# Visual-Inertial SLAM

Vaibhav Bishi
*Department of Electrical and Computer Engineering*
*University of California, San Diego*
La Jolla, USA
vbishi@ucsd.edu

***Abstract*—This project presents an approach for implementing Visual-Inertial Simultaneous Localization and Mapping (VI-SLAM) for a moving vehicle using an Extended Kalman Filter (EKF) with the sensor data from an Inertial Measurement Unit (IMU) and a stereo camera. The results show pretty good estimation of the vehicle trajectory and the landmarks around it.**

***Keywords—Visual-Inertial SLAM, Extended Kalman filter, IMU, stereo camera***

## I. Introduction

In this modern age of science and technology, the presence of autonomous vehicles and robots is becoming ever more profound by the day. With increasing complexities in task, we are inclined to develop better models and techniques for these autonomous agents to deal with different situations. One such problem which is associated with autonomous navigation is the Simultaneous Localization and Mapping (SLAM) problem. SLAM is a technique that utilizes the raw data obtained via sensors from the outside world directly to perform mapping and localization simultaneously. In this problem, an autonomous agent has to be able to identify its own state in an environment while simultaneously having to build the map around it. This is much like a chicken and an egg problem, giving it is a form of parameter estimation problem in some sense as we have to deal with probabilities.

In this project, we solve the SLAM problem for a vehicle using sensor data provided by an IMU (inertial data) and stereo camera (visual data), thus making it a Visual-Inertial SLAM problem. We implement the Extended Kalman Filter (EKF) approach which is a special case of a more general Bayes filter to achieve our objective, where the model is linearized considering Gaussian noise. We are able to get a very reliable estimate of the vehicle trajectory by its IMU pose along with the different landmark locations in the map as seen by the stereo camera.

## II. Problem Formulation

### A. Problem Statement

We are given data from two sensors mounted on the vehicle: the IMU which provides inertial measurements and the stereo camera which provides the visual measurements. Our objective is to estimate the localized state of the vehicle in the map while simultaneously estimating the locations of the landmarks in the map over time.

### B. Dataset

We are given two different datasets that correspond to different vehicle trajectories and landmark positions in the form of files '03.npz' and '10.npz'. We particularly have data from two kinds of sensors namely the IMU and stereo camera. The IMU provides inertial measurements in the form of linear and angular velocities in 3D coordinates in the IMU frame at different timestamps. The stereo camera provides visual measurements in the form of pixel coordinates of the landmark positions as observed by the left and right cameras at timestamps synchronized with the IMU measurements. Different parameters for pose transformations and stereo camera model are also provided.

### C. SLAM Problem

The localization problem deals with predicting the position of an agent in a given map, while the mapping problem deals with generating a map of the agent's surroundings as it progresses. The SLAM problem is a culmination of both localization and mapping with a limited understanding of both. Thus, given a sequence of control inputs $u_{0:T-1}$ and observations $z_T$, we are required to generate a sequence of the agent's states $x_{0:T}$ and the map $m$ surrounding it. We can relate these parameters by exploiting the decomposition of the joint probability distribution with Markov independence assumption as:

$$p(x_{0:T}, m, z_{0:T}, u_{0:T-1}) = p_0(x_0, m)\prod_{t=0}^{T} p_h(z_t|x_t, m)\prod_{t=1}^{T} p_f(x_t|x_{t-1}, u_{t-1})\prod_{t=0}^{T-1} p(u_t|x_t)$$

### D. EKF Approach

Our task is to solve the SLAM problem using EKF, which is a special kind of Bayes filter. The Bayes filter is a probabilistic inference technique for estimating the state of a dynamical system. Here, we will have to keep track of $p_{t|t}(x_t)$ and $p_{t+1|t}(x_{t+1})$ using a prediction step to incorporate the control inputs and an update step to incorporate the measurements.

Prediction step:

$$p_{t+1|t}(x) = \int p_f(x|s, u_t)p_{t|t}(s)ds$$

Update step:

$$p_{t+1|t+1}(x) = \frac{p_h(z_{t+1}|x)p_{t+1|t}(x)}{\int p_h(z_{t+1}|s)p_{t+1|t}(s)ds}$$

The EKF that we will implement is a special kind of Bayes filter. In Kalman filtering, the non-linear motion and observation models are linearized and have Gaussian noise in them. The EKF forces the predicted and updated Probability Distribution Functions (PDFs) to be Gaussian using a first-order Taylor series approximation to the motion and observation models around the state and noise means.

### E. Objectives

Firstly, given an IMU measurement $u_t = [v_t \ \ \omega_t]^T \in \mathbb{R}^6$, where $v_t$ and $\omega_t$ are the linear and angular velocities respectively, we need to predict the IMU pose $T_t \in SE(3)$ and get the vehicle trajectory over the time $t$.

