ECE 276A Project 3: Visual-Inertial SLAM

Winston Chou
PID: A17460970

***Abstract*—This project focuses on implementing a visual-inertial simultaneous localization and mapping (SLAM) system using an extended Kalman filter (EKF). The system integrates measurements from an inertial measurement unit (IMU) and a stereo camera to estimate both the trajectory of the robot and the positions of static landmarks in the environment. The IMU provides linear and angular velocity data, while the stereo camera captures visual features with precomputed correspondences between left and right frames. The SLAM process involves two main steps: an EKF prediction step based on IMU kinematics to estimate the robot's pose and an EKF update step using visual observations to refine landmark positions. The project assumes known extrinsic and intrinsic calibration parameters for the sensors. Results demonstrate the effectiveness of the proposed approach in estimating accurate robot trajectories and mapping landmark positions, despite challenges such as noisy measurements and partial observability.**

Keywords—SLAM, Markov Chain, pose estimation, multi-sensor fusion, IMU, RGBD camera

I. INTRODUCTION

Simultaneous localization and mapping (SLAM) is a fundamental problem in robotics that involves estimating a robot's trajectory while simultaneously building a map of its environment. This capability is critical for autonomous navigation in various applications, including self-driving cars, drones, and robotic exploration. SLAM is particularly challenging due to uncertainties in sensor measurements, dynamic environments, and computational complexity.

This project aims to solve the SLAM problem using data from an IMU and a stereo camera mounted on a vehicle. The IMU provides linear and angular velocity measurements, while the stereo camera captures visual features used for landmark mapping. By leveraging an extended Kalman filter (EKF), the system integrates these sensor inputs to estimate both the robot's pose and the positions of static landmarks in 3D space.

The proposed approach consists of two key components: an EKF prediction step that uses IMU kinematics to estimate the robot's trajectory over time, and an EKF update step that incorporates stereo camera observations to refine landmark positions. Known extrinsic calibration between the IMU and camera frames, as well as intrinsic camera parameters, are utilized to ensure accurate sensor fusion.

This report details the implementation of the visual-inertial SLAM system, including problem formulation, technical approach, results, and analysis. The findings highlight the potential of combining visual and inertial data to achieve robust localization and mapping in real-world scenarios.

Fig. 1. Sensor Setup. Clearpath Jackal robots on MIT's campus, equipped withIMU, 3-D LIDAR scanner, and an RGBD camera, Realsense D455. [Source: ECE276A_PR3.pdf]

II. PROBLEM FORMULATION

A. Simultaneous Localization And Mapping

SLAM can be described as a probabilistic process modeled as a Markov Chain. At discrete time steps t, the robot's pose x_t evolves based on the control input u_t and motion noise w_t. The pose at the next time step $t+1$ is determined by a probabilistic function:

$$x_{t+1} = f(x_t, u_t, w_t) \sim p_f(\cdot\,|x_t, u_t)$$

Where w_t represents motion noise. This relationship is commonly referred to as the **motion model**.

To map the environment, the robot collects observations at each time step. Let the observation at time t be denoted as z_t and the environment as m. The sensor observations follow a probabilistic relationship:

$$z_t = h(x_t, m, v_t) \sim p_h(\cdot\,|x_t, m)$$

Where v_t represents observation noise. This relationship is commonly referred to as the **observation model**.

The SLAM problem involves estimating the environment m and the robot's poses x_t using observations $z_0, \dots, z_t$ and control inputs $u_0, \dots, u_{t-1}$ at each time step t. The goal is to compute the joint probability distribution of the environment and robot poses conditioned on the observations and control inputs. This probabilistic formulation captures the uncertainty inherent in SLAM:

$$p(m, x_t | z_{0:t}, u_{0:t-1})$$

The relationship between the environment and robot poses is challenging to determine directly. However, leveraging the Markov assumptions allows for decomposing the joint probability density function into manageable components. This decomposition includes terms for the initial state of the robot and environment, observation likelihoods based on sensor readings, and motion probabilities derived from control inputs and previous poses.

$$p(x_{0:t}, z_{0:t}, u_{0:t-1}, m)$$
$$= p_{0|0}(x_0, m) \prod_{i=0}^{t} p_h(z_i | x_i, m) \prod_{i=1}^{t} p_f(x_i | x_{i-1}, u_{i-1})$$

