Project 3: Visual-Inertial SLAM

Mazeyu Ji, PID:A59023027

***Abstract*—This project addresses the challenge of simultaneous localization and mapping (SLAM) using visual-inertial data, a critical component in the field of robotics for enabling autonomous navigation in complex environments. By integrating measurements from an inertial measurement unit (IMU) and a stereo camera system, we implement an extended Kalman filter (EKF) based approach to achieve accurate SLAM. Our methodology encompasses two main phases: firstly, an IMU localization via EKF prediction, leveraging SE(3) kinematics for estimating the IMU's pose over time; secondly, landmark mapping through EKF updates, utilizing visual feature observations to estimate the positions of static landmarks within the environment. This dual approach allows for the effective combination of inertial and visual data, addressing the challenges of environmental mapping and robot localization simultaneously. The project outcomes demonstrate the feasibility and effectiveness of visual-inertial SLAM in providing a comprehensive understanding of the surroundings. This work not only showcases the implementation of a complex SLAM system but also lays the groundwork for future enhancements in robotic sensing and estimation.**

***Index Terms*—Visual-inertial SLAM, Extended Kalman filter, Visual feature, SE3 Kinematics**

I. Introduction

THE challenge of understanding and interacting with complex environments is central to the advancement of robotics, especially in the context of autonomous navigation. Simultaneous Localization and Mapping (SLAM) stands as a cornerstone in this field, enabling robots to concurrently build a map of their environment while also determining their position within it. Among the various approaches to SLAM, Visual-Inertial SLAM (VI-SLAM) has emerged as particularly promising, thanks to its ability to combine the high-rate inertial measurements with the rich spatial information provided by visual sensors. This integration offers a robust solution to the SLAM problem, capable of dealing with the inherent limitations present when either modality is used in isolation.

The primary objective of this project was to implement a VI-SLAM system leveraging an Extended Kalman Filter (EKF), using data from an inertial measurement unit (IMU) and a stereo camera. This approach involves predicting the trajectory of the IMU using kinematic equations and updating this trajectory with landmark positions obtained from visual feature observations. This dual process not only aids in precise localization but also in the detailed mapping of surrounding landmarks, which are crucial for navigation and decision-making in autonomous systems.

The significance of VI-SLAM extends beyond academic interest, finding practical applications in various domains, including autonomous vehicles, drone navigation, and robotic exploration, where accurate and reliable environmental understanding is crucial. Our work aims to explore the potentials of EKF in VI-SLAM to contribute to the ongoing efforts in improving robotic autonomy and environmental interaction. This report presents an overview of our approach, from problem formulation to the technical methodology and the results achieved, providing insights into the challenges faced and the solutions adopted in pursuit of efficient visual-inertial SLAM.

II. Problem Formulation

In our scenario, we focus on the problem of localizing a vehicle equipped with a stereo RGB camera and Inertial Measurement Unit (IMU) while concurrently mapping the environment. Specifically, we are tasked with estimating the state vector $x_{0:t}$ representing the vehicle's pose up to a time t, based on the control inputs $u_{0:t-1}$ and observations $z_{0:t}$. Additionally, we aim to construct a spatial map $m \in \mathbb{R}^{3\times M}$ detailing M landmarks.

A. SLAM with Bayes Filter

To achieve the goal of estimating position of landmarks and the pose of the robot simultaneously, the state transition of the vehicle is conceptualized as a Markov chain, where the state x_t at each timestep is a function of its previous state x_{t-1}, the control input u_t, and the observation z_t. The Bayes Filter maintains two distinct probabilities:

Prediction Probability: This is the forecasted probability distribution for the future state, expressed as $p_{t+1|t}(x_{t+1}) = p(x_{t+1}|z_t, u_t)$, which predicts the next state based on the current observations and controls.

Update Probability: After the prediction, this updated probability distribution incorporates the latest observations, formulated as $p_{t+1|t+1}(x_{t+1}) = p(x_{t+1}|z_{t+1}, u_t)$, effectively refining the prediction with new evidence.

This process leverages the Markov assumption combined with the principles of conditional probability. The prediction phase uses information from the vehicle's motion model to estimate its position at time t, and the update phase then adjusts this estimate to align with the latest observed data.

