Visual-Inertial SLAM

Saurabh Himmatlal Mirani (A53319557)
Sensing and Estimation in Robotics
University of California, San Deigo
smirani@ucsd.edu

I. Introduction

Recent advances in autonomous mobile robots has allowed robots to be involved in many applications including planetary exploration, search and rescue operations etc. For accomplishing a generic task, it is important for the robot to autonomously navigate, use the information acquired through various sensors and a model of the surrounding environment. This problem of estimating both the robot pose and the environment representation at the same time is usually defined as Simultaneous Localization And Mapping (SLAM). SLAM problem has become one of the most important research area of robotics, and the key to realize autonomous navigation. Localization needs the information of the environment map, while mapping relies on the robot's position and pose. Due to the accuracy and cost of sensors as well as the unknown environment, both localization and mapping problems are challenging, which call for the technique for localization and mapping at the same time [1].

SLAM addresses the problem of building a map of the environment using sensor data obtained from a mobile robot. The mobile robot is subject to error, hence the mapping problem creates a robot localization problem; hence the name SLAM. The ability to simultaneously localize a robot and accurately map its environment is considered by many to be a key prerequisite of truly autonomous robots [2].

In this paper, SLAM is implemented using

The rest of the paper is organized as follows, Section II describes Problem Formulation, Section III describes the Technical Approach, experimental results are reported in section IV, Section V concludes the paper.

Kalman filtering, also known as linear quadratic estimation (LQE), is an algorithm that uses a series of measurements observed over time, containing statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each time-frame. Using a Kalman filter assumes that the errors are Gaussian. In summary, the following assumptions are made about random processes: Physical random phenomena may be thought of as due to primary random sources exciting dynamic systems. The primary sources are assumed to be independent Gaussian random processes with zero mean; the dynamic systems will be linear. The random processes are therefore described by models

The three main steps in this paper are [3]:

1) IMU Localization using EKF Prediction
2) Landmark Mapping via EKF Update
3) Visual-Inertial SLAM

II. Problem Formulation

Given a series of controls u_t and sensor observations z_t over discrete time steps t, the SLAM problem is to compute an estimate of the robot's location x_t and a map of the environment m_t (here positions of the landmarks in world frame). All quantities are usually probabilistic, so the objective is to compute:

$$P(m_{t+1}, x_{t+1} | z_{1:t+1}, u_{1:t})$$

Applying Bayes' rule gives a framework for sequentially updating the location posteriors, given a map and a transition function $P(x_t|x_{t-1})$,

$$P(x_t|z_{1:t}, u_{1:t}, m_t) =$$
$$\sum_{m_{t-1}} P(z_t|x_t, m_t, u_{1:t}) \sum_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}|m_t, z_{1:t-1}, u_{1:t})/\eta$$

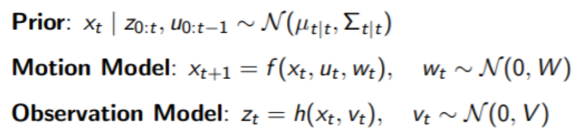
Similarly the map can be updated sequentially by

$$P(m_t|x_t, z_{1:t}, u_{1:t}) =$$
$$\sum_{x_t} \sum_{m_t} P(m_t|x_t, m_{t-1}, z_t, u_{1:t}) P(m_{t-1}, x_t|z_{1:t-1}, m_{t-1}, u_{1:t})$$

III. Technical Approach:

The above problem is solve using Extended Kalman Filter (EKF).
The EKF predict and update equations are stated below. The use of these equations to solve the problem are discussed later.

A. EKF Equations:



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

Motion Model: $x_{t+1} = f(x_t, u_t, w_t), \quad w_t \sim \mathcal{N}(0, W)$

Observation Model: $z_t = h(x_t, v_t), \quad v_t \sim \mathcal{N}(0, V)$

where,
x_t: state at time t
$z_{0:t}$: observations from 0 to t
$u_{0:t-1}$: observations from 0 to $t-1$

μ: mean
Σ: standard deviation

First-order Taylor series approximation to the motion and observation model is used in EKF:

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

- **EKF Prediction Step**
  The Jacobians are:

  $$F_t := \frac{df}{d\mathbf{x}}(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, \mathbf{0}) \text{ and } Q_t := \frac{df}{d\mathbf{w}}(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, \mathbf{0})$$