In practical SLAM implementations, maximum likelihood estimation (MLE) is often used to find optimal values for the robot's trajectory $x_{0:t}$ and the environment m. The optimization process involves maximizing the sum of log-likelihoods for sensor observations and motion probabilities over all time steps. This approach helps determine both the robot's poses and a map of the environment effectively.

$$\max_{x_{0:t},m} \sum_{i=0}^{t} \log\left(p_h(z_i|x_i,m)\right) + \sum_{i=1}^{t} \log\left(p_f(x_i|x_{i-1},u_{i-1})\right)$$

B. Bayesian Filtering

Bayes filtering is a probabilistic method used to estimate the state of dynamic systems, such as a robot's pose, by combining information from control inputs and observations. This technique relies on the Markov assumptions and Bayes' rule to infer the system's state over time. The Bayes filter operates in two main steps:

1) *Prediction Step:*
   In this step, the prior probability distribution of the system state at time t, along with the control input, is used to predict the state at the next time step $t+1$. The motion model governs this process, and the predicted probability distribution is computed by integrating over all possible states at time t:

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

2) *Update Step:*
   Once a new observation is received at time $t+1$, the predicted probability distribution is updated using the observation model. This step incorporates measurement information to refine the estimate of the system's state at $t+1$. The posterior distribution is normalized by dividing by a constant factor (marginal likelihood) to ensure it remains a valid probability distribution:

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

C. Landmark-based Mapping

Landmark-based mapping focuses on creating a map of the environment using noisy and uncertain sensor observations, assuming the robot's poses are known. The environment is modeled as a set of static landmarks, with each landmark represented by its location in 3D space. A landmark's position is denoted as m_j, where j ranges from 1 to M, the total number of landmarks. Collectively, the landmarks are represented as a matrix $m \in \mathbb{R}^{3\times M}$, with each landmark specified by three numerical values corresponding to its coordinates.

The robot can detect landmarks at each time step t, and the observations are denoted as z_t. Since multiple landmarks may be sensed simultaneously, z_t represents a composite observation encompassing all detected landmarks at that time.

The goal of this mapping process is to estimate the locations of the landmarks based on the robot's pose x_t and the observations z_t. This estimation relies on an observation model, which defines the probabilistic relationship between the observations, robot pose, environment landmarks, and measurement noise.

$$p(z_t|x_t,m,\Delta_t)$$

An index map Δ_t is used to track which landmarks correspond to specific observations. At each time step t, the robot observes N_t landmarks, with each observation denoted as $z_{t,i} \in \mathbb{R}^4$ (Homogeneous Coordinates), where $i = 1,\dots,N_t$. The data association map $\Delta_t(j)$ specifies that the j^{th} landmark corresponds to the observation indexed by $i = \Delta_t(j)$.

D. Sensors Setup

The proposed solution addresses the SLAM problem using data from an IMU and a stereo camera mounted on a vehicle. The IMU provides measurements of linear velocity ($v_t \in \mathbb{R}^3$) and angular velocity ($\omega_t \in \mathbb{R}^3$), both expressed in the IMU's frame of reference. The stereo camera captures visual data, with precomputed visual features and correspondences established between the left and right camera frames as well as across time steps (data association).

At each time step t, the visual features are represented as $z_t \in \mathbb{R}^{4M}$, where each column corresponds to a landmark. Specifically, the i^{th} column contains the pixel coordinates of the i^{th} landmark in both the left and right camera images. If a landmark is not observable at time t, its corresponding column in z_t is set to $[-1 \quad -1 \quad -1 \quad -1]^T$.

The system assumes that both followings are known:

1) *The transformation from the IMU frame to the stereo camera's optical frame* (${}_OT_I \in SE(3)$) *is known (extrinsic calibration)*
2) *The stereo camera calibration matrix* (M_{stereo}) *is also known (intrinsic calibration). The calibration matrix is defined as:*

$$M_{stereo} = \begin{bmatrix} fs_u^{Left} & 0 & c_u^{Left} & 0 \\ 0 & fs_v^{Left} & c_v^{Left} & 0 \\ fs_u^{Right} & 0 & c_u^{Right} & -fs_u^{Right}b \\ 0 & fs_v^{Right} & c_v^{Right} & 0 \end{bmatrix}$$

Where f is the focal length, s_u, s_v are pixel scaling, c_u and c_v are the principal points, and b is the stereo baseline.

