Visual-Inertial Simultaneous Localization and Mapping (SLAM) via the Extended Kalman Filter (EKF)

1st Kai Chuen Tan
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, United States
kctan@ucsd.edu, A59011493

I. Introduction

Navigating around an unexplored and unfamiliar environment is not only a challenging task for autonomous vehicles and mobile robots but also a crucial task that enables humans to explore uncharted hazardous territories without the need of risking any human's life. One of the most common hardware that autonomous vehicles and mobile robots used to navigate in outdoor environments is the global positioning system (GPS). The development of GPS navigation technology for autonomous vehicles including unmanned aerial vehicles (UAV), unmanned ground vehicle (UGV), and other automobile robotics has been matured over the years [1]; with the real-time kinematic (RTK) GPS and the global navigation satellite system (GNSS) on an automobile robot, the centimeter-level positioning accuracy can be achieved [2]. In this day and age, most of the UAVs' navigation system uses GPS to navigate safely in the sky; for example, a private drone delivery company in the United States, Zipline, integrated a custom-built GPS navigation system into their UAV system to perform safe flight operation and accurate medical supplies deliveries like blood platelets [3].

Nevertheless, in GPS-denied environments like indoor environments and urban canyons, the navigation tasks difficulty for autonomous vehicles and automobile robots increases. Autonomous vehicles and automobile robots that heavily rely on the GPS as their navigation system will not be able to navigate around safely in GPS-denied environments. If autonomous UAVs and UGVs are able to navigate in a hazardous indoor environment, UGVs and UAVs will be able to perform search-and-rescue missions and safe autonomous flights to save lives. In the previous sensing and estimation robotics project, the particle filter in simultaneously localization and mapping (SLAM) with the differential-drive motion and scan-grid correlation observation model is implemented to localize a vehicle and to map its surrounding environment simultaneously using the vehicle sensors including encoders, fiber optic gyro (FOG), 2-D light detection and ranging (LIDAR), and an RGB stereo camera to collect data that helps the vehicle to be aware of its surrounding. However, the price of a LIDAR sensor is approximately USD 500, which is considered expensive; on the other hand, eight cameras that are equipped by one Tesla's self-driving car only cost less than USD 100, which is cheaper than a LIDAR sensor [4]. Therefore, in this project, the visual-inertial simultaneous localization and mapping (SLAM) via an extended Kalman filter (EFK) is implemented to localize a vehicle and to map its surrounding environment simultaneously by using only two vehicle sensors, which are an inertial measurement unit (IMU) sensor and a stereo camera to collect data that helps the vehicle to be aware of its surrounding.

The visual-inertial SLAM via the EKF algorithm is divided into three main steps. First, to estimate the pose of a vehicle at a future time with the linear and angular velocities measurements from the IMU sensor, the IMU localization via EKF prediction step is implemented. Then, the landmark mapping via the EKF update step function is also implemented to estimate and update the landmark positions of visual features observed in the images in the world frame. Finally, a complete visual-inertial SLAM is performed by implementing the visual-inertial SLAM algorithm that consists of both the EKF prediction step and EKF update step to plot and estimate both the vehicle trajectory and landmark positions in the world on a map. The paper is organized as follows. §II presents the problem formulation. §III describes the technical approach to the IMU localization via the EKF prediction step, the landmark mapping via the EKF update step, and the visual-inertial SLAM. Lastly, §IV presents the results of the IMU localization via the EKF prediction step, the results of the landmark mapping via the EKF update step, the results of the visual-inertial SLAM, and the discussion of the results.

II. Problem Formulation

A. Simultaneous Localization and Mapping (SLAM)

The main objective of the visual-inertial simultaneous localization and mapping (SLAM) via the extended Kalman filter is to localize a vehicle using the IMU sensor and map the environment with feature observations from a vehicle stereo camera. Therefore, in the SLAM problem, several Markov assumptions can be made given that the sequences of control

inputs $\mathbf{u}_{0:T}$ and observations $\mathbf{z}_{0:T}$ are known and observed, respectively, and the sequences of the vehicle state $\mathbf{x}_{0:T}$ and map states $\mathbf{m}_{0:T}$ are unknown and hidden, respectively, where $t \in \mathbb{N}_0$, is the discrete-time steps in nanoseconds, $\forall t = 0, 1, 2, 3, \cdots, T$, where T is the end of the time in nanoseconds. The Markov assumptions are listed in the following [5]:

- The vehicle state, $\mathbf{x}_{t+1}$ only depends on the previous input $\mathbf{u}_t$ and $\mathbf{x}_t$.
- The map state, $\mathbf{m}_{t+1}$ only depends on the previous map state, $\mathbf{m}_t$.
- The map state, $\mathbf{m}_t$, and the vehicle state, $\mathbf{x}_t$ may affect each other's motion.
- The observation $\mathbf{z}_t$ only depends on the vehicle state, $\mathbf{x}_t$, and the map state, $\mathbf{m}_t$.

The motion model of the vehicle can be formulated as follows:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t) \sim p_f(\cdot|\mathbf{x}_t, \mathbf{u}_t) \tag{1}$$

where f is a non-linear function of the motion model, p_f is the probability density function of the motion model that describes the vehicle motion to a new state, $\mathbf{x}_{t+1}$, after the control input, $\mathbf{u}_t$ is applied at state $\mathbf{x}_t$, and $\mathbf{w}_t$ is the motion noise.

Besides the motion model of the vehicle, there is an observation model since the vehicle map the environment with feature observations from a vehicle stereo camera, and the observation model is formulated as follows:

$$\mathbf{z}_t = h(\mathbf{x}_t, \mathbf{m}_t, \mathbf{v}_t) \sim p_h(\cdot|\mathbf{x}_t, \mathbf{m}_t) \tag{2}$$

where h is the function of the observation model, p_h is the probability density function of the observation model that describes the vehicle observation, $\mathbf{z}_t$ depending on $\mathbf{x}_t$, and $\mathbf{m}_t$, and $\mathbf{v}_t$ is the observation noise.