  Then the predicted mean and co-variances are:

  $$\mu_{t+1|t} = f(\mu_{t|t}, u_t, 0)$$
  $$\Sigma_{t+1|t} = F_t \Sigma_{t|t} F_t^T + Q_t W Q_t^T$$

  where W is the co-variance of noise.
- **EKF Update Step:** The Jacobians are:

  $$H_{t+1} := \frac{dh}{d\mathbf{x}}(\boldsymbol{\mu}_{t+1|t}, \mathbf{0}) \text{ and } R_{t+1} := \frac{dh}{d\mathbf{v}}(\boldsymbol{\mu}_{t+1|t}, \mathbf{0})$$

  Then the updated mean and co-variances are:

  $$\mu_{t+1|t+1} = \mu_{t+1|t} + K_{t+1|t}(z_{t+1} - h(\mu_{t+1|t}, 0))$$
  $$\Sigma_{t+1|t+1} = (I - K_{t+1|t} H_{t+1}) \Sigma_{t+1|t}$$

  where, the Kalman gain is,

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

For the given problem statement, with synchronized measurements from IMU and stereo camera, we adopt the EKF model by Visual-Inertial SLAM.

1) **IMU Localization using EKF Prediction:** Assumption: The homogenous coordinates of the landmarks in the world frame are known and are $m \in \mathbb{R}^{4\text{xM}}$. Given the IMU measurements u_t where $u_t = [v_t^T \ \omega_t^T]^T$ and the visual feature observations z_t estimate the inverse IMU pose $U_t =_w T_i^{-1}$, where $T \in$ SE(3) over time.
   The prior is,

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

   EKF Prediction Step:

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

   where

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

2) **Landmark Mapping via EKF Update:**
   Assumption: The IMU pose over time is known. Given the feature observations z_t we need to estimate the homogeneous coordinates m in the world frame of the landmarks that generated the visual observations.
   The prior is:

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

   The observation model with measurement noise $v_{t,i} \sim N(0, V)$:

   $$\mathbf{z}_{t,i} = h(U_t, \mathbf{m}_j) + \mathbf{v}_{t,i} := M\pi\left({}_O T_I U_t \underline{\mathbf{m}}_j\right) + \mathbf{v}_{t,i}$$

   The EKF Update step is thus,

   $$K_t = \Sigma_t H_t^\top \left(H_t \Sigma_t H_t^\top + I \otimes V\right)^{-1} \qquad I \otimes V := \begin{bmatrix} V & & \\ & \ddots & \\ & & V \end{bmatrix}$$
   $$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + K_t(\mathbf{z}_t - \tilde{\mathbf{z}}_t)$$
   $$\Sigma_{t+1} = (I - K_t H_t)\Sigma_t$$

   where H_t is the Jacobian of $\tilde{z}_{t,i}$ with respect to m_j evaluated at $\mu_{t,j}$

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

   where,

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

   M is the calibration matrix

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

   ${}_o T_i \in$ SE(3) is extrinsic

3) **Visual-Inertial SLAM:** The IMU pose prediction step from part (1) and the landmark update step from part (2) are combined with the IMU update step based on the stereo camera observation model to obtain a complete Visual-Inertial SLAM algorithm

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

   The predicted observation is

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

The Jacobian of $\tilde{z}_{t+1,i}$ with respect to U_{t+1} evaluated at $\mu_{t+1|t}$ is:

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

The EKF update step is thus,

$$\begin{aligned} K_{t+1|t} &= \Sigma_{t+1|t}H_{t+1|t}^{\top}\left(H_{t+1|t}\Sigma_{t+1|t}H_{t+1|t}^{\top} + I \otimes V\right)^{-1} \\ \boldsymbol{\mu}_{t+1|t+1} &= \exp\left(\left(K_{t+1|t}(\mathbf{z}_{t+1} - \tilde{\mathbf{z}}_{t+1})\right)^{\wedge}\right)\boldsymbol{\mu}_{t+1|t} \\ \Sigma_{t+1|t+1} &= (I - K_{t+1|t}H_{t+1|t})\Sigma_{t+1|t} \end{aligned} \qquad H_{t+1|t} = \begin{bmatrix} H_{1,t+1|t} \\ \vdots \\ H_{N_{t+1},t+1|t} \end{bmatrix}$$

B. Data:

1) **IMU:** The IMU provides the linear velocity and the angular velocity of the robot in 3D.

$$u_t = [v_t \ w_t]^T$$

where $v_t \in \mathbb{R}^3$ is the linear velocity and $w_t \in \mathbb{R}^3$ is the angular velocity. These are measured in the body frame of IMU.