III. Technical Approach

The implementation of the project is presented as follows:

A. Extended Kalman Filter

Extended Kalman Filter (EKF) is a nonlinear extension of the Kalman Filter designed to handle systems with nonlinear dynamics and observations. It operates by linearizing the system around the current estimate of the mean and covariance using a moment-matching approach. The EKF is fundamentally a Bayes filter, relying on several key assumptions:

- The prior probability density function (pdf), $p_{t|t}$, is Gaussian.
- The state transition model is affected by Gaussian noise, expressed as:

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

- The observation model is also influenced by Gaussian noise:

$$z_t = h(x_t, m, v_t), \qquad v_t \sim \mathcal{N}(0, V)$$

- The process noise w_t and measurement noise v_t are independent of each other, the state x_t, and across time steps.
- The posterior pdf is approximated as Gaussian using moment matching.

The primary challenge in applying the EKF to nonlinear systems lies in the fact that the predicted and updated pdfs are not inherently Gaussian and cannot be computed in closed form. To address this, moment matching is employed to approximate these pdfs as Gaussians by evaluating their first and second moments (mean and covariance). This ensures that the EKF maintains a consistent probabilistic framework while handling nonlinearity effectively.

The Extended Kalman Filter (EKF) uses a first-order Taylor series expansion to approximate the integrals required for implementing a nonlinear Kalman Filter. The motion model is approximated as follows:

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

Where:

$$\boldsymbol{F_t} = \frac{\partial f}{\partial x}(\mu_{t|t}, u_t, 0)$$

$$\boldsymbol{Q_t} = \frac{\partial f}{\partial w}(\mu_{t|t}, u_t, 0)$$

Similarly, the observation model is approximated as:

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

Where:

$$\boldsymbol{H_{t+1}} = \frac{\partial h}{\partial x}(\mu_{t+1|t}, 0)$$

$$\boldsymbol{R_{t+1}} = \frac{\partial h}{\partial v}(\mu_{t+1|t}, 0)$$

Based on these approximations, the EKF models are defined as follows:

Prior:

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

Motion Model:

The state transition includes Gaussian process noise:

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

Linearization parameters:

$$\boldsymbol{F_t} \coloneqq \frac{\partial f}{\partial x}(\mu_{t|t}, u_t, 0)$$

$$\boldsymbol{Q_t} \coloneqq \frac{\partial f}{\partial w}(\mu_{t|t}, u_t, 0)$$

Observation Model:

Observations are affected by Gaussian measurement noise:

$$z_t = h(x_t, v_t), \qquad v_t \sim \mathcal{N}(0, V)$$

Linearization parameters:

$$\boldsymbol{H_t} \coloneqq \frac{\partial h}{\partial x}(\mu_{t|t-1}, 0)$$

$$\boldsymbol{R_t} \coloneqq \frac{\partial h}{\partial v}(\mu_{t|t-1}, 0)$$

Prediction Step:

Predicted Mean and Covariance

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

Update Step:

Kalman Gain

$$\boldsymbol{K_{t+1|t}} \coloneqq \Sigma_{t+1|t} \boldsymbol{H_{t+1}}^T \left(\boldsymbol{H_{t+1}} \Sigma_{t+1|t} \boldsymbol{H_{t+1}}^T + \boldsymbol{R_{t+1}} V \boldsymbol{R_{t+1}}^T\right)^{-1}$$

Posterior Mean & Covariance

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

B. Visual Mapping: Landmark Mapping via EKF update

In the context of the landmark-based visual mapping problem, it is assumed that the inverse IMU pose ${}_W T_{I,t}{}^{-1} \in SE(3)$ is known. Additionally, the landmarks are assumed to be static, and the data association $\Delta_t : \{1, \dots, M\} \to \{1, \dots, N_t\}$ —which specifies which landmarks are observed at each time step t— is pre-computed by an external algorithm.

The observation model incorporates a Gaussian prior and observation noise and is expressed as follows:

Prior:

$$m | z_{0:t} \sim \mathcal{N}(\mu_t, \Sigma_t)$$
$$\mu_t \in \mathbb{R}^{3M}, \qquad \Sigma_t \in \mathbb{R}^{3M \times 3M}$$

Observation Model:

$$z_{t,i} = h({}_W T_{I,t}{}^{-1}, m_j) + v_{t,i} = M_{stereo} \boldsymbol{\pi}({}_O T_I \, {}_W T_{I,t}{}^{-1} \underline{m}_j) + v_{t,i}$$

$$m | z_{0:t} \sim \mathcal{N}(\mu_t, \Sigma_t)$$
$$v_{t,i} \sim \mathcal{N}(0, V)$$