SLAM is a parameter estimation problem to determine the environment map, $\mathbf{m}$ and the vehicle poses, $\mathbf{x}_t$ given a dataset of the vehicle inputs $\mathbf{u}_{0:T-1}$ and observations $\mathbf{z}_{0:T}$. Hence, the SLAM problem can be formulated in a probabilistic form as shown below:

$$p(\mathbf{m}, \mathbf{x}_t|\mathbf{z}_{0:T}, \mathbf{u}_{0:T-1}) \tag{3}$$

The objectives of the SLAM problem also exploit the decomposition of the joint probability density function without considering the control policy term due to the Markov assumptions listed above as shown in the following:

$$p(\mathbf{x}_{0:T}, \mathbf{m}, \mathbf{z}_{0:T}, \mathbf{u}_{0:T-1}) = p_{0|0}(\mathbf{x}_0, \mathbf{m}) \prod_{t=0}^{T} p_h(\mathbf{z}_t|\mathbf{x}_t, \mathbf{m}_t) \prod_{t=1}^{T} p_f(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{u}_{t-1}) \tag{4}$$

In this project, the maximum likelihood estimation (MLE) approach is applied to determine the optimal environment map, $\mathbf{m}$ and the vehicle poses, $\mathbf{x}_{0:T}$ as shown in the problem formulation below:

$$\textbf{MLE}: \max_{\mathbf{x}_{0:T}, \mathbf{m}} \sum_{t=0}^{T} \log p_h(\mathbf{z}_t|\mathbf{x}_t, \mathbf{m}) + \sum_{t=1}^{T} \log p_f(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{u}_{t-1}) \tag{5}$$

1) Bayes Filtering

Bayes filtering is a probabilistic inference technique for estimating the state, $\mathbf{x}_t$ of the dynamical system of the vehicle that combines evidence from control inputs and observations using the Markov assumptions as listed above and Bayes rule [5]. The Bayes filtering algorithm requires two main steps, which are the prediction and update steps, to keep track of the updated probability density function (pdf), and the predicted pdf over a discrete state set X as defined below:

$$\textbf{Updated pdf}: p_{t|t}(\mathbf{x}_t) = p(\mathbf{x}_t|\mathbf{z}_{0:T}, \mathbf{u}_{0:T-1}) \tag{6}$$

$$\textbf{Predicted pdf}: p_{t+1|t}(\mathbf{x}_{t+1}) = p(\mathbf{x}_{t+1}|\mathbf{z}_{0:T}, \mathbf{u}_{0:T}) \tag{7}$$

To compute the predicted pdf $p_{t+1|t}$ over $\mathbf{x}_{t+1}$, the vehicle motion model pdf p_f is applied with the given prior pdf $p_{t|t}$ over $\mathbf{x}_t$ and the control input $\mathbf{u}_t$ as formulated below:

$$p_{t+1|t}(\mathbf{x}) = \int p_f(\mathbf{x}|\mathbf{s}, \mathbf{u}_t) p_{t|t}(\mathbf{s}) d\mathbf{s} \tag{8}$$

To obtain the updated pdf $p_{t+1|t+1}$ over $\mathbf{x}_{t+1}$ by using the vehicle observation model pdf, p_h with the given predicted pdf $p_{t+1|t}$ over $\mathbf{x}_{t+1}$ and the measurement $\mathbf{z}_{t+1}$ as formulated below:

$$p_{t+1|t+1}(\mathbf{x}) = \frac{p_h(\mathbf{z}_{t+1}|\mathbf{x}) p_{t+1|t}(\mathbf{x})}{\int p_h(\mathbf{z}_{t+1}|\mathbf{s}) p_{t+1|t}(\mathbf{s}) d\mathbf{s}} \tag{9}$$

where $\mathbf{s} \in X$.

2) Landmark Mapping

For landmark mapping problem, the poses of the vehicle over time, $\mathbf{x}_t$ are assumed to be known, and the landmarks $\mathbf{m}$ are assumed to be static. Given the observations over time, $\mathbf{z}_t \in \mathbb{R}^{4 \times N_t}, \forall t = 0, \cdots, T$, the main objective of the landmark mapping is to estimate the coordinates of the landmarks, $\mathbf{m} \in \mathbb{R}^{3 \times M}$ that generated them, where M is the total number of static landmarks, N_t is the number of observed landmarks at time t, $\mathbf{m}_i \in \mathbb{R}^3$ is the landmark 3D-coordinate in a 3D-space, $\mathbf{z}_{t,i} \in R^4$ is the i^{th} observation. $\mathbf{z}_t$ is a general notation for a composed observation from multiple landmarks since the vehicle can observe multiple landmarks at a single time step, t. The landmark mapping can be formulated as follows:

$$p(\mathbf{m}|\mathbf{z}_{0:T}, \mathbf{x}_{0:T}) \tag{10}$$

B. Vehicle Sensors Configuration and Parameters

The vehicle is equipped with an IMU sensor and a stereo camera to tackle the visual-inertial SLAM problem by using the IMU sensor's linear and angular velocity measurements to estimate the pose of the vehicle and utilizing the stereo camera feature observations to map the environment. The IMU control inputs over time, $\mathbf{u}_t$ contain linear velocity, $v_t \in \mathbb{R}^3$ and angular velocity, $\omega_t \in \mathbb{R}^3$ in the IMU sensor's body frame as presented below:

$$\mathbf{u}_t = \begin{bmatrix} \mathbf{v}_t \\ \boldsymbol{\omega}_t \end{bmatrix} \in \mathbb{R}^{2\times 3} \tag{11}$$

The stereo camera data association, $\Delta_t : \{1, \cdots, M\} \rightarrow \{1, \cdots, N_t\}$ stipulating that landmark j corresponds to stereo camera observation, $\mathbf{z}_{t,i}$ with $i = \Delta_t(j)$ at time t is pre-computed by an external algorithm, where $\mathbf{z}_{t,i}$ contains the left and right camera images pixel coordinates of landmark i. The pixel coordinates of detected features with the correspondences between the left and right camera frames are denoted as $\mathbf{z}_t \in \mathbb{R}^{4\times M}$ as shown in Figure 1. If the landmarks $\mathbf{m}_i$ are not observable at time t, the observation $\mathbf{z}_{t,i}$ will be equal to $[-1, -1, -1, -1]^\top$.