Secondly, given stereo camera measurements $z_t \in \mathbb{R}^{4N_t}$, we need to estimate the homogeneous coordinates for the landmarks in the world frame $m \in \mathbb{R}^{4M}$ and get the locations of landmarks in the world along with the predicted trajectory from before. Here, $N_t$ is the number of observed visual features and $M$ is the number of landmarks.

Finally, given all the measurement data $v_t$, $\omega_t$ and $z_t$, we need to implement SLAM to get the vehicle trajectory and landmark locations in the world over time.

## III. Technical Approach

### A. IMU Localization via EKF prediction

Our first goal is to predict the IMU pose using the inertial measurements from the IMU in order to get the vehicle trajectory. We will implement the EKF prediction steps to achieve this. Firstly, we consider the IMU pose $T_t$ to have a mean $\mu_{t|t} \in SE(3)$ and covariance $\Sigma_{t|t} \in \mathbb{R}^{6X6}$ with the prior mean as identity matrix and prior covariance as zero matrix. We take the motion noise as a Gaussian random variable $w_t \sim N(0, W)$, where $W$ is a diagonal matrix with variance values of 0.03 and 0.002 for linear and angular velocities respectively, which is typically needed for the later parts in the project. For dead-reckoning however, these values are taken to be zero. This gives us a good idea of how the vehicle trajectory is supposed to look like without any errors induced.

The pose that can be obtained as:

$$T_t = \mu_{t|t}\exp(\delta\mu_{t|t})^\wedge$$

So, the motion model for $\mu_{t|t}$ with perturbation $\delta\mu_{t|t}$ with time discretization $\tau_t$ is given as:

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

Now, we implement the EKF prediction steps as:

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

where

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

Here, $u_t$ is the control input.

### B. Landmark Mapping via EKF update.

Our next objective is to estimate the landmark coordinates in the world using the visual measurements from the stereo camera. We will implement the EKF update steps to achieve this. Since, the number of observed landmarks are quite high, we find the best 500 and 1000 features based on the number of their occurrences to reduce computational complexity. We consider the landmark world frame coordinates (where the values for a particular landmark is averaged over all of its occurrences) to have a mean $\mu_t \in \mathbb{R}^{3M}$ and covariance $\Sigma_t \in \mathbb{R}^{3MX3M}$. Using the stereo calibration matrix $K_s$, extrinsics $oT_I \in SE(3)$, the IMU pose $T_{t+1}$ and the camera observation $z_{t+1} \in \mathbb{R}^{4N_{t+1}}$, we find the following quantities:

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

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

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

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

where,

Projection function and its derivative:

$$\pi(\mathbf{q}) := \frac{1}{q_3}\mathbf{q} \in \mathbb{R}^4 \qquad \frac{d\pi}{d\mathbf{q}}(\mathbf{q}) = \frac{1}{q_3}\begin{bmatrix} 1 & 0 & -\frac{q_1}{q_3} & 0 \\ 0 & 1 & -\frac{q_2}{q_3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{q_4}{q_3} & 1 \end{bmatrix} \in \mathbb{R}^{4\times4}$$

Now, we implement the EKF update steps as:

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

Here, $V$ is the observation noise.

Also, we converted the stereo camera observations in the form of pixels in left and right coordinates into the landmark coordinates in optical frame using the stereo camera model as:

$$\begin{bmatrix} u_L \\ v_L \\ u_R \\ v_R \end{bmatrix} = \underbrace{\begin{bmatrix} fs_u & 0 & c_u & 0 \\ 0 & fs_v & c_v & 0 \\ fs_u & 0 & c_u & -fs_u b \\ 0 & fs_v & c_v & 0 \end{bmatrix}}_{M} \frac{1}{z}\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

Once we have the landmark coordinates in optical frame ($P_o$), we can get the landmark coordinates in the world frame ($P_W$) as:

$$P_W = {}_wT_{IMU} \, {}_oT_{IMU} \, P_o$$

### C. VI-SLAM

Our final objective is to implement the VI-SLAM by combining both the EKF prediction and update steps to estimate the final vehicle localization and landmark mapping. In order to do this, we now need to simultaneously perform the prediction and update steps one after the other. The prediction step stays the same as in the first part, but now with non-zero motion noise, while the update steps have certain changes in the equations.

