 10.npz which has odometry and stereo camera data collected by a driving car whose trajectory is to be visualized. Instead of the complete set of features a subset was used wherein every 5th, 10th and 20th feature was used to get the trajectory. It was observed that changing features impacted the final trajectory.

A. Dead Reckoning

The results of dead reckoning are shown in Fig 6. Here only IMU pose was computed using EKF predict equations. As the IMU sensor is responsible for trajectory, it's covariance was $\Sigma as 0.01I$

B. Visual Mapping

Here the updated features were placed alone with the IMU predicted trajectory. No update was done to IMU pose i.e it's mean and covariance. The results are shown in Fig. 7 where the mapping is done using 1/5 features and 1/20 features (for Dataset 10). It is observed that the landmarks are close to the trajectory but some are also quite spread out. (Trajectory is in red)

C. Visual Inertial SLAM

By only performing dead reckoning and mapping, the trajectory is not very accurate as the correlation (off diagonal terms) between IMU pose and landmarks is lost. This loss is prevented in Visual Inertial SLAM by keeping a joint covariance matrix and joint Jacobian matrix. Then the equations of Update are used using a block Kalman Gain and joint Jacobian. By changing motion and observation noise, the trajectories change as SLAM is dependent on noise in the observations. It has no ground truth, so noise affects the trajectories incrementally. Fig 9 and 10 have the trajectories (in red) obtained after performing SLAM using every 5th and every 20th feature. As seen in Dataset 10, changing features also changes the trajectories. The graphs also show that features are more concentrated than when it was with just performing mapping.

D. Noteworthy Observations

- To account for computation power, a subset of features were used.
- Varying the hyperparameters like Observation Noise V is essential in preventing the Kalman Gain invertible matrix. An Identity matrix was multiplied by 100 to ensure the term in Kalman Gain remains invertible.
- Reducing the number features from every 5th to 10th to 20th changes the trajectory.
- Using Joseph Form of the covariance update equation makes the Kalman Gain equation invertible $\Sigma_{t+1|t+1} = (I - K_{t+1|t}H_{t+1})\Sigma_{t+1|t}(I - K_{t+1|t}H_{t+1})^\top + K_{t+1|t}VK^\top_{t+1|t}$instead of $\Sigma_{t+1|t+1} = (I - K_{t+1|t}H_{t+1})\Sigma_{t+1|t}$
- Adding a slight offset to joint Covariance off diagonals terms made the trajectory more smooth.
- The features were closer to the trajectory in Visual Intertial SLAM rather than just mapping. This indicated that there was some correlation between IMU pose and visual landmarks.

References

[1] https://natanaso.github.io/ece276a/ref/ECE276A13VISLAM.pdf

[2] Chang, Chen Zhu, Hua Li, Menggang You, Shaoze. (2018). A Review of Visual-Inertial Simultaneous Localization and Mapping from Filtering-Based and Optimization-Based Perspectives. Robotics. 7. 45. 10.3390/robotics7030045.