Fig. 1: Visual features matched across the left-right camera frames (left) and across time (right)

In this project, the stereo camera extrinsic calibration is the transformation matrix from the left camera frame to the IMU frame, which denoted as ${}_IT_C \in SE(3)$, and the stereo camera intrinsic calibration matrix, K_s are known. The stereo camera intrinsic calibration matrix, K_s is defined as follows:

$$K_s = \begin{bmatrix} fs_u & 0 & c_u & 0 \\ 0 & fs_v & c_b & 0 \\ fs_u & 0 & c_u & -fs_ub \\ 0 & fs_v & c_b & 0 \end{bmatrix} \tag{12}$$

where f is the focal length in meters, b is the stereo baseline in meters, s_u, s_v are pixel scaling in pixels per meter, and c_u, c_v are the principal point in pixels.

III. Technical Approach

A. Extended Kalman Filter

The Kalman filter is an estimation algorithm that produces estimates of hidden variables based on inaccurate and uncertain measurements and predicts the future system state based on the past estimation [6]. The extended Kalman filter (EKF) is a non-linear version of the Kalman filter that applies the moment matching method to linearize about an estimate of the current mean and covariance. The non-linear Kalman filter is a Bayes filter from §II with the following assumptions [7]:

- The prior pdf $p_{t|t}$ is Gaussian.
- The motion model is with Gaussian noise $\mathbf{w}_t$.
- The observation model is with Gaussian noise $\mathbf{v}_t$.
- The motion noise $\mathbf{w}_t$ and observation noise $\mathbf{v}_t$ are independent of each other, of the state $\mathbf{x}_t$, and across time.
- The predicted and updated pdfs are forced to be Gaussian via approximation.

The main challenge of the non-linear Kalman filter is that the predicted and updated pdfs are not Gaussian; hence, it can no longer be evaluated in closed form. To tackle this challenge, the moment matching approach is used to force the predicted and updated pdfs to be Gaussian by evaluating their first and second moments and approximating them with Gaussians with the same moments. The EKF uses a first-order Taylor series approximation to the motion model, $f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t)$ and the observation model, $h(\mathbf{x}_{t+1}, \mathbf{v}_{t+1})$, around the state and noise means as shown in the following:

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

$$f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t) \approx f(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, 0) + \mathbf{F}_t(\mathbf{x}_t - \boldsymbol{\mu}_{t|t}) + \mathbf{Q}_t\mathbf{w}_t \tag{13}$$

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

$$h(\mathbf{x}_{t+1}, \mathbf{v}_{t+1}) \approx h(\boldsymbol{\mu}_{t+1|t}, 0) + \mathbf{H}_{t+1}(\mathbf{x}_{t+1} - \boldsymbol{\mu}_{t+1|t}) + \mathbf{R}_{t+1}\mathbf{v}_{t+1} \tag{14}$$

Then, the EKF models can be formulated from the Equation 13 and Equation 14 as presented below:

Prior:

$$\mathbf{x}_t | \mathbf{z}_{0:T}, \mathbf{u}_{0:T-1} \sim \mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}) \tag{15}$$

Motion Model:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t), \quad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}) \tag{16}$$

$$\mathbf{F}_t = \frac{df}{d\mathbf{x}}(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, 0) \tag{17}$$

$$\mathbf{Q}_t = \frac{df}{d\mathbf{w}}(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, 0) \tag{18}$$

Observation Model:

$$\mathbf{z}_t = h(\mathbf{x}_t, \mathbf{v}_t), \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \tag{19}$$

$$\mathbf{H}_t = \frac{dh}{d\mathbf{x}}(\boldsymbol{\mu}_{t|t-1}, 0) \tag{20}$$

$$\mathbf{R}_t = \frac{dh}{d\mathbf{v}}(\boldsymbol{\mu}_{t|t-1}, 0) \tag{21}$$

EKF Prediction Step:

$$\boldsymbol{\mu}_{t+1|t} = f(\boldsymbol{\mu}_{t|t}, \mathbf{u}_t, 0) \tag{22}$$

$$\boldsymbol{\Sigma}_{t+1|t} = \mathbf{F}_t \boldsymbol{\Sigma}_{t|t} \mathbf{F}_t^\top + \mathbf{Q}_t \mathbf{W} \mathbf{Q}^\top \tag{23}$$

EKF Update Step:

$$\boldsymbol{\mu}_{t+1|t+1} = \boldsymbol{\mu}_{t+1|t} + \mathbf{K}_{t+1|t}(\mathbf{z}_{t+1} - h(\boldsymbol{\mu}_{t+1|t}, 0)) \tag{24}$$

$$\boldsymbol{\Sigma}_{t+1|t+1} = (I - \mathbf{K}_{t+1|t}\mathbf{H}_{t+1})\boldsymbol{\Sigma}_{t+1|t} \tag{25}$$

Kalman Gain:

$$\mathbf{K}_{t+1|t} = \boldsymbol{\Sigma}_{t+1|t}\mathbf{H}_{t+1}^\top (\mathbf{H}_{t+1}\boldsymbol{\Sigma}_{t+1|t}\mathbf{H}_{t+1}^\top + \mathbf{R}_{t+1}\mathbf{V}\mathbf{R}_{t+1}^\top)^{-1} \tag{26}$$