2) **Stereo Camera:** Using a feature detection algorithm on a stereo camera model, features are obtained. The corresponding features are matched in both the cameras, left and right and the pixel coordinates of a feature are given by,

$$z = [u_l \ v_l \ u_r \ v_r]$$

where, l and r represent left and right cameras

3) **Time synchronisation:** Both the sensors, i.e. IMU and Stereo camera are time synchronised.

4) **Intrinsic Calibration:** The relationship between the optical frame and the pixel values is specified by the intrinsic calibration matrix given by,

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

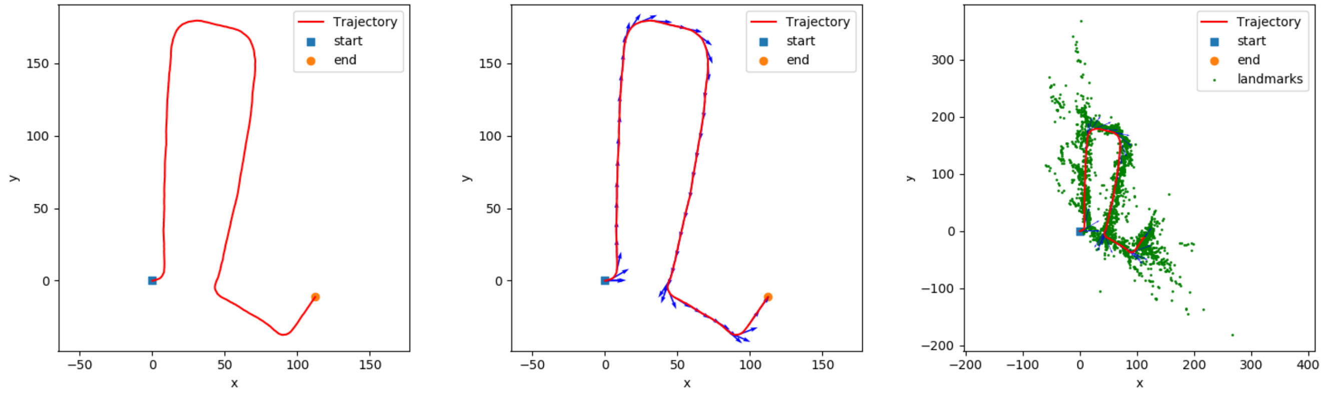
5) **Extrinsic calibration:** The transformation ${}_cT_i \in SE(3)$ from the IMU to left camera frame

C. Feature selection:

All the features can result in over-fitting. Hence, the algorithm should be generalized so that it avoids over-fitting, since noise in the data shouldn't affect our SLAM largely. In order to achieve this, best features are chosen in way such that maximum update is possible. Update is essential step for trajectory and landmark position correction. Hence, data is first sorted according to the number of times a feature is observed (in descending order, maximum times observed first, least observed last). Then 1st n features are chosen. In this paper, using $\frac{1}{7}$th of total features as best features yielded significantly better results. Thus, all features are not considered and only $\frac{1}{7}$th of total features are considered.