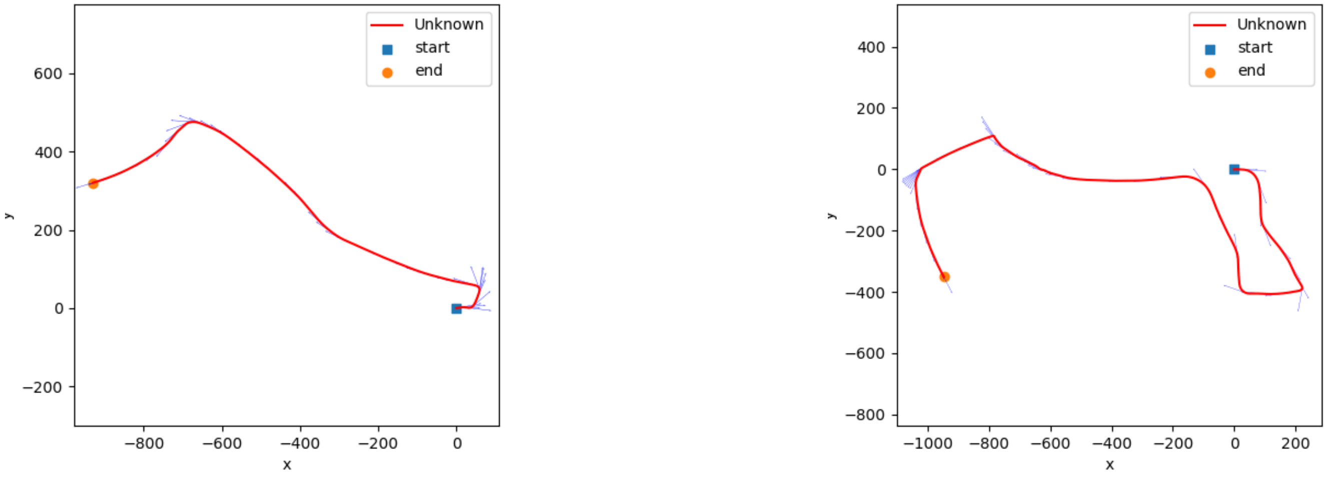


Fig. 6. Dead Reckoning result for 03 and 10 dataset

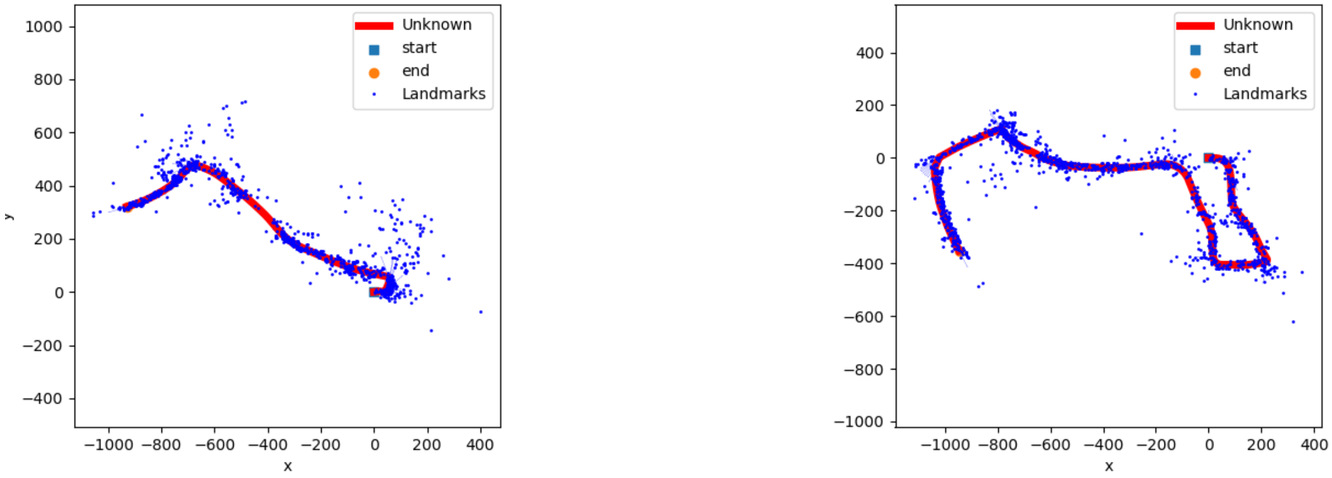


Fig. 7. Visual Mapping for Dataset 03 and Dataset 10 using every 5th landmark

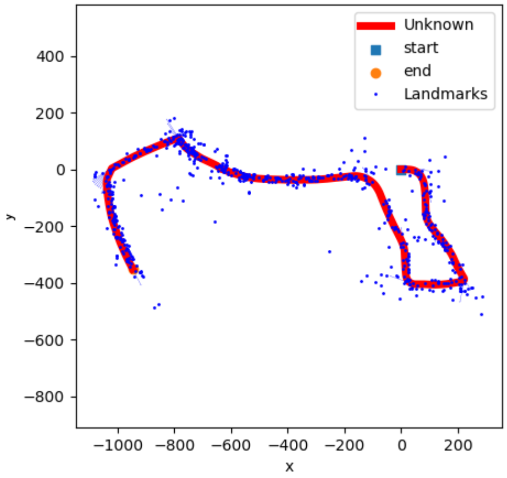