B. IMU Localization via the EKF Prediction

The objective of the IMU localization via the EKF prediction step is to estimate the inverse IMU pose of the vehicle, $\mathbf{T}_t \in SE(3)$ over time given the IMU's linear velocity, $v_t \in \mathbb{R}^3$ and angular velocity, $\omega_t \in \mathbb{R}^3$ measurements, $\mathbf{u}_t$ as shown in the following:

$$\mathbf{T}_t = {}_W\mathbf{T}_{I,t}^{-1} \in SE(3) \tag{27}$$

$$\mathbf{u}_t = \begin{bmatrix} v_t^\top \\ \omega_t^\top \end{bmatrix} \tag{28}$$

To solve the IMU localization problem, first, the Gaussian prior needs to be formulated as shown in the following equations:

$$\mathbf{T}_t | \mathbf{z}_{0:T}, \mathbf{u}_{0:T-1} \sim \mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}) \tag{29}$$

$$\mathbf{T}_t = \boldsymbol{\mu}_{t|t} \exp(\hat{\delta\boldsymbol{\mu}_{t|t}}) \tag{30}$$

$$\delta\boldsymbol{\mu}_{t|t} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{t|t}) \tag{31}$$

where, $\boldsymbol{\mu}_{t|t} \in SE(3)$, and $\boldsymbol{\Sigma}_{t|t}) \in \mathbb{R}^{6\times6}$. The $\boldsymbol{\Sigma}_{t|t})$ is a 6 $\times$ 6 matrix because only the 6 degrees of freedom of $\mathbf{T}_t$ are changing. Then, the EKF prediction step with motion model with the Gaussian noise, $\mathbf{w}_t$, can be formulated as shown below:

$$\mathbf{T}_{t+1|t} = \exp(-\tau(\mathbf{u}_t + \mathbf{w}_t)^\wedge)\mathbf{T}_t, \quad \mathbf{w}_t \sim \mathcal{N}(0, \mathbf{W}) \tag{32}$$

To separate the effect of the noise, $\mathbf{w}_t$ from the motion of the deterministic part of $\mathbf{T}_t$ [7], the EKF motion model can be re-written in terms of nominal kinematics of $\boldsymbol{\mu}_{t|t}$ and perturbation kinematics of $\delta\boldsymbol{\mu}_{t|t}$ with time discretization τ_t as presented below:

$$\boldsymbol{\mu}_{t+1|t} = \exp(-\tau\hat{\mathbf{u}}_t)\boldsymbol{\mu}_{t|t} \tag{33}$$

$$\delta\boldsymbol{\mu}_{t+1|t} = \exp(-\tau\overset{\curlywedge}{\mathbf{u}}_t)\delta\boldsymbol{\mu}_{t|t} + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}(0, \mathbf{W}) \tag{34}$$

where,

$$\hat{\mathbf{u}}_t = \begin{bmatrix} \hat{\omega}_t & v_t \\ \mathbf{0}^\top & 0 \end{bmatrix} \in \mathbb{R}^{4\times4} \tag{35}$$

$$\overset{\curlywedge}{\mathbf{u}}_t = \begin{bmatrix} \hat{\omega}_t & \hat{v}_t \\ \mathbf{0}^\top & \hat{\omega}_t \end{bmatrix} \in \mathbb{R}^{6\times6} \tag{36}$$

where $\hat{v}_t \in \mathbb{R}^3$ is the linear velocity skew-symmetric matrix, and $\hat{\omega}_t \in \mathbb{R}^3$ is the angular velocity skew-symmetric matrix. Next, with Equation 33 and Equation 34 above, the EKF prediction step with motion model with the Gaussian noise, $\mathbf{w}_t$, can be re-formulated as shown below:

$$\boldsymbol{\mu}_{t+1|t} = \boldsymbol{\mu}_{t|t} \exp(\tau\hat{\mathbf{u}}_t) \tag{37}$$

$$\begin{aligned} \boldsymbol{\Sigma}_{t+1|t} &= \mathbb{E}[\delta\boldsymbol{\mu}_{t+1|t}\delta\boldsymbol{\mu}_{t+1|t}^\top] \\ \boldsymbol{\Sigma}_{t+1|t} &= \exp(-\tau\overset{\curlywedge}{\mathbf{u}}_t)\boldsymbol{\Sigma}_{t|t} \exp(-\tau\overset{\curlywedge}{\mathbf{u}}_t)^\top + \mathbf{W} \end{aligned} \tag{38}$$

The Rodrigues' formula is applied to find a closed-form expression for the exponential map from $\mathfrak{so}(3)$ to $SO(3)$ by using the "*expm()*" function from the "*scipy.linalg*" Python package. From the IMU localization via the prediction step, the poses of the IMU in the world frame can be obtained by taking the inverse of all predicted mean, $\mathbf{u}_{t+1|t}$ since all the predicted mean, $\mathbf{u}_{t+1|t}$ are the inverse IMU poses in the IMU frame.