B. Visual Mapping

The mapping component aims to determine the spatial coordinates $m \in \mathbb{R}^{3M}$ of landmarks from the visual observations $z_t \in \mathbb{R}^{4N_t}$, where N_t indicates the count of visible landmarks at time t. With a pre-established correspondence between landmarks and their visual detections, the mapping challenge narrows down to maximizing the observational probability $p(z_t|x_t, m)$ for landmark localization.

C. Visual-Inertial Odometry

Incorporating both the visual and inertial measurements, the vehicle's pose $T_t \in SE(3)$ is to be estimated. This is achieved through an EKF that combines the motion model leveraging IMU data (linear v_t and angular ω_t velocities) with the observation model using visual data. The odometry estimation feeds into the SLAM process, enabling a comprehensive and robust mapping and localization system.

III. Technical Approach

A. EKF SLAM

The Extended Kalman Filter (EKF) is an extension of the Kalman Filter for nonlinear systems. It linearizes about the current mean and covariance to provide a Gaussian approximation to the true posterior distribution. The EKF consists of two steps: prediction and update. In the prediction step, the EKF propagates the state and covariance forward in time. In the update step, the filter incorporates the new measurement to refine the state estimate.

The prior distribution of the state x_t given all past measurements $z_{0:t}$ and controls $u_{0:t-1}$ is assumed to be Gaussian with mean $\mu_{t|t}$ and covariance $\Sigma_{t|t}$:

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

The motion model predicts the next state x_{t+1} using the current state x_t, control input u_t, and process noise w_t which follows a zero-mean Gaussian distribution with covariance W:

$$\begin{aligned} x_{t+1} &= f(x_t, u_t, w_t), & w_t &\sim \mathcal{N}(0, W) \\ F_t &:= \frac{\partial f}{\partial x}(\mu_{t|t}, u_t, 0), & Q_t &:= \frac{\partial f}{\partial w}(\mu_{t|t}, u_t, 0) \end{aligned}$$

The observation model relates the predicted state x_t to the measurement z_t with measurement noise v_t, which is also assumed to follow a zero-mean Gaussian distribution with covariance V:

$$\begin{aligned} z_t &= h(x_t, v_t), & v_t &\sim \mathcal{N}(0, V) \\ H_t &:= \frac{\partial h}{\partial x}(\mu_{t|t-1}, 0), & R_t &:= \frac{\partial h}{\partial v}(\mu_{t|t-1}, 0) \end{aligned}$$

The prediction equations project the state and covariance estimates from time t to $t+1$ without yet incorporating the measurement z_{t+1}:

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

The update equations adjust the predicted state $\mu_{t+1|t}$ and covariance $\Sigma_{t+1|t}$ using the Kalman gain K_{t+1} and the discrepancy between the actual measurement z_{t+1} and the predicted measurement $h(\mu_{t+1|t}, 0)$:

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

The Kalman gain K_{t+1} is computed to minimize the a posteriori error covariance and is given by:

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

B. Visual Mapping

Visual mapping involves estimating the positions of landmarks in the environment using observations from a visual sensor. The Extended Kalman Filter (EKF) facilitates this by providing a statistical framework for integrating new observations and updating the estimates of landmark positions.

Firstly, consider the mapping-only problem where the IMU pose $T_t := {}_wT_{l,t} \in SE(3)$ is known. The objective is to use observations z_t to estimate the landmark coordinates m. It is assumed that the landmarks are static, meaning they do not move over time, thus a motion model for m is not required.

Data association Δ_t matches the observed features to known landmarks. This can be provided by an external algorithm.