Fig. 8. Visual Mapping for Dataset 10 using every 20th landmark

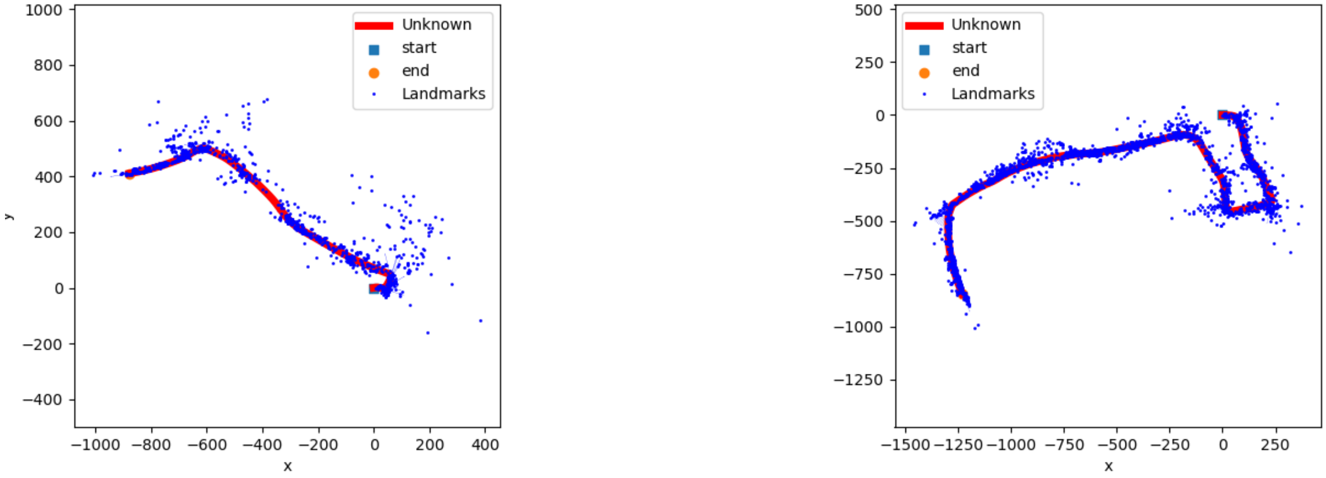


Fig. 9. VI Slam for Dataset 03 and Dataset 10 using every 5th landmark

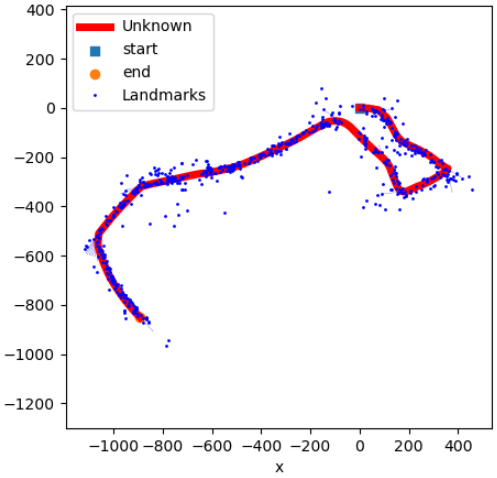


Fig. 10. VI Slam for Dataset 10 using every 20th landmark