C. Landmark Mapping via the EKF Update

Before solving the landmark mapping problem, there are several assumptions to make. First, the IMU pose, $\mathbf{T}_t$ from Equation 27 is known. Second, the data association $\Delta_t : \{1, \cdots, M\} \rightarrow \{1, \cdots, N_t\}$ stipulating that landmark j corresponds to stereo camera observation, $\mathbf{z}_{t,i}$ with $i = \Delta_t(j)$ at time t is provided by an external algorithm as mentioned in §II. Third, the landmarks $\mathbf{m}$ are static. With these landmark mapping assumptions, the landmark coordinates, $\mathbf{m}$ can be estimated given the observations over time, $\mathbf{z}_t$. The Gaussian prior of the observation model can formulated as shown below:

$$\mathbf{m} | \mathbf{z}_{0:T} \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad \boldsymbol{\mu}_t \in \mathbb{R}^{3M}, \quad \boldsymbol{\Sigma}_t \in \mathbb{R}^{3M\times3M} \tag{39}$$

where $\boldsymbol{\mu}_t$ is the landmark position estimates at time t, and $\boldsymbol{\Sigma}_t$ is the landmark covariance estimate at time t. The EKF observation model with the Gaussian measurement noise can be formulated as follows:

$$\begin{aligned}\mathbf{z}_{t,i} &= h(\mathbf{T}_t, \mathbf{m}_j) + \mathbf{v}_{t,i} \\ \mathbf{z}_{t,i} &= K_s \pi({}_C T_I T_t^{-1} \underline{\mathbf{m}}_j) + \mathbf{v}_{t,i}, \quad \mathbf{v}_{t,i} \sim \mathcal{N}(0, \mathbf{V})\end{aligned} \tag{40}$$

where $\underline{\mathbf{m}}_j$ is the landmark homogeneous coordinate as shown below:

$$\underline{\mathbf{m}}_j = \begin{bmatrix} \mathbf{m}_j \\ 1 \end{bmatrix} \tag{41}$$

The projection function, $\pi(\mathbf{q})$, and the derivative of the projection function, $\pi'(\mathbf{q})$ can be defined as follows:

$$\pi(\mathbf{q}) = \frac{1}{q_3}\mathbf{q} \in \mathbb{R}^4 \tag{42}$$

$$\pi'(\mathbf{q}) = \frac{d\pi}{d\mathbf{q}}(\mathbf{q}) = \frac{1}{q_3}\begin{bmatrix} 1 & 0 & -\frac{q_1}{q_3} & 0 \\ 0 & 1 & -\frac{q_2}{q_3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{q_4}{q_3} & 1 \end{bmatrix} \tag{43}$$

Then, all of the observations, $\mathbf{z}_{t,i}, \forall i = 1, \cdots, N_t$, are stacked as a $4N_t$ vector, at time t, denoted as $\mathbf{z}_t \in \mathbb{R}^{4N_t}$. The EKF observation model can be re-formulated as shown below:

$$\mathbf{z}_t = K_s \pi({}_C T_I T_t^{-1} \underline{\mathbf{m}}) + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I} \otimes \mathbf{V}) \tag{44}$$

$$\mathbf{I} \otimes \mathbf{V} = \begin{bmatrix} \mathbf{V} & & \\ & \ddots & \\ & & \mathbf{V} \end{bmatrix} \tag{45}$$

The coordinates for each landmark, $[x_{o,j}, y_{o,j}, z_{o,j}]^\top \in \mathbb{R}^3$ in the camera optical frame need to be determined with the initial observation by applying the inverse of the stereo camera model and calculating the disparity between the left and right cameras as shown in the following equations:

$$\begin{bmatrix} u_L \\ v_L \\ u_R \\ v_R \end{bmatrix} = K_s \frac{1}{z_{o,j}} \begin{bmatrix} x_{o,j} \\ y_{o,j} \\ z_{o,j} \\ 1 \end{bmatrix} \tag{46}$$

$$\begin{bmatrix} x_{o,j} \\ y_{o,j} \\ z_{o,j} \end{bmatrix} = \begin{bmatrix} \frac{(u_L - c_u)b}{u_L - u_R} \\ \frac{(v_L - c_v)(-fs_u b)}{fs_v(u_L - u_R)} \\ \frac{fs_u b}{u_L - u_R} \end{bmatrix} \tag{47}$$

where, u_L, v_L is the pixel coordinate in the left image, and u_R, v_R is the pixel coordinate in the right image. Then, the predicted observations, $\tilde{\mathbf{z}}_{t+1,i}$, based on the landmark position estimate at time t, $\boldsymbol{\mu}$, and known correspondences Δ_{t+1} is formulated below:

$$\tilde{\mathbf{z}}_{t+1,i} = K_s \pi({}_C T_I T_{t+1}^{-1} \underline{\boldsymbol{\mu}}_{t,j}) \in \mathbb{R}^4, \quad \forall i = 1, \cdots, N_{t+1} \tag{48}$$

Next, the Jacobian of the predicted observation, $\tilde{\mathbf{z}}_{t+1,i}$, denoted as $\mathbf{H}_{t+1,i,j}$, with respect to the current feature $\mathbf{m}_j$ evaluated at its corresponding landmark position, $\underline{\boldsymbol{\mu}}_{t,j}$ can be computed as follows:

$$\mathbf{H}_{t+1,i,j} = \begin{cases} K_s \pi'({}_C T_I T_{t+1}^{-1} \underline{\boldsymbol{\mu}}_{t,j}) {}_C T_I T_{t+1}^{-1} P^\top & \Delta_t(j) = i \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{49}$$

where $P = [\mathbf{I}, \mathbf{0}] \in \mathbb{R}^{3\times 4}$ is the projection matrix. The EKF update step can be then performed as presented below:

$$\mathbf{K}_{t+1|t} = \boldsymbol{\Sigma}_t \mathbf{H}_{t+1}^\top (\mathbf{H}_{t+1} \boldsymbol{\Sigma}_t \mathbf{H}_{t+1}^\top + \mathbf{I} \otimes \mathbf{V})^{-1} \tag{50}$$

$$\boldsymbol{\mu}_{t+1|t+1} = \boldsymbol{\mu}_t + \mathbf{K}_{t+1}(\mathbf{z}_{t+1} - \tilde{\mathbf{z}}_{t+1}) \tag{51}$$

$$\boldsymbol{\Sigma}_{t+1|t+1} = (\mathbf{I} - \mathbf{K}_{t+1}\mathbf{H}_{t+1})\boldsymbol{\Sigma}_t \tag{52}$$

where $\mathbf{K}_{t+1|t}$ is the Kalman gain, $\boldsymbol{\mu}_{t+1|t+1}$ is the updated landmark positions at the world frame, and $\boldsymbol{\Sigma}_{t+1|t+1}$ is the updated landmark covariances.