The observation model includes measurement noise $v_{t,i} \sim \mathcal{N}(0, V)$ and is given by the equation:

$$\mathbf{z}_{t,i} = h\left(T_t, \mathbf{m}_j\right) + \mathbf{v}_{t,i} := K_s \pi \left({}_oT_l T_t^{-1} \underline{\mathbf{m}}_j \right) + \mathbf{v}_{t,i}$$

K_s is the calibration matrix for the stereo camera:

$$K_s = \begin{bmatrix} f_u & 0 & c_u & 0 \\ 0 & f_v & c_v & 0 \\ f_u & 0 & c_u & -f_u b \\ 0 & f_v & c_v & 0 \end{bmatrix}$$

Landmark positions are expressed in homogeneous coordinates and projected onto the image plane using a projection function π:

$$\underline{\mathbf{m}}_j := \begin{bmatrix} \mathbf{m}_j \\ 1 \end{bmatrix}, \quad \pi(q) := \frac{1}{q_3} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ 1 \end{bmatrix}$$

The derivative of π is crucial for the Jacobian calculations in the EKF update:

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

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

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

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

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

The the EKF update procedures are as follow:

$$\begin{aligned} K_{t+1} &= \Sigma_t H_{t+1}^\top \left(H_{t+1} \Sigma_t H_{t+1}^\top + I \otimes V \right)^{-1} \\ \mu_{t+1} &= \mu_t + K_{t+1} \left(z_{t+1} - \hat{z}_{t+1} \right) \\ \Sigma_{t+1} &= \left(I - K_{t+1} H_{t+1} \right) \Sigma_t \end{aligned}$$

Note that $I \otimes V$ represents a block diagonal matrix with V along the diagonal, which accounts for the measurement noise in the update step.

C. Visual-Inertial Odometry

For the task of visual odometry, our objective is to characterize the pose using a Gaussian distribution. Considering $T \in SE(3)$, our approach involves introducing a perturbation to the pose. We propose that $T_t|z_{0:t}, u_{0:t-1}$ follows a distribution $\mathcal{N}(\mu_{t|t}, \Sigma_{t|t})$, where $\mu_{t|t}$ is an element of $SE(3)$ and $\Sigma_{t|t}$ is a member of $\mathbb{R}^{6\times6}$. Consequently, $T_t = \mu_{t|t}\exp(\hat{\delta\mu_{t|t}})$, where the perturbation $\delta\mu_{t|t}$ adheres to $\mathcal{N}(0, \Sigma_{t|t})$.

The motion model governing the vehicle, along with the prediction phase, given the control input $u_t = \begin{bmatrix} v_t \\ \omega_t \end{bmatrix} \in \mathbb{R}^6$, is articulated as:

$$\mu_{t+1|t} = \mu_{t|t}\exp\left(\tau_t \hat{\mathbf{u}}_t\right)$$

The covariance matrix update with noise W is described as:

$$\Sigma_{t+1|t} = \exp\left(-\tau \mathbf{u}_t^{\curlywedge}\right)\Sigma_{t|t}\exp\left(-\tau \mathbf{u}_t^{\curlywedge}\right)^T + W$$

where

$$\mathbf{u}_t = \begin{bmatrix} \mathbf{v}_t \\ \boldsymbol{\omega}_t \end{bmatrix} \in \mathbb{R}^6$$

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

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

The observational model parallels the one employed in the visual mapping stage, but here, the vehicle's pose is denoted by $\mu_{t+1|t}$. Therefore,

$$\tilde{\mathbf{z}}_{t+1,i} = K_s\pi(oT_l\mu_{t+1|t}^{-1}\underline{\mathbf{m}}_j)$$

In a manner akin to the visual mapping correction step, the pose update equations are formulated as follows:

$$\begin{aligned} K_{t+1} &= \Sigma_{t+1|t}H_{t+1}^\top(H_{t+1}\Sigma_{t+1|t}H_{t+1}^\top + I)^{-1} \\ \mu_{t+1|t+1} &= \mu_{t+1|t}\exp((K_{t+1}(z_{t+1} - \hat{z}_{t+1}))) \\ \Sigma_{t+1|t+1} &= (I - K_{t+1}H_{t+1})\Sigma_{t+1|t} \end{aligned}$$

The matrix H, representing the Jacobian of the observational model with respect to $\mu_{t+1|t}$, is expressed as:

$$H_{t+1,i} = -K_s\frac{\partial\pi}{\partial q}(oT_l\mu_{t+1|t}^{-1}\underline{\mathbf{m}}_j)oT_l(\mu_{t+1|t}^{-1}\underline{\mathbf{m}}_j)^{\odot}$$