Homogeneous Coordinates:

$$\underline{m}_j = \begin{bmatrix} m_j{}^T & 1^T \end{bmatrix}^T$$

Where:

- $\mu_t \in \mathbb{R}^{3M}$ represents the expected locations of all landmarks.
- $\Sigma_t \in \mathbb{R}^{3M \times 3M}$ is the covariance matrix of the estimate.
- M_{stereo} is the calibration matrix mentioned in Problem Formulation

The projection function $\boldsymbol{\pi}$ and its derivative are defined as:

$$\boldsymbol{\pi}(q) = \frac{1}{q_3} q \in \mathbb{R}^4$$

$$\frac{d\pi}{dq}(q) = \frac{1}{q_3}\begin{bmatrix} 1 & 0 & -\frac{q_1}{q_3} & 0 \\ 0 & 1 & -\frac{q_2}{q_3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{q_4}{q_3} & 1 \end{bmatrix}$$

All observations at time step t can be stacked into a single vector:

$$z_t = M_{stereo}\boldsymbol{\pi}\left({}_oT_I \, {}_WT_{I,t}{}^{-1}\underline{m}\right) + v_t$$
$$v_t \sim \mathcal{N}(0, I \otimes V)$$

where $\otimes$ is the Kronecker product.

For the Extended Kalman Filter (EKF), the derivative of the observation model with respect to the landmark positions must be computed. Considering a small perturbation $\delta\mu_{t,j}$ for the landmark j:

$$m_j = \mu_{t,j} + \delta\mu_{t,j}$$

Using a first-order Taylor series approximation, the observation model for feature i at time step t becomes:

$$z_{t,i} = M_{stereo}\boldsymbol{\pi}\left({}_oT_I \, {}_WT_{I,t}{}^{-1}\left(\underline{\mu_{t,j}} + \delta\underline{\mu_{t,j}}\right)\right) + v_{t,i}$$
$$= M_{stereo}\boldsymbol{\pi}\left({}_oT_I \, {}_WT_{I,t}{}^{-1}\left(\underline{\mu_{t,j}} + \mathbf{P}^T\delta\underline{\mu_{t,j}}\right)\right) + v_{t,i}$$
$$\approx M_{stereo}\boldsymbol{\pi}\left({}_oT_I \, {}_WT_{I,t}{}^{-1}\underline{\mu_{t,j}}\right)$$
$$+ \; M_{stereo}\frac{d\pi}{dq}\left({}_oT_I \, {}_WT_{I,t}{}^{-1}\underline{\mu_{t,j}}\right) {}_oT_I \, {}_WT_{I,t}{}^{-1}\mathbf{P}^T\delta\underline{\mu_{t,j}} + v_{t,i}$$

Where $\mathbf{P}$ is defined as $\mathbf{P} = [I \quad 0] \in \mathbb{R}^{3\times4}$.

EKF Update Steps:

The EKF update steps for landmark mapping are as follows:

Predicted Observations:

$$\tilde{z}_{t,i} = M_{stereo}\boldsymbol{\pi}\left({}_oT_I \, {}_WT_{I,t}{}^{-1}\underline{\mu_{t,j}}\right) \in \mathbb{R}^4$$

for all observed features $i = 1, \dots, N_t$.

Observation Matrix:

$$\boldsymbol{H_{t,i,j}} = \begin{cases} M_{stereo}\frac{d\pi}{dq}\left({}_oT_I \, {}_WT_{I,t}{}^{-1}\underline{\mu_{t,j}}\right) {}_oT_I \, {}_WT_{I,t}{}^{-1}\mathbf{P}^{\mathrm{T}}, & i = j \\ 0, & i \neq j \end{cases}$$

The full matrix is stacked as: $\boldsymbol{H_t} \in \mathbb{R}^{4N_t\times3M}$.

Kalman Gain:

$$\boldsymbol{K_t} = \Sigma_t\boldsymbol{H_t}^T\left(\boldsymbol{H_t}\Sigma_t\boldsymbol{H_t}^T + I \otimes V\right)^{-1}$$

State Update:

$$\mu_{t+1} = \mu_t + \boldsymbol{K_t}(z_t - \tilde{z}_{t,i})$$
$$\Sigma_{t+1} = (I - \boldsymbol{K_t}\boldsymbol{H_t})\Sigma_t$$

Since landmarks are assumed to be static, there is no prediction step for their positions. Each landmark's position is initialized when it is first observed using:

$$z_{t,i} = M_{stereo}\boldsymbol{\pi}\left({}_oT_I \, {}_WT_{I,t}{}^{-1}\underline{m}_j\right)$$

C. EKF-Based Visual-Inertial Odometry

The localization problem aims to estimate the inverse IMU pose of the robot, $P_t = {}_WT_{I,t} \in SE(3)$, using IMU measurements $u_t = [{v_t}^T, {\omega_t}^T]^T$, visual feature observations $z_{0:t}$, and landmark coordinates $m \in \mathbb{R}^{3\times M}$ in the world frame. Similar to the visual mapping problem, data association between observations and landmarks is assumed to be pre-computed by an external algorithm.