D. Visual-Inertial Simultaneous Localization and Mapping (SLAM)

To localize the pose of the vehicle and map the landmarks simultaneously based on the IMU sensor and the stereo camera measurements, the IMU prediction step and the landmark update step are combined, and then the IMU update step can be implemented based on the stereo-camera observation model. The joint estimated state, $\boldsymbol{\mu}_{joint}$ and covariance, $\boldsymbol{\Sigma}_{joint}$ between the landmark map and the IMU pose must be defined as shown below since they are dependent:

$$\boldsymbol{\mu}_{joint} = \begin{bmatrix} \boldsymbol{\mu}_l \\ \boldsymbol{\mu}_r \end{bmatrix} \in \mathbb{R}^{(3M+6)} \tag{53}$$

$$\boldsymbol{\Sigma}_{joint} \in \mathbb{R}^{(3M+6)\times(3M+6)} \tag{54}$$

where $\boldsymbol{\mu}_l$ is the predicted landmark position from Equation 39, $\boldsymbol{\mu}_r$ is the estimated inverse IMU pose with six degrees of freedom from Equation 33. The visual-inertial SLAM EKF prediction step with the joint estimated state and covariance can be derived as follows:

$$\boldsymbol{\mu}_{t+1|t,joint} = \begin{bmatrix} \boldsymbol{\mu}_{t+1|t,l} \\ \boldsymbol{\mu}_{t+1|t,r} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_{t+1|t,l} \\ \exp(-\tau \hat{\mathbf{u}}_t) \boldsymbol{\mu}_{t|t,r} \end{bmatrix} \tag{55}$$

$$\boldsymbol{\Sigma}_{t+1|t,joint} = \mathbf{F}_{t,joint} \boldsymbol{\Sigma}_{t|t,joint} \mathbf{F}_{t,joint}^\top + \mathbf{W}_{joint} \tag{56}$$

$$\mathbf{F}_{t,joint} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \exp(-\tau \hat{\mathbf{u}}_t) \end{bmatrix} \tag{57}$$

$$\mathbf{W}_{joint} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} \tag{58}$$

The virtual-inertial SLAM EKF update step is the combination of both landmark mapping update step with the predicted observation, $\tilde{\mathbf{z}}_{t+1,i}$, from Equation 48 and the IMU update step, which can be formulated with the joint Jacobian matrix, $\mathbf{H}_{t+1|t,joint}$ as follows:

Jacobian Matrix:

$$\mathbf{H}_{t+1|t,joint} = [\mathbf{H}_{t+1|t,l}, \mathbf{H}_{t+1|t,r}] \in \mathbb{R}^{(4N_t \times (3M+6))} \tag{59}$$

$$\mathbf{H}_{t+1|t,i,r} = -K_s \pi'({}_C T_I \boldsymbol{\mu}^{-1}_{t+1|t,r} \underline{\mathbf{m}}_j)\, {}_C T_I (\boldsymbol{\mu}^{-1}_{t+1|t,r} \underline{\mathbf{m}}_j)^{\odot} \tag{60}$$

$$\mathbf{H}_{t+1|t,r} = \begin{bmatrix} \mathbf{H}_{t+1|t,1,r} \\ \vdots \\ \mathbf{H}_{t+1|t,N_t,r} \end{bmatrix} \tag{61}$$

$$\underline{\mathbf{m}}_j^{\odot} = \begin{bmatrix} \mathbf{m}_j \\ \lambda \end{bmatrix}^{\odot} = \begin{bmatrix} \mathbf{I} & -\hat{\mathbf{m}}_j \\ \mathbf{0} & 0 \end{bmatrix} \in \mathbb{R}^{4\times 6} \tag{62}$$

where, $\mathbf{H}_{t+1|t,l}$ is from Equation 49.

EKF Update Step:

$$\mathbf{K}_{t+1,joint} = \boldsymbol{\Sigma}_{t+1|t,joint} \mathbf{H}^{\top}_{t+1,joint} (\mathbf{H}_{t+1,joint} \boldsymbol{\Sigma}_{t+1|t,joint} \mathbf{H}^{\top}_{t+1,joint} + \mathbf{I} \otimes \mathbf{V})^{-1} \tag{63}$$

$$\boldsymbol{\mu}_{t+1|t+1,joint} = \boldsymbol{\mu}_{t+1|t,joint} \exp((\mathbf{K}_{t+1,joint}(\mathbf{z}_{t+1} - \tilde{\mathbf{z}}_{t+1}))^{\wedge}) \tag{64}$$

$$\boldsymbol{\Sigma}_{t+1|t+1,joint} = (\mathbf{I} - \mathbf{K}_{t+1,joint} \mathbf{H}_{t+1,joint}) \boldsymbol{\Sigma}_{t|t+1,joint} \tag{65}$$

After the joint estimate mean, $\boldsymbol{\mu}_{t+1|t+1,joint}$, and the joint estimate covariance, $\boldsymbol{\Sigma}_{t+1|t+1,joint}$ are updated, the $\boldsymbol{\mu}_{t+1|t+1,r}$ from $\boldsymbol{\mu}_{t+1|t+1,joint}$ is the inverse IMU pose, and the $\boldsymbol{\mu}_{t+1|t+1,l}$ from $\boldsymbol{\mu}_{t+1|t+1,joint}$ is the landmarks coordinates in the world frame.