IV. Results and Discussion

A. Dead Reckoning

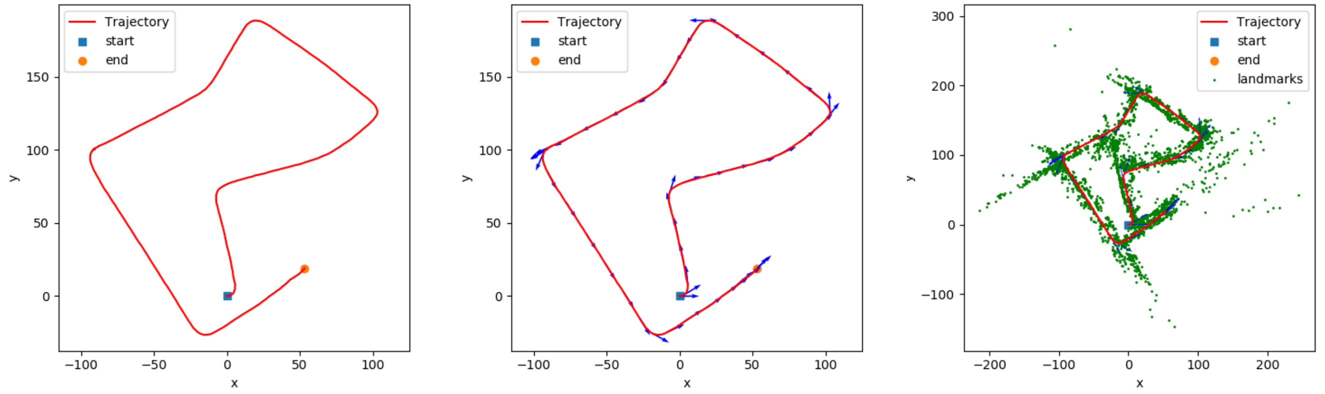


(a) Robot Trajectory

(b) Robot orientation

(c) Landmarks

Fig. 1. Dataset 0022

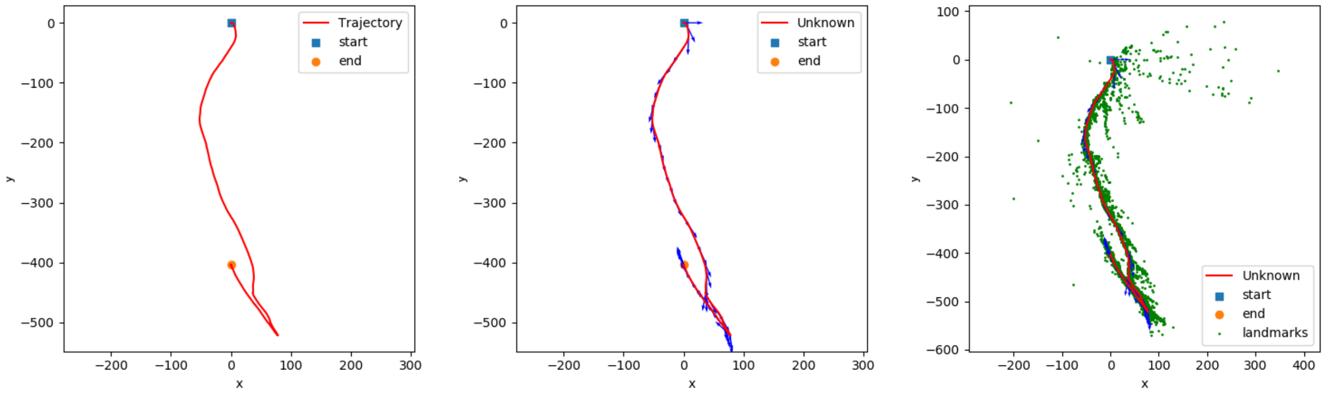


(a) Robot Trajectory

(b) Robot orientation

(c) Landmarks

Fig. 2. Dataset 0027



(a) Robot Trajectory

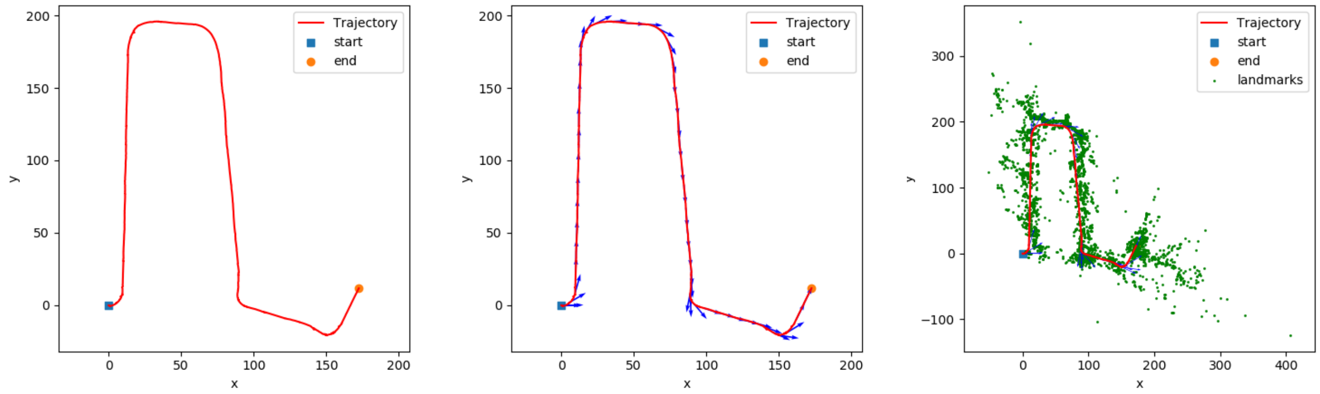
