Visual-Inertial SLAM: A Project Review

Hao-Yuan Tang
Department of Electrical Engineering
University of California, San Diego
h6tang@ucsd.edu

I. Introduction

In this project, Extended Kalman filter algorithm (EKF) was used to tackle simultaneous localization and mapping (SLAM)problem in visual-inertial sensor configuration. SLAM can be separated into two parts: localization given a surrounding map of the environment for estimating robot trajectory by odometry reading, and mapping surrounding with robot state simultaneously by two stereo cameras.

EKF, which is a form of Bayesian filter, can effectively model a robot's predicted locations in the world as Gaussian distribution based on past input controls and current observations. The EKF was used when we try to model non-linear observation and motion model, which provided more widely applicable scenario than particle filter.

In problem formulation session, we will discuss on essential parts of EKF SLAM in mathematical sense. In technical approach session, detailed implementation of EKF SLAM was given, such as transformation of coordinate systems, car model localization using odometry and surrounding map using stereo cameras.

II. Problem Formulation

A. SLAM

The visual SLAM problem can be separated into localization by odometry and mapping by stereo cameras parts. Markov assumption was given for model simplification, i.e., the robots state $\boldsymbol{x}_{t+1}$ only depends on the previous input $\boldsymbol{u}_t$ and state $\boldsymbol{x}_t$, and the map state $\boldsymbol{m}_{t+1}$ only depends on the previous map state $\boldsymbol{m}_t$. Given robot's pose $\boldsymbol{x}_t$ and control input $\boldsymbol{u}_t$ at discrete time steps t, the pose in next time step can be described as probabilistic function:

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

where $\boldsymbol{w}_t$ is motion noise and this model was called motion model.

For mapping, the robot can sense environment via camera observations at each time step. In a similar way, observation model, which described camera observations within an environment, can be defined in probabilistic:

$$\boldsymbol{z}_{t+1} = h(\boldsymbol{z}_t, \boldsymbol{v}_t) \sim p_f(\cdot\,|\boldsymbol{x}_t, \boldsymbol{m}) \tag{2}$$

where $\boldsymbol{v}_t$ is motion noise and this model was called observation model.

With motion model and observation model, SLAM is the problem to determine the environment and robot poses from observations and control inputs at each time step t. The objective to compute can be written in probabilistic form:

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

B. Bayes Filtering

Bayes filtering is a probabilistic inference technique for estimating the state $\boldsymbol{x}_t$of dynamic systems combining observations and control inputs with Markov assumptions and Bayes rule. The Bayes filter relies on two steps to keep track of $p_{t|t}(\boldsymbol{x}_t)$ and $p_{t|t+1}(\boldsymbol{x}_{t+1})$.

For prediction step, given a prior density of robot state $p_{t|t}(\boldsymbol{x}_t)$ and the control input $\boldsymbol{u}_t$, we use the motion model Eq. (1) to compute the predicted density $p_{t|t+1}(\boldsymbol{x}_{t+1})$ as the following equation:

$$p_{t+1|t}(\boldsymbol{x}_t) = \int p_f(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{u}_t) p_{t|t}(\boldsymbol{s})\, d\boldsymbol{s} \tag{4}$$

For update step, given the predicted density $p_{t+1|t}(\boldsymbol{x}_{t+1})$ and the measurement $\boldsymbol{z}_{t+1}$, we use the observation model Eq. (2) to incorporate the measurement information and obtain the posterior $p_{t+1|t+1}(\boldsymbol{x}_{t+1})$ as following:

$$p_{t+1|t+1}(\boldsymbol{x}) = \frac{p_h(\boldsymbol{z}_{t+1}|\boldsymbol{x}) p_{t+1|t}(\boldsymbol{x})}{\int p_h(\boldsymbol{z}_{t+1}|\boldsymbol{s}) p_{t+1|t}(\boldsymbol{s}) ds} \tag{5}$$

C. Mapping of Landmarks

The landmark-based mapping aims to generate a map of the surrounding environment from sensor observations **z** with the assumption that the poses of the robot x are known. Also, uncertainty of system was modeled by adding noise into observation model.