The motion model incorporates a Gaussian prior and process noise:

Prior:

$$P_t|z_{0:t}, u_{0:t-1} \sim \mathcal{N}(\mu_{t|t}, \Sigma_{t|t})$$
$$\mu_{t|t} \in SE(3)$$
$$\Sigma_{t|t} \in \mathbb{R}^{6\times6}$$

Motion Model:

$$P_{t+1} = P_t \exp\left(\tau(u_t + w_t)^\wedge\right)$$
$$u_t = [{v_t}^T, {\omega_t}^T]^T$$
$$P_t|z_{0:t}, u_{0:t-1} \sim \mathcal{N}(\mu_{t|t}, \Sigma_{t|t})$$
$$w_t \sim \mathcal{N}(0, W)$$

where:

- $u_t = [{v_t}^T, {\omega_t}^T]^T$ combines linear and angular velocity.
- τ is the time difference between two consecutive frames.
- $w_t \sim \mathcal{N}(0, W)$ represents Gaussian process noise.

To separate the deterministic motion from noise, discrete-time perturbation techniques are used to rewrite the motion model in terms of nominal kinematics and zero-mean perturbations:

$$\mu_{t+1|t} = \mu_{t|t}\exp(\tau u_t{}^\wedge)$$
$$\delta\mu_{t+1|t} = \exp(-\tau u_t{}^\curlywedge)\,\delta\mu_{t|t} + \sqrt{\tau}w_t$$

Where

$$u_t = \begin{bmatrix} v_t \\ \omega_t \end{bmatrix} \in \mathbb{R}^6$$
$$u_t{}^\wedge = \begin{bmatrix} \omega_t{}^\wedge & v_t \\ 0^T & 0 \end{bmatrix} \in \mathbb{R}^{4\times4}$$
$$u_t{}^\curlywedge = \begin{bmatrix} \omega_t{}^\wedge & v_t{}^\wedge \\ 0 & \omega_t{}^\wedge \end{bmatrix} \in \mathbb{R}^{6\times6}$$

With the motion model defined, the Extended Kalman Filter prediction steps are listed as follows:

$$\mu_{t+1|t} = \mu_{t|t}\exp(\tau u_t{}^\wedge)$$
$$\Sigma_{t+1|t} = \mathbb{E}\left[\delta\mu_{t+1|t}\delta\mu_{t+1|t}{}^T\right]$$
$$= \exp(-\tau u_t{}^\curlywedge)\,\delta\mu_{t|t}\exp(-\tau u_t{}^\curlywedge)^T + \tau W$$

Observation Model:

The observation model is consistent with the visual mapping problem. For each observation at time step $t+1$, a first-order Taylor series approximation is applied to account for perturbations in the inverse IMU pose:

$$z_{t+1,i} = M_{stereo}\boldsymbol{\pi}\left({}_oT_I\left(\mu_{t+1|t}\exp(\delta\mu_{t+1|t}{}^\wedge)\right)^{-1}\underline{m}_j\right) + v_{t+1,i}$$

$$
\begin{aligned}
&\approx M_{stereo}\boldsymbol{\pi}\left({}_{O}T_{I}(I - \delta\mu_{t+1|t}\wedge)\mu_{t+1|t}{}^{-1}\underline{m}_{j}\right) + v_{t+1,i} \\
&= M_{stereo}\boldsymbol{\pi}\Big({}_{O}T_{I}\mu_{t+1|t}{}^{-1}\underline{m}_{j} \\
&\qquad - {}_{O}T_{I}\left(\mu_{t+1|t}{}^{-1}\underline{m}_{j}\right)^{\odot}\delta\mu_{t+1|t}\Big) + v_{t+1,i} \\
&\approx M_{stereo}\boldsymbol{\pi}\left({}_{O}T_{I}\mu_{t+1|t}{}^{-1}\underline{m}_{j}\right) \\
&- M_{stereo}\frac{\boldsymbol{d\pi}}{\boldsymbol{dq}}\left({}_{O}T_{I}\mu_{t+1|t}{}^{-1}\underline{m}_{j}\right) {}_{O}T_{I}\left(\mu_{t+1|t}{}^{-1}\underline{m}_{j}\right)^{\odot}\delta\mu_{t+1|t} \\
&+ v_{t,i}
\end{aligned}
$$