(b) Robot orientation

(c) Landmarks

Fig. 3. Dataset 0034

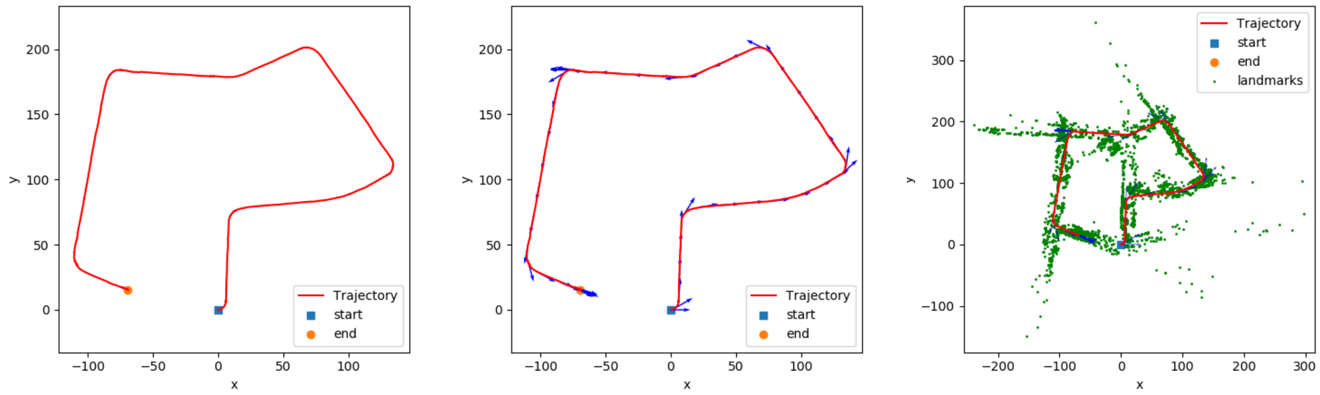
B. Visual-Inertial SLAM

Motion noise used was $W = diag(0.5, 0.5, 0.5, 0.05, 0.05, 0.05)$ 0.5m/s and 0.05 rad/s and observation noise used was $V = diag(15, 15, 15)$ 15 pixels



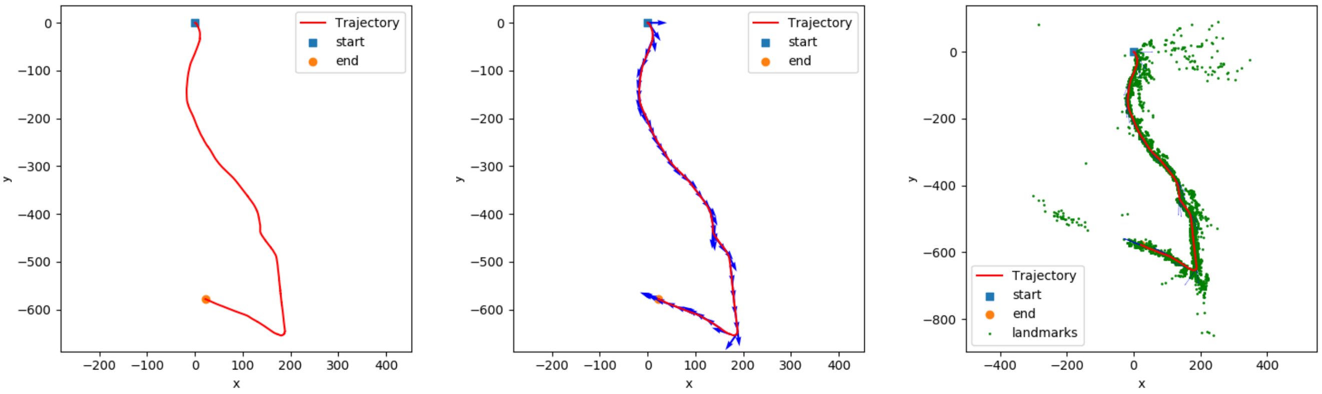
(a) Robot Trajectory (b) Robot orientation (c) Landmarks