IV. Results and Discussion

A. IMU Localization via the EKF Prediction Step

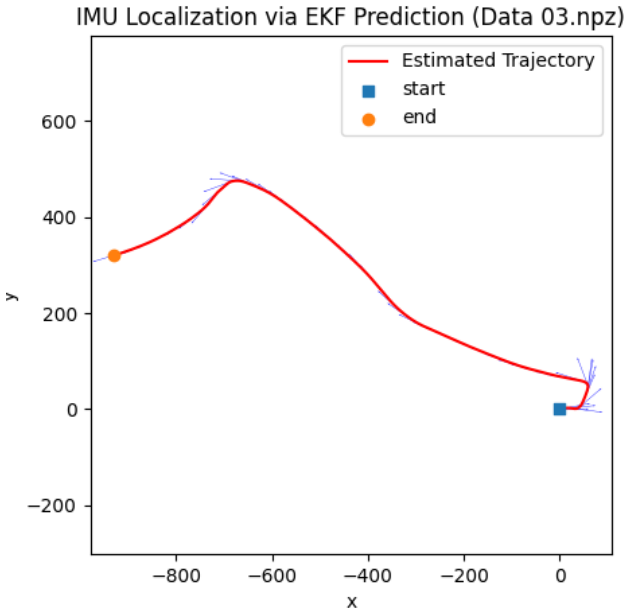

Fig. 2: IMU Localization via EKF Prediction (Data 03.npz)

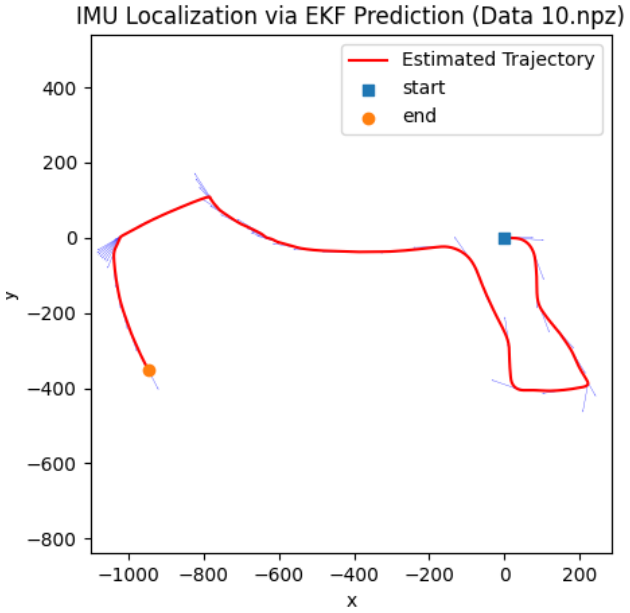

Fig. 3: IMU Localization via EKF Prediction (Data 10.npz)

According to the Figure 2 and Figure 3, the predicted vehicle trajectories using IMU localization via the EKF prediction step seem reasonably accurate, and the predicted vehicle trajectory is validated with each dataset video [8].

B. Landmark Mapping via the EKF Update Step

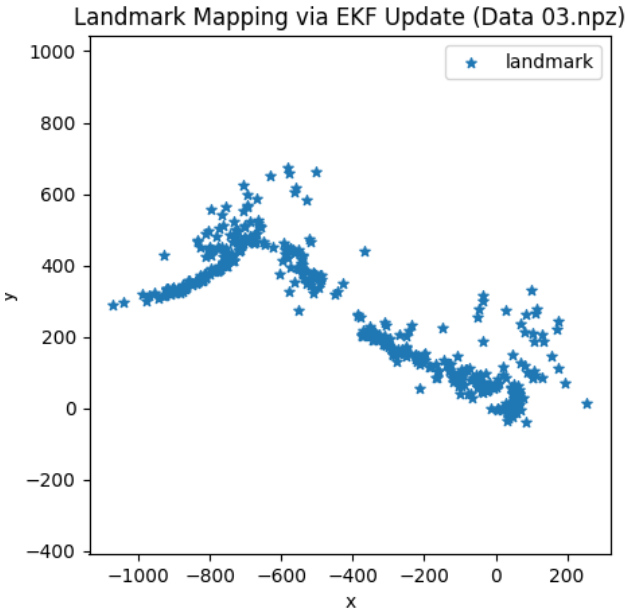

Fig. 4: Landmark Mapping via EKF Update (Data 03.npz)

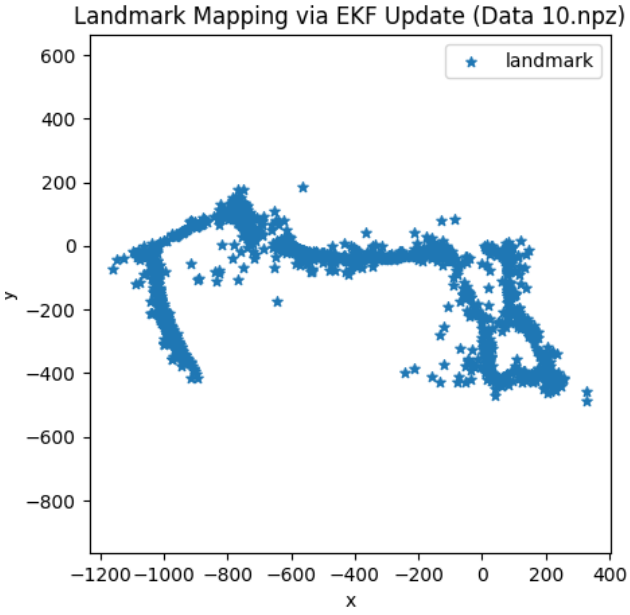

Fig. 5: Landmark Mapping via EKF Update (Data 10.npz)

Figure 4 and Figure 5 illustrate the landmark coordinates in the world frame, and the landmark mapping via the EKF update step is completed by skipping every 5 features to speed up the computational speed. Figure 4 and Figure 5 also show that there are some landmark outliers from the IMU localization via the EKF prediction step because the landmark mapping via the EKF update step does not consider the correlation between the IMU pose and the landmarks pose.