Where for any homogeneous coordinates $s \in \mathbb{R}^4$:

$$
\begin{bmatrix} s \\ I \end{bmatrix}^{\odot} = \begin{bmatrix} I & s\wedge \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 6}
$$

EKF Update Steps:

The EKF update steps for visual-inertial odometry are as follows:

$$
\begin{gathered}
P_{t+1}|z_{0:t}, u_{0:t} \sim \mathcal{N}(\mu_{t+1|t}, \Sigma_{t+1|t}) \\
\mu_{t+1|t} \in SE(3) \\
\Sigma_{t+1|t} \in \mathbb{R}^{6\times 6}
\end{gathered}
$$

Predicted Observations:

$$
\tilde{z}_{t+1,i} = M_{stereo}\boldsymbol{\pi}\left({}_{O}T_{I}\mu_{t+1|t}{}^{-1}\underline{m}_{j}\right) \in \mathbb{R}^4
$$

for all observed features $i = 1, \dots, N_{t+1}$.

Observation Matrix:

$$
\boldsymbol{H_{t+1,i}} = -M_{stereo}\frac{\boldsymbol{d\pi}}{\boldsymbol{dq}}\left({}_{O}T_{I}\mu_{t+1|t}{}^{-1}\underline{m}_{j}\right) {}_{O}T_{I}\left(\mu_{t+1|t}{}^{-1}\underline{m}_{j}\right)^{\odot}
$$

The full matrix is stacked as:

$$
\boldsymbol{H_{t+1|t}} = \begin{bmatrix} \boldsymbol{H_{t+1,1}} \\ \boldsymbol{H_{t+1,2}} \\ \vdots \\ \boldsymbol{H_{t+1,N_{t+1}}} \end{bmatrix} \in \mathbb{R}^{4N_{t+1}\times 6}
$$

Kalman Gain:

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

State Update:

$$
\begin{aligned}
\mu_{t+1|t+1} &= \mu_{t+1|t}\exp\left((\boldsymbol{K_{t+1|t}}(z_{t+1} - \tilde{z}_{t+1,i}))\wedge\right) \\
\Sigma_{t+1|t+1} &= (I - \boldsymbol{K_{t+1|t}}\boldsymbol{H_{t+1|t}})\Sigma_{t+1|t}
\end{aligned}
$$

D. EKF-Based Visual-Inertial SLAM

To simultaneously estimate the robot's pose and the positions of landmarks, the proposed approach merges the prediction and update steps from EKF-based visual mapping and visual-inertial odometry. The joint state and covariance are defined under the Gaussian assumption as follows:

$$
\begin{gathered}
\mu = \begin{bmatrix} \mu_m \\ \mu_p \end{bmatrix} \in \mathbb{R}^{6+3M} \\
\Sigma \in \mathbb{R}^{(6+3M)\times(6+3M)}
\end{gathered}
$$

where μ_m is the estimated landmark position and μ_p is the estimated six degrees of freedom of IMU robot pose.

Prediction Step:

Since landmarks are assumed static, the prediction step only applies to the IMU pose. The joint state and covariance are updated using the IMU measurements u_t as follows:

$$
\begin{gathered}
\mu_{t+1|t} = \begin{bmatrix} \mu_{m,t+1|t} \\ \mu_{p,t+1|t} \end{bmatrix} = \begin{bmatrix} \mu_{m,t|t} \\ \mu_{p,t|t}\exp(\tau u_t\wedge) \end{bmatrix} \\
\Sigma_{t+1|t} = \boldsymbol{F_t}\Sigma_{t|t}\boldsymbol{F_t}^{T} + \boldsymbol{W} \\
\boldsymbol{F_t} = \begin{bmatrix} I & 0 \\ 0 & \exp(-\tau u_t\wedge) \end{bmatrix} \\
\boldsymbol{W} = \begin{bmatrix} 0 & 0 \\ 0 & \tau W_p \end{bmatrix}
\end{gathered}
$$

where Wp is the process noise covariance.