Here, the composition operator $\circ$ is defined as:

$$\begin{bmatrix} s \\ 1 \end{bmatrix}^{\odot} = \begin{bmatrix} I & -\hat{s} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times6}$$

D. Visual-Inertial SLAM

To concurrently estimate the positions of landmarks and the robot's pose, we propose integrating the prediction and update stages of both Extended Kalman Filter (EKF)-based visual mapping and visual-inertial odometry. Initially, we define the joint estimated state and covariance assuming Gaussian distribution as:

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

$$\Sigma \in \mathbb{R}^{(6+3M)\times(6+3M)}$$

Here, μ_m represents the estimated landmark positions, and μ_p denotes the estimated six degrees of freedom of the inverse IMU pose as discussed earlier.

The predict step of the Extended Kalman Filter on the joint estimated state and covariance is derived solely from the visual-inertial odometry predict step, assuming all landmarks are static. The equations considering the IMU measurement u_t are formulated as follows:

$$\boldsymbol{\mu}_{t+1|t} = \begin{bmatrix} \boldsymbol{\mu}_{m,t+1|t} \\ \exp\left(-\tau\hat{\mathbf{u}}_t\right)\boldsymbol{\mu}_{p,t|t} \end{bmatrix}$$

$$\Sigma_{t+1|t} = F_t\Sigma_{t|t}F_t^\top + W$$

$$F_t = \begin{bmatrix} I & 0 \\ 0 & \exp(-\tau u_t^{\curlywedge}) \end{bmatrix}$$

The update step integrates both the visual mapping and visual-inertial odometry update processes. The equations for the update step are provided below:

Predicted observations:

$$\tilde{\mathbf{z}}_{t+1,i} = K_s\pi(oT_l\mu_{t+1|t}^{-1}\underline{\mathbf{m}}_j)$$

Observation matrix:

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

where H_m and H_p represent the mapping process and visual odometry process, respectively, as discussed earlier.

IV. Results

In this chapter, we will present the results of all processes involved in the SLAM procedure, including the IMU trajectory, the landmark mapping results and the EKF-SLAM results with different settings. We alse meticulously compare and analyze all the results to obtain a deeper understanding of Visual-inertial SLAM system. Furthermore, dynamic demonstrations of the map construction are provided, with the results accessible via provided links [1].

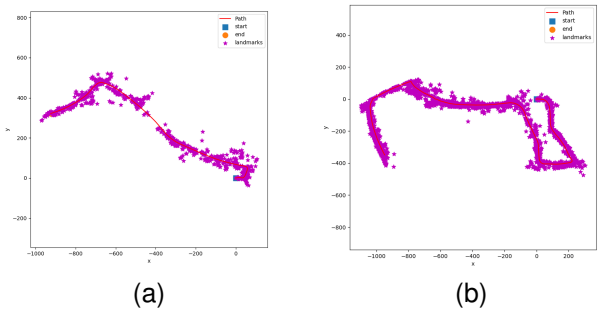

Fig. 1. The robot trajectory predicted by SE3 Kinematics and landmark mapping(noise covariance=2.0) results. The left picture is generated with the dataset 03 and the right one is from the dataset 10.

A. IMU prediction and Landmark Mapping

In this part, we use SE3 Kinematics to predict the robot pose and assume that the predicted IMU trajectory is correct. The results are shown in the Fig. 1. Initially, the system's ability to estimate the robot's trajectory using only IMU data was examined. The trajectory generated from the IMU measurements alone appeared very smooth, indicating a consistent and steady estimation of movement over time without the direct influence of visual data for correction. This smoothness is a testament to the reliability of the IMU in capturing the continuous motion of the robot. The subsequent phase involved mapping the environment's landmarks using the previously estimated IMU trajectory. In this process, the positions of visual landmarks were estimated and observed to change over time as the system incorporated more data. This dynamism in landmark estimation underscores the system's adaptive nature, gradually refining its understanding of the environment with continuous input. The overall results from this stage were positive, demonstrating the system's capability to map out landmarks relative to the robot's trajectory effectively. The validation of the system's performance and the accuracy of the trajectory and landmark mapping was conducted through the analysis of video recordings of the robot's navigation. It confirmed that the system could achieve a coherent and consistent mapping of the environment.