The environment is represented by *M* static landmarks, and each of them is characterized by its location in the space denoted as $\boldsymbol{m}_i$. These landmarks are considered as points in the 3D space and can be specified by three numerical values where $\boldsymbol{m}_i \in \mathbb{R}^3$ and $\boldsymbol{m} \in \mathbb{R}^{3\times M}$.

The robot can sense the landmarks at each time step t, with observation $\boldsymbol{z}_t$, and since the robot can sense multiple landmarks at a single time step, $\boldsymbol{z}_t$ is a general notation for composed observation from multiple landmarks. The goal of the mapping problem is then to estimate the locations of landmarks based on the pose of robot $\boldsymbol{x}_t$ and the observation $\boldsymbol{z}_t$. Therefore, we can define the observation model:

$$p(\boldsymbol{z}_t|\boldsymbol{x}_t, \boldsymbol{m}, \boldsymbol{n}_t) \tag{6}$$

where $\boldsymbol{n}_t$ is the index map with $|\boldsymbol{n}_t| = N_t$ and N_t is the number of observed landmarks at time *t*.

D. Sensors Configuration

Our proposed solution aims to solve the SLAM problem with observations from an IMU and a stereo camera installed in a car. The IMU observations contain the linear velocity $\boldsymbol{v}_t \in \mathbb{R}^3$ and angular velocity $\boldsymbol{\omega}_t \in \mathbb{R}^3$ in IMU frame.

To collect landmarks, two calibrated stereo camera were install on the top of car and data are pre-computed to extract the visual features and find the correspondence between left and right camera frames across time steps. Fig.1 shows the visual features extracted in this configuration. The visual features at time t is denoted as $\boldsymbol{z}_t \in \mathbb{R}^{4\times M}$, where the i_{th} column contains the pixel coordinates of landmark *i* in the left and right camera images. One thing to note is which landmarks showed up in specific time step, in other words, there might be some landmarks that are not observable at different time step.

In our configuration, we assume that the transformation from the IMU to the camera optical frame belongs to special Euclidean group SE(3) (extrinsic parameters) and the stereo camera calibration matrix M (intrinsic parameters) are known. The camera calibration matrix is defined as the following equation:

$$M = \begin{bmatrix} fs_u & 0 & c_u & 0 \\ 0 & fs_v & c_v & 0 \\ fs_u & 0 & c_u & -fs_u b \\ 0 & fs_v & c_v & 0 \end{bmatrix} \tag{7}$$

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

III. Technical Approach

A. Extended Kalman Filter

The Extended Kalman Filter (EKF) is a nonlinear version of the Kalman Filter, which linearizes about an estimate of the current mean and covariance with a moment matching approach. The nonlinear Kalman Filter is a Bayes filter with the following assumptions:

1. The prior pdf is Gaussian distribution
2. The state $\boldsymbol{x}_{t+1}$ is affected by Gaussian noise, that is:
$$\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{w}_t),\ \boldsymbol{w}_t \sim \mathcal{N}(0, \boldsymbol{W}) \tag{8}$$
3. The observation $\boldsymbol{z}_t$ is affected by Gaussian noise, that is:
$$\boldsymbol{z}_{t+1} = h(\boldsymbol{z}_t, \boldsymbol{v}_t), \boldsymbol{v}_t \sim \mathcal{N}(0, \boldsymbol{V}) \tag{9}$$
4. The process noise $\boldsymbol{w}_t$ and measurement noise $\boldsymbol{v}_t$ are independent of each other
5. The posterior pdf is forced to be Gaussian via approximation

The challenge of the nonlinear Kalman Filter is that the predicted and updated pdfs are not Gaussian and cannot be evaluated in closed form. Using moment matching, we can force the predicted and updated pdfs to be Gaussian by evaluating their first and second moments and modeling them with Gaussians with the same moments.