Firstly, we find the following quantities:

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

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

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

$$H_{t+1,i} = -K_s \frac{d\pi}{d\mathbf{q}}\left({}_oT_I \boldsymbol{\mu}_{t+1|t}^{-1} \underline{\mathbf{m}}_j\right) {}_oT_I \left(\boldsymbol{\mu}_{t+1|t}^{-1} \underline{\mathbf{m}}_j\right)^\odot \in \mathbb{R}^{4\times6}$$

where,

homogeneous coordinates $\underline{\mathbf{s}} \in \mathbb{R}^4$ and $\hat{\boldsymbol{\xi}} \in \mathfrak{se}(3)$:

$$\hat{\boldsymbol{\xi}}\underline{\mathbf{s}} = \underline{\mathbf{s}}^{\odot}\boldsymbol{\xi} \qquad \begin{bmatrix}\mathbf{s}\\1\end{bmatrix}^{\odot} := \begin{bmatrix}I & -\hat{\mathbf{s}}\\0 & 0\end{bmatrix} \in \mathbb{R}^{4\times 6}$$

Now, the EKF update steps are given as:

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

Here, V is the observation noise which is a diagonal matrix with variances equal to 5.
We have now implemented the VI-SLAM completely.

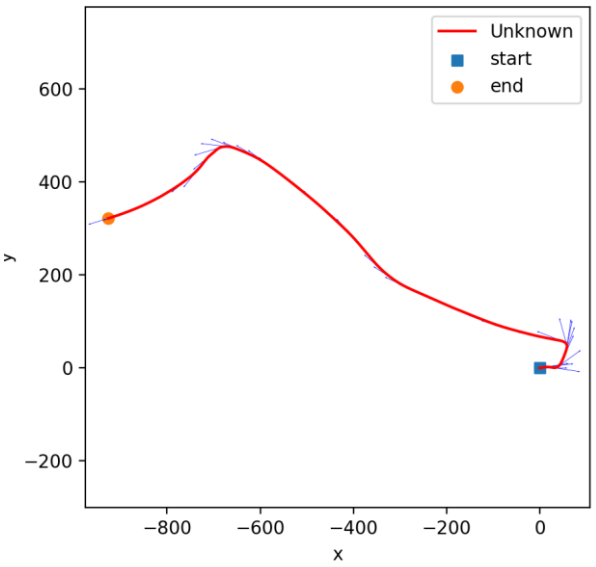
## IV. Results

In our project, we have implemented the VI-SLAM for two different datasets that correspond to different trajectories and landmark locations given in the two files '03.npz' and '10.npz'.

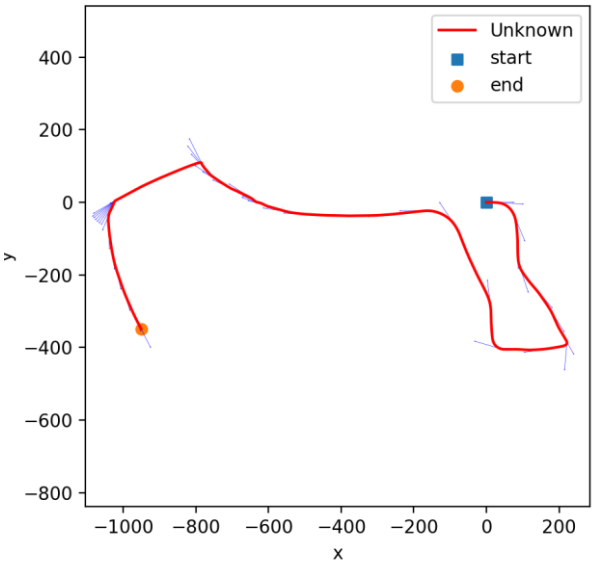
For part (a), we have plotted the errorless trajectories (dead-reckoning) for both the cases. For part (b), we have obtained the scatter plot for the updated landmark locations along with the vehicle trajectory over time, but now with non-zero Gaussian noise as defined earlier, for both cases. We have done this for the best 500 and also the best 1000 visual features in both cases. For part (c), we have plotted the final updated trajectories along with the landmark world coordinates achieved by VI-SLAM for dataset '03.npz'. We have tried different parameters for motion and observation model noise by tweaking the values and finally decided on these values. We have done this again for the best 500 and 1000 features. However, even after a lot of experimentation, we have been unable to plot the final trajectory with landmark mapping for dataset '10.npz' due to eventual singularity of the IMU pose covariance matrix. This happened at different time-steps on different runs of the code.