B. EKF-SLAM Results

In the second part of our analysis, we explore the outcomes of the complete visual-inertial SLAM system by integrating IMU predictions with landmark updates and robot pose corrections. First, we did multiple tests on the adjustment of the prediction error covariance and finally set the parameter as 0.00001, which generates the most smooth looking trajectory. Then we examined the effects of varying observation noise levels on pose updates, the results are demonstrated in the Fig. 2,. With a higher assumed observation noise, the system tends to favor IMU predictions, resulting in a trajectory closely resembling that obtained solely from IMU data. Conversely, a lower noise assumption leads to significant pose adjustments based on visual observations, altering the localization result noticeably. Particularly in the dataset 03, discrepancies were most evident in sections with scarce reliable visual landmarks,

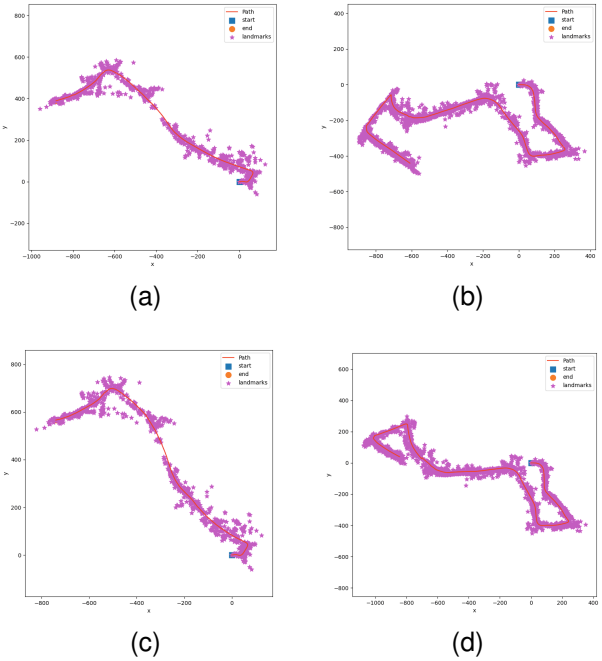

Fig. 2. The robot trajectory generated with the complete EKF SLAM with different parameter settings. The left pictures are generated with the dataset 03 and the right pictures is from the dataset 10. The upper pictures show the results with large observation noise assumption(noise covariance=2.0) while the bottom pictures assume a small noise(noise covariance=0.1).

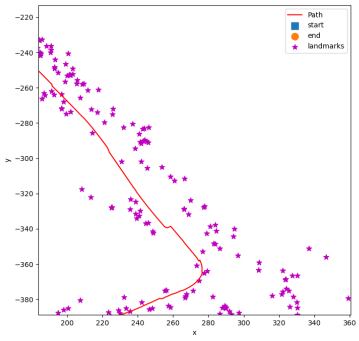

Fig. 3. Trajectory jumping generated during the SLAM process.

leading to drifts. A closer inspection of the trajectory revealed that visual localization introduces jumps and jaggedness, as is shown in Fig. 3. Additionally, it's crucial that during the SLAM process, we filtered out landmarks that deviated significantly from previous predictions. Without implementing this step, numerous mismatches would lead to localization errors. All these findings suggest that the observation model require elaborate design to enhance the system's performance.

V. Conclusion

In conclusion, this project provided a practical exploration into visual-inertial SLAM, leveraging an EKF framework. While it demonstrated the basic integration of visual and inertial sensors for SLAM, it also highlighted the complexities and challenges inherent to this approach. The results, although

promising, point towards areas for improvement, especially in computational efficiency and landmark estimation accuracy. This work serves as a foundational step, underscoring the need for further research and optimization to enhance the capabilities of SLAM systems in real-world applications.

REFERENCES

[1] Mazeyu Ji. The localization and mapping results [Video]. https://github.com/jimazeyu/img_lib/tree/main/EKF_SLAM.