C. Visual-Inertial Simultaneous Localization and Mapping (SLAM)

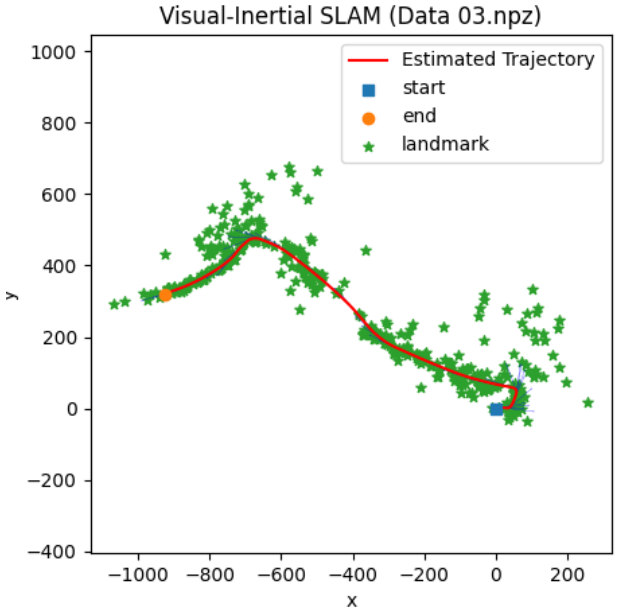

Fig. 6: Visual-Inertial SLAM (Data 03.npz)

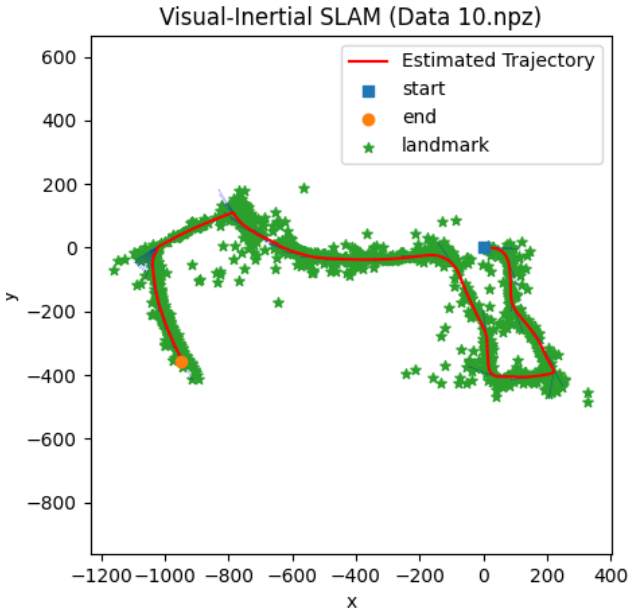

Fig. 7: Visual-Inertial SLAM (Data 10.npz)

Figure 6 and Figure 7 illustrate the predicted and updated vehicle trajectory and landmark positions in the world frame. Every 5 features are skipped in the process to speed up the computational speed for the visual-inertial SLAM; By comparing the visual-inertial SLAM results with the landmark mapping results and the IMU localization results, they are comparatively similar because the motion and observation noises parameters are reasonable and the same. In this project, selecting the optimal noises for both the motion and observation noises is challenging and important because choosing unreasonable high motion and observation noises will output vastly different results. The noises hyper-parameters are presented in Table 1. Furthermore, the visual-inertial SLAM feature selection approach is arbitrary, and the approach depends on knowing the number of features in advance, but in real-life practical cases, the number of features at a time step, t is not known. Hence, instead of skipping every certain number of features along the process, observable features can be selected based on how often the observable features appear.

REFERENCES

[1] Kazunori Ohno, Takashi Tsubouchi, Bunji Shigematsu Shin'ichi Yuta (2004) Differential GPS and odometry-based outdoor navigation of a mobile robot, Advanced Robotics, 18:6, 611-635, DOI: 10.1163/1568553041257431

[2] Ferreira A, Matias B, Almeida J, Silva E. Real-time GNSS precise positioning: RTKLIB for ROS. International Journal of Advanced Robotic Systems. May 2020. doi:10.1177/1729881420904526

[3] Zipline. (n.d.). Technology with Purpose. Zipline. Retrieved February 24, 2022, from https://flyzipline.com/technology/

[4] "Self-driving technology: Lidar vs camera," Vested Finance, 18-Jan-2022. [Online]. Available: https://vested.co.in/blog/self-driving-technology-lidar-vs-camera/. [Accessed: 09-Mar-2022].

[5] Atanasov, N. (2022, February 10). ECE276A: Sensing amp; Estimation in Robotics Lecture 9: Bayesian Filtering. UCSD ECE276A: Sensing amp; Estimation in Robotics (Winter 2022). Retrieved March 11, 2022, from https://natanaso.github.io/ece276a/ref/ECE276A_9_BayesianFiltering.pdf

[6] A. Becker, "Online Kalman Filter Tutorial," Kalman Filter Tutorial. [Online]. Available: https://www.kalmanfilter.net/default.aspx. [Accessed: 11-Mar-2022].

[7] Atanasov, N. (2022, February 10). ECE276A: Sensing amp; Estimation in Robotics Lecture 12: Extended and Unscented Kalman Filtering. UCSD ECE276A: Sensing amp; Estimation in Robotics (Winter 2022). Retrieved March 11, 2022, from https://natanaso.github.io/ece276a/ref/ECE276A_9_BayesianFiltering.pdf

[8] Q. Feng, "PR3 original image sequence videos," Piazza, 07-Mar-2022. [Online]. Available: https://piazza.com/class/kxfgt4eags92p5?cid=381. [Accessed: 12-Mar-2022].

TABLE I: Hyper-parameters for the Visual-Inertial SLAM Algorithm

Hyper-parameter	**Description of the Hyper-parameter**	**Value**
W	6×6 Motion Noise Variance Matrix	**I**
V	$3M \times 3M$ Observation Noise Variance Matrix	100 **I**