Update Step:

The update step combines the visual mapping update with visual-inertial odometry. The equations are as follows:

Predicted Observations:

$$
\tilde{z}_{t+1,i} = M_{stereo}\boldsymbol{\pi}\left({}_{O}T_{I}\mu_{p,t+1|t}{}^{-1}\mu_{m,t+1|t}\right) \in \mathbb{R}^4
$$

for all observed features $i = 1, \dots, N_{t+1}$.

Observation Matrices:

$$
\boldsymbol{H_{m,t+1,i}} = M_{stereo}\frac{\boldsymbol{d\pi}}{\boldsymbol{dq}}\left({}_{O}T_{I}\mu_{p,t+1|t}{}^{-1}\underline{\mu}_{t,j}\right) {}_{O}T_{I}\,{}_{W}T_{I,t}{}^{-1}P^{T}
$$

$$
\begin{aligned}
&\boldsymbol{H_{p,t+1,i}} \\
&= -M_{stereo}\frac{\boldsymbol{d\pi}}{\boldsymbol{dq}}\left({}_{O}T_{I}\mu_{p,t+1|t}{}^{-1}\underline{m}_{j}\right) {}_{O}T_{I}\left(\mu_{p,t+1|t}{}^{-1}\underline{m}_{j}\right)^{\odot}
\end{aligned}
$$

The full matrix is stacked as:

$$
\begin{aligned}
\boldsymbol{H_{t+1|t}} &= \begin{bmatrix} \boldsymbol{H_{m,t+1|t}} & \boldsymbol{H_{p,t+1|t}} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{H_{m,t+1,1}} & \boldsymbol{H_{p,t+1,1}} \\ \boldsymbol{H_{m,t+1,2}} & \boldsymbol{H_{p,t+1,2}} \\ \vdots & \vdots \\ \boldsymbol{H_{m,t+1,N_{t+1}}} & \boldsymbol{H_{p,t+1,N_{t+1}}} \end{bmatrix} \in \mathbb{R}^{4N_{t+1}\times(3M+6)}
\end{aligned}
$$

Kalman Gain:

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

State Update:

$$
\begin{aligned}
\mu_{t+1|t+1} &= \begin{bmatrix} \mu_{m,t+1|t+1} \\ \mu_{p,t+1|t+1} \end{bmatrix} \\
&= \begin{bmatrix} \mu_{m,t+1|t} + \boldsymbol{K_{t+1|t}}(z_t - \tilde{z}_{t,i}) \\ \mu_{p,t+1|t}\exp\left((\boldsymbol{K_{t+1|t}}(z_{t+1} - \tilde{z}_{t+1,i}))\wedge\right) \end{bmatrix} \\
\Sigma_{t+1|t+1} &= (I - \boldsymbol{K_{t+1|t}}\boldsymbol{H_{t+1|t}})\Sigma_{t+1|t}
\end{aligned}
$$

This combined EKF approach leverages both IMU measurements for pose prediction and stereo camera observations for landmark updates, enabling simultaneous estimation of robot trajectory and environmental mapping.

IV. RESULTS

The proposed EKF-based SLAM algorithm has been tested on three datasets, all collected from real-world driving scenarios. The hyperparameters used in the algorithm remain consistent across all datasets, except for adjustments to the number of landmarks. This decision was made to simulate realistic scenarios where the algorithm operates online and does not allow for manual fine-tuning of parameters for every unknown environment. Table 1 contains details of the selected hyperparameters.

Table 1: Hyperparameters

Prior landmark estimate covariance, Σ_m	$0.01 * I \in R^{3M \times 3M}$
Prior pose estimate covariance, Σ_p	$0.001 * I \in R^{6 \times 6}$
Observation noise covariance, V	$0.05 * I \in R^{4 \times 4}$
Process noise covariance, W_p	$0.001 * I \in R^{6 \times 6}$

The algorithm's performance is compared for two cases:

1. **IMU-based landmark mapping**:
    - In this case, the trajectory is estimated solely using IMU measurements and EKF prediction, without any observations from visual features.
    - This approach primarily focuses on estimating the positions of landmarks.
2. **Visual EKF SLAM**:
    - This case simultaneously predicts and updates both the positions of landmarks and the robot pose.

The hyperparameters were calibrated using the second test case. The results from this test case showed the parameters can be tuned better as covariance matrices can significantly affect the outcome of Kalman Gain.