We notice that trajectories obtained in part (b) are almost the same as the dead-reckoned trajectories in part (a). It is worth noticing that, for both the datasets, upon increasing the number of landmarks to be considered for the plot, the trajectories look much closer to the dead-reckoned trajectory especially during turns. This can be attributed to the fact that more number of updated landmark features reinforces our vehicle trajectory. Also, more number of landmark points can be seen clustered together making the trajectory more believable. We see that, in the final plots obtained in part (c) for the dataset '03.npz', the plot with higher features has better accuracy and continuity than the one with lesser features. Although, it can be seen that the trajectories obtained are somewhat erratic in some places, the general trend of the trajectory along with the mapped landmarks seem reasonable. This can be due to the fact that we are taking very few features out of all and can be resolved by taking better noise parameters. The reason for unavailability of the plots for dataset '10.npz' can be because of less optimum parameters in terms of noise in motion and observation models which could have eventually led to the singularity in the covariance matrix. All the different plots obtained have been attached below.

Dead-reckoning trajectory plots:

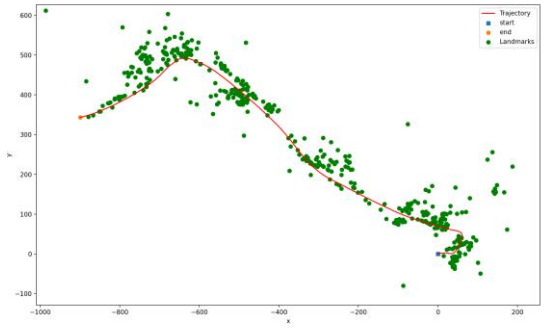



Dead reckoning trajectory for '03.npz'



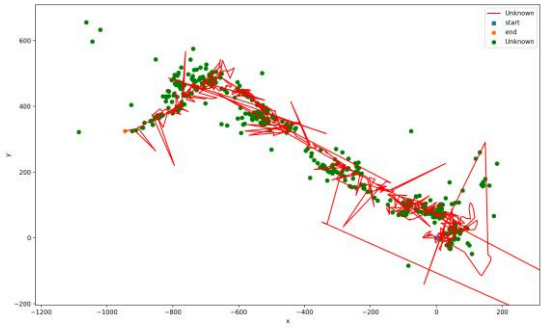

Dead reckoning trajectory for '10.npz'

Trajectory with landmark mapping for ’03.npz’:

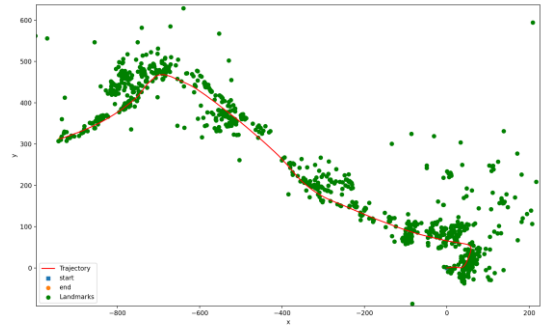



Trajectory with 500 landmarks

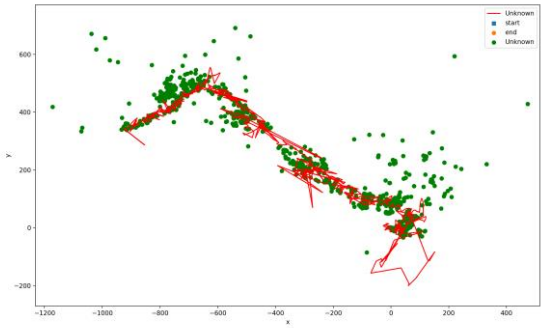
VI-SLAM trajectory with landmark mapping for ’03.npz’:




Trajectory with 500 landmarks

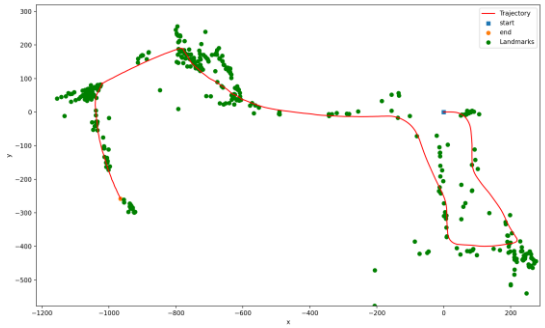



Trajectory with 1000 landmarks

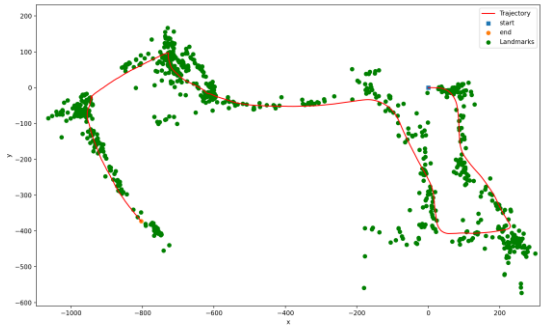



Trajectory with 1000 landmarks

Trajectory with landmark mapping for ’10.npz’:




Trajectory with 500 landmarks




Trajectory with 1000 landmarks