The Extended Kalman Filter uses a first-order Taylor series to approximate the integrals required to implement the nonlinear Kalman Filter. Thus, the approximation of the motion model would be as follow:

$$\begin{aligned} f(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{w}_t) &\approx f(\boldsymbol{\mu}_{t|t}, \boldsymbol{u}_t, 0) + \left[\frac{df}{d\boldsymbol{x}}(\boldsymbol{\mu}_{t|t}, \boldsymbol{u}_t, 0)\right](\boldsymbol{x}_t - \boldsymbol{\mu}_{t|t}) \\ &\quad + \left[\frac{df}{d\boldsymbol{w}}(\boldsymbol{\mu}_{t|t}, \boldsymbol{u}_t, 0)\right](\boldsymbol{w}_t - 0) \\ &\approx f(\boldsymbol{\mu}_{t|t}, \boldsymbol{u}_t, 0) + \mathbf{F}_t(\boldsymbol{x}_t - \boldsymbol{\mu}_{t|t}) + \mathbf{Q}_t \boldsymbol{w}_t \end{aligned} \tag{10}$$

where $\boldsymbol{F}_t$ and $\boldsymbol{Q}_t$ are Jacobians w.r.t. $\boldsymbol{x}$ and w. Since $\boldsymbol{x}_t$ and $\boldsymbol{w}_t$ are independent of each other, and because the approximation given in Eq. (12) is linear, we know the distribution is Gaussian, i.e., $\boldsymbol{x}_{t+1} \sim \mathcal{N}(\boldsymbol{\mu}_{t+1|t}, \Sigma_{t+1|t})$:

$$\boldsymbol{\mu}_{t+1|t} = f(\boldsymbol{\mu}_{t|t}, \boldsymbol{u}_t, 0) \tag{11}$$

$$\Sigma_{t+1|t} = F_t \Sigma_{t|t} F_t^T + Q_t W Q_t^T \tag{12}$$

We begin our localization problem in any given timestep t using our IMU readings to get some idea of our new position in space, after a control input $\boldsymbol{u}_t$ has been applied. Our motion model, in this system, is a discretized version of nominal and perturbed kinematics. This allows us, with some time discretization, to employ the exponential map:

$$\boldsymbol{\mu}_{t+1|t} = \exp(-\tau \hat{\boldsymbol{u}}_t) \boldsymbol{\mu}_{t|t} \tag{13}$$

where $\hat{\boldsymbol{u}}_t \in \mathbb{R}^{4\times 4}$ represents the hat map of the velocity $\boldsymbol{u}_t = [\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t]^T \in \mathbb{R}^6$ composed of the linear velocity vector $\boldsymbol{\lambda}_t \in \mathbb{R}^3$ and rotational velocity vector $\boldsymbol{\omega}_t \in \mathbb{R}^3$:

$$\hat{\boldsymbol{u}}_t = \begin{bmatrix} \hat{\boldsymbol{\omega}}_t & \hat{\boldsymbol{\lambda}}_t \\ \mathbf{0} & 0 \end{bmatrix} \in \mathbb{R}^{4\times 4} \tag{14}$$

Here, for any vector $\boldsymbol{s} \in \mathbb{R}^3$, $\hat{\boldsymbol{s}}$ is the corresponding skew symmetric matrix. Similarly, the approximation of the observation model would be described as follow:

$$\begin{aligned} h(\boldsymbol{x}_{t+1}, \boldsymbol{v}_{t+1}) &\approx h(\boldsymbol{\mu}_{t+1|t}, 0) + \left[\frac{dh}{d\boldsymbol{x}}(\boldsymbol{\mu}_{t|t}, \boldsymbol{u}_t, 0)\right](\boldsymbol{x}_{t+1} - \boldsymbol{\mu}_{t+1|t}) \\ &\quad + \left[\frac{df}{d\boldsymbol{v}}(\boldsymbol{\mu}_{t+1|t}, 0)\right](\boldsymbol{v}_{t+1} - 0) \\ &\approx h(\boldsymbol{\mu}_{t+1|t}, 0) + \mathbf{H}_{t+1}(\boldsymbol{x}_{t+1} - \boldsymbol{\mu}_{t+1|t}) + \mathbf{R}_{t+1} \boldsymbol{v}_{t+1} \end{aligned} \tag{15}$$

where $\boldsymbol{H}_{t+1}$and $\boldsymbol{R}_{t+1}$ is the Jacobian w.r.t. $\boldsymbol{x}$ and $\boldsymbol{v}$. Based on the equations above, the models of Extended Kalman Filter will be the following equations:

1. Prior:

$$\boldsymbol{x}_t | \boldsymbol{z}_{0:t}, \boldsymbol{u}_{0:t-1} \sim \mathcal{N}(\boldsymbol{\mu}_{t|t}, \Sigma_{t|t}) \tag{16}$$

2. Motion model:

$$\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t, \boldsymbol{u}_t, \boldsymbol{w}_t), \boldsymbol{w}_t \sim \mathcal{N}(0, \boldsymbol{W}) \tag{17}$$

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

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

3. Observation Model:

$$\boldsymbol{z}_t = h(\boldsymbol{z}_t, \boldsymbol{v}_t), \boldsymbol{v}_t \sim \mathcal{N}(0, \boldsymbol{V}) \tag{18}$$

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

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

4. Prediction:

$$\mu_{t+1|t} = f\left(\mu_{t|t}, \boldsymbol{u}_t, 0\right) \quad (19)$$

$$\Sigma_{t+1|t} = F_t \Sigma_{t|t} F_t^T + Q_t W Q_t^T \quad (20)$$

5. Update:

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

$$\Sigma_{t+1|t+1} = (\mathbf{I} + \mathbf{K}_{\mathrm{t+1|t}} \mathbf{H}_{\mathrm{t+1}}) \Sigma_{t+1|t} \quad (21)$$

6. Kalman gain:

$$\mathbf{K}_{\mathrm{t+1|t}} \coloneqq \Sigma_{t+1|t} \mathbf{H}_{t+1}^T (\mathbf{H}_{\mathrm{t+1}} \Sigma_t \mathbf{H}_{t+1}^T + \mathbf{R}_{\mathbf{V}} \mathbf{R}_{\mathrm{t+1}}^{\mathrm{T}})^{-1} (22)$$

B. EKF-based Visual Mapping

For the landmark-based visual mapping problem, we assume that the inverse IMU pose is known. Moreover, as described above, we assume that the landmarks are static and the data association π_t= {1, …, M} → {1, …, N_t} stipulating which landmarks were observed at each time t is pre-computed by an external algorithm. The observation model can be written with the Gaussian prior and observation noise:
Prior:

$$\mathrm{m}|z_{0:t} \sim \mathcal{N}(\mu_t, \Sigma_t) \quad (23)$$

$$\mu_t \in \mathbb{R}^{3M}$$

$$\Sigma_t \in \mathbb{R}^{3M \times 3M}$$

Observation model:

$$z_{t,i} = h\left(\mathrm{U_t}, \mathrm{m}_j\right) + \mathrm{v}_{t,i}$$

$$= M\pi\left({}_oT_i\mathrm{U_t}, \underline{\mathrm{m}}_j\right) + \mathrm{v}_{t,i} \quad (24)$$

$$\mathrm{v}_{t,i} \sim \mathcal{N}(0, \mathrm{V})$$

where $\mu_t \in \mathbb{R}^{3M}$is the expectation of locations of all landmarks, $\Sigma_t \in \mathbb{R}^{3M \times 3M}$is the covariance of the estimate, and M is the calibration matrix in Eq. 7. And π refers to the projection function:

$$\pi(\mathbf{q}) \coloneqq \frac{1}{\mathrm{q}_3}\mathbf{q} \in \mathbb{R}^4 \quad (25)$$

and whose derivative is given by:

$$\frac{d\pi}{d\mathbf{q}} = \frac{1}{q_3}\begin{bmatrix} 1 & 0 & -\frac{q_1}{q_3} & 0 \\ 0 & 1 & -\frac{q_2}{q_3} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{q_4}{q_3} & 1 \end{bmatrix} \quad (23)$$