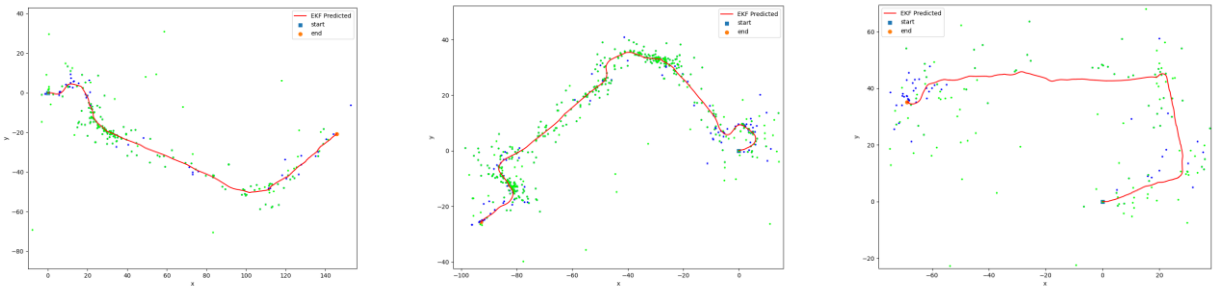
The results of the EKF-based SLAM algorithm are shown in Figures 2 and 3, where Figure 2 illustrates the trajectory produced solely by the IMU-based prediction, and Figure 3 depicts the trajectory generated using the EKF SLAM method. In Figure 3, the red line represents the estimated robot trajectory, and the blue arrows indicate the positions of visual landmarks. It is evident that the EKF SLAM trajectory exhibits significant instability, with erratic corrections and landmarks densely scattered near the origin. This behavior suggests that the landmark updates, likely influenced by noisy observations, destabilize the trajectory and cause divergence. In contrast, Figure 2 demonstrates a much smoother trajectory, indicating that the IMU motion model alone provides stable predictions. However, the lack of corrections from visual observations means that the trajectory gradually drifts and the associated landmark estimates remain less accurate. These results highlight the importance of properly tuning the EKF parameters and accurately modeling observation noise to balance the reliance on IMU data and visual feature updates for reliable SLAM performance.

Unfortunately, due to the absence of ground truth data, the algorithm's performance could not be quantitatively evaluated. For example, incorporating high-precision GPS data could have provided ground truth for the robot trajectory. This would enable performance metrics for the SLAM algorithm and potential improvements in the future.

Conclusion

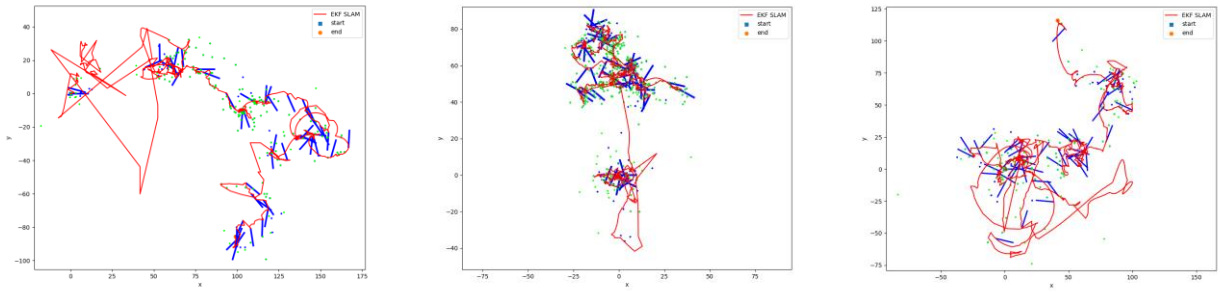
Hyperparameter Calibration: Tested on a closed-loop driving scenario, the algorithm demonstrates its ability to produce reliable trajectory estimations and landmark distributions.

- IMU-only mapping struggles due to the lack of corrections from visual observations.
- Visual SLAM provides better trajectory and landmark estimates through simultaneous updating.
- The absence of ground truth data (e.g., high-precision GPS) limits quantitative analysis of algorithm performance.



(a) dataset 00 (b) dataset 01 (c) dataset 02

Fig. 2. 1. IMU-based landmark mapping



(a) dataset 00 (b) dataset 01 (c) dataset 02

Fig. 3. Visual EKF SLAM

REFERENCES

[1] N. Atanasov, UCSD ECE276A: Sensing & Estimation in Robotics (Winter 2025), https://natanaso.github.io/ece276a/schedule.html.