Fig. 4. Dataset 0022: Using all features



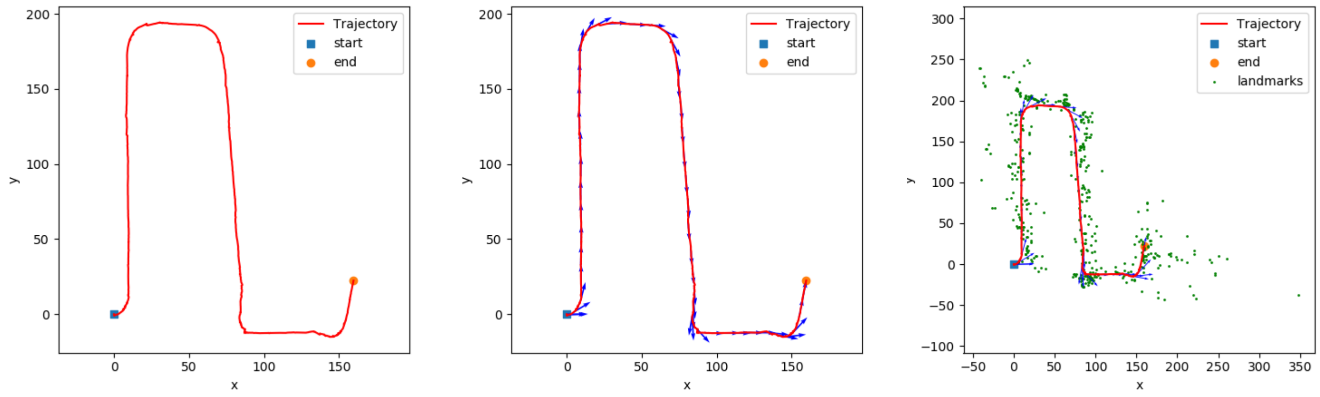
(a) Robot Trajectory (b) Robot orientation (c) Landmarks

Fig. 5. Dataset 0027: Using all features



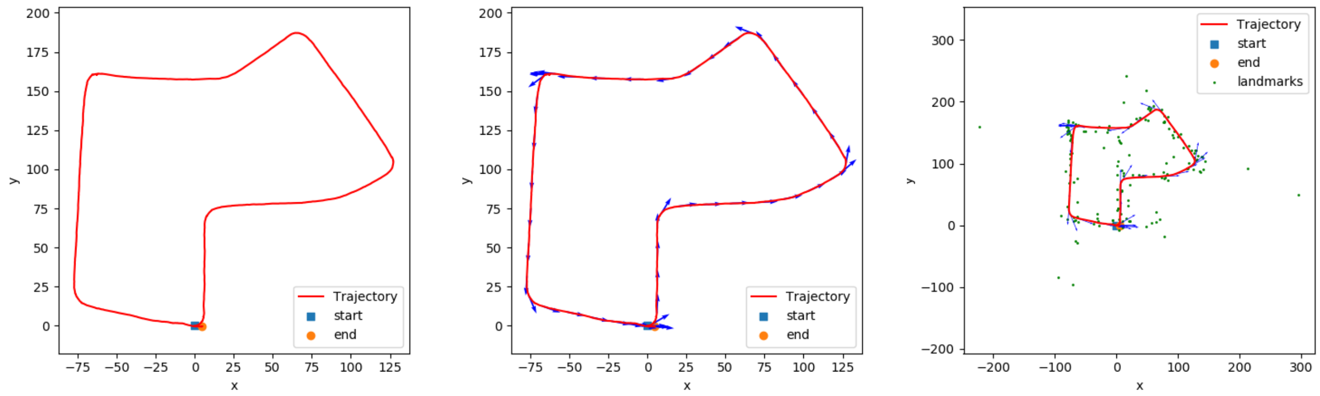
(a) Robot Trajectory (b) Robot orientation (c) Landmarks

Fig. 6. Dataset 0034: Using all features



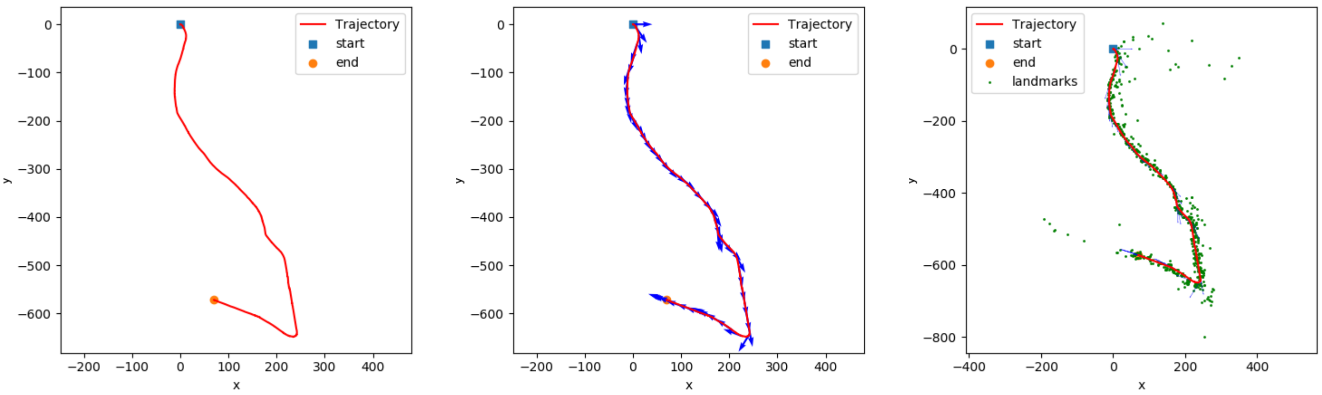
(a) Robot Trajectory (b) Robot orientation (c) Landmarks

Fig. 7. Dataset 0027: Using 460 best features



(a) Robot Trajectory (b) Robot orientation (c) Landmarks

Fig. 8. Dataset 0027: Using 565 best features



(a) Robot Trajectory (b) Robot orientation (c) Landmarks

Fig. 9. Dataset 0034: Using 688 best features

C. *Discussion*

As it can be seen, the results of dead reckoning are poor. Visual Inertial SLAM when used with all features, does improve the results significantly. However, due to over-fitting, it is essential to use only some (good) features out of all. Using best features yielded great results. Fig. 8 clearly shows a loop closure (almost). The end point is slight ahead of the start point which is the actual case as seen in the video, the car moves forward after reaching the start point. Even in Fig. 9, a small bulge can be seen around the middle of the trajectory, which corresponds to the actual case as seen in video. Hence, the results obtained are accurate.

V. Conclusion and Future Work

Visual-Inertial SLAM improves the results drastically. It helps in estimating what the actual pose of robot would have been by and hence determine accurate positions of landmarks in the environment. This thus solves the chicken egg problem of mapping and localization.

Using best features only by determining the number of times it was observed helps in correcting trajectory but then information is lost. There is a trade-off between information gain and trajectory correction. Information gain is not so important here but may be important in cluttered environment, for example indoor SLAM where we have many rooms. This can be considered for future work.

Acknowledgment

I thank Mr. Shiladitya Biswas, Mr. Sushruth Nagesh, Ms. Asfiya Baig, Ms. Savitha Srinivasan, Ms. Shivani Athavale, Mr. Ayon Biswas for collaborating with me on this project. We held many whiteboard discussion sessions.

References

[1] Mei Wu, Hongbin Ma, Mengyin Fu, and Chenguang Yang. Particle filter based simultaneous localization and mapping using landmarks with rplidar. In Honghai Liu, Naoyuki Kubota, Xiangyang Zhu, Rüdiger Dillmann, and Dalin Zhou, editors, *Intelligent Robotics and Applications*, pages 592–603, Cham, 2015. Springer International Publishing.

[2] Michael Montemerlo. *FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem with Unknown Data Association*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, July 2003.

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