We can stack all the observation $z_{t,i}$ as a single vector $z_t \in \mathbb{R}^{4N_t}$ and rewrite the observation model equation:

$$z_t = M\pi\left({}_oT_i\mathrm{U_t}\underline{\mathrm{m}}\right) + \mathrm{v_t} \quad (27)$$

$$\mathrm{v}_t \sim \mathcal{N}(0, \mathrm{I} \otimes \mathrm{V}) \quad (28)$$

where ⊗ is the Kronecker product.
For Extended Kalman Filter, we need to derive the differentiation of the observation model with respect to m evaluated with μ_t. Consider a small perturbation $\delta\mu_{t,j}$ for the location of landmark j:

$$\mathrm{m}_j = \mu_{t,j} + \delta\mu_{t,j} \quad (29)$$

The first-order Taylor series approximation to observation time t can be then written as:

$$\begin{aligned} \mathbf{z}_{t,i} &= \mathbf{M}\pi({}_O\mathbf{T}_I\mathbf{U}_t(\boldsymbol{\mu}_{t,j} + \delta\boldsymbol{\mu}_{t,j})) + \mathbf{v}_{t,i} \\ &= \mathbf{M}\pi({}_O\mathbf{T}_I\mathbf{U}_t(\underline{\boldsymbol{\mu}}_{t,j} + \mathbf{P}^T\delta\boldsymbol{\mu}_{t,j})) + \mathbf{v}_{t,i} \\ &\approx \mathbf{M}\pi({}_O\mathbf{T}_I\mathbf{U}_t\underline{\boldsymbol{\mu}}_{t,j})+ \\ &\mathbf{M}\frac{d\pi}{d\mathbf{q}}({}_O\mathbf{T}_I\mathbf{U}_t\underline{\boldsymbol{\mu}}_{t,j}){}_O\mathbf{T}_I\mathbf{U}_t\mathbf{P}^T\delta\boldsymbol{\mu}_{t,j} + \mathbf{v}_{t,i} \end{aligned} \quad (30)$$

where P $= [\mathbf{I}, \mathbf{0}] \in \mathbb{R}^{3\times4}$is the projection matrix.
The stereo camera Jacobian is given by:

$$H_{t,i,j} = M\frac{d\pi}{d\mathbf{q}}({}_oT_{iw}T_i^{-1}\underline{\boldsymbol{\mu}}_{t,j}){}_oT_{iw}T_i^{-1}P^T \quad (31)$$

and the EKF update:

$$\mathrm{K}_t = \Sigma_t H_t^T\left(H_t\Sigma_t H_t^T + \mathrm{I} \otimes \mathrm{V}\right)^{-1} \quad (32)$$

$$\mu_{t+1} = \mu_t + \mathrm{K}_t(z_t - \tilde{z}_t) \quad (33)$$

$$\Sigma_{t+1} = (\mathbf{I} - \mathrm{K}_t H_t)\Sigma_t \quad (34)$$

Since we assume all the landmarks are static, there is no need to perform prediction steps in EKF-based visual mapping. For each landmark, we initialize its position at the first timestep t we observe it. The following equation with the observation $z_{t,i}$ is used to compute the landmark initialization:

$$z_{t,i} = M\pi\left({}_oT_i\mathrm{U_t}, \underline{\mathrm{m}}_j\right) \quad (35)$$

C. EKF-based Visual-Inertial Odometry

The localization problem aims to estimate the inverse IMU pose of the robot given the IMU measurements $\mathrm{u}_t = [\mathrm{v}_t^T, \omega_t^T]$, the visual feature observation $z_{0:t}$, and the landmark coordinates $\underline{\mathrm{m}}$ in the world frame. With the same assumption in the visual mapping problem, data association between observations and landmarks is computed by an external algorithm. The motion model can be written with the Gaussian prior and process noise:

Prior:

$$\mathrm{U_t}|z_{0:t}, u_{0:t-1} \sim \mathcal{N}(\mu_{t|t}, \Sigma_{\boldsymbol{t|t}}) \tag{36}$$

$$\mu_{t|t} \in SE(3)$$

$$\Sigma_{t|t} \in \mathbb{R}^{6\times 6}$$

Motion model:

$$\mathrm{U_{t+1}} = \exp\left(-\tau((u_t + w_t))^\wedge\right)\mathrm{U_t} \tag{37}$$

$$\mathrm{w}_{t,i} \sim \mathcal{N}(0, \mathrm{W})$$

where τ is time difference in seconds between two contiguous frames. In order to separate the effect of the noise w_t from the motion of the deterministic part of $wTi, t = \mathrm{U}_t^{-1}$, we take advantages of the discrete-time perturbation idea in and rewrite the motion model in terms of nominal kinematics and zero-mean perturbation kinematics:

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

The observation model is the same as the definition in the visual mapping problem. In order to derive the update steps for Extended Kalman Filter, we need the observation model Jacobian $H_{t+1|t} \in \mathbb{R}^{4N_t\times 6}$ with respect to the inverse IMU pose $\mathrm{U_t}$ evaluated at $\mu_{t|t}$. Similar to the update steps in EKF-based visual mapping, the update steps for visual-inertial odometry can be composed of the following equations:

Prior:

$$\mathrm{U_{t+1}}|z_{0:t}, u_{0:t} \sim \mathcal{N}(\mu_{t+1|t}, \Sigma_{t+1|t}) \tag{39}$$

$$\mu_{t|t} \in SE(3)$$

$$\Sigma_{t|t} \in \mathbb{R}^{6\times 6}$$

Predicted observations:

$$\tilde{z}_{t+1,i} = M\pi({}_oT_i\mu_{t+1|t}, \underline{\mathrm{m}}_j) \tag{40}$$

Observation matrix:

$$\begin{aligned} \mathbf{H}_{i,t+1|t} &= \mathrm{M}\frac{d\pi}{d\mathbf{q}}({}_O\mathrm{T}_I\boldsymbol{\mu}_{t+1|t}\mathbf{m}_j){}_O\mathrm{T}_I(\boldsymbol{\mu}_{t+1|t}\mathbf{m}_j)^\odot \\ \mathbf{H}_{t+1|t} &= \begin{bmatrix} \mathbf{H}_{1,t+1|t} \\ \mathbf{H}_{2,t+1|t} \\ \vdots \\ \mathbf{H}_{N_{t+1},t+1|t} \end{bmatrix} \in \mathbb{R}^{4N_t\times 6} \end{aligned} \tag{41}$$

EKF Update:

$$\mathrm{K}_{t+1|t} = \Sigma_{t+1|t}H_{t+1|t}^T(H_{t+1|t}\Sigma_{t+1|t}H_{t+1|t}^T + \mathrm{I}\otimes\mathrm{V})^{-1}$$

$$\mu_{t+1|t+1} = \exp\left(K_{t+1|t}(\widehat{\boldsymbol{z}_{t+1}} - \tilde{\boldsymbol{z}}_{t+1})\right)\mu_{t+1|t} \tag{42}$$

$$\Sigma_{t+1|t+1} = (\mathrm{I} - K_{t+1|t}H_{t+1|t})\Sigma_{t+1|t} \tag{43}$$

D. EKF SLAM

To achieve the goal of estimating position of landmarks and the pose of the robot simultaneously, the proposed idea is to merge the predict and update steps of Extended Kalman Filter based visual mapping and visual-inertial odometry. First of all, the joint estimated state and covariance under the Gaussian assumption are defined as follows:

$$\mu = \begin{bmatrix}\mu_m \\ \mu_p\end{bmatrix} \in \mathbb{R}^{6\times 6} \tag{44}$$

$$\Sigma_{t|t} \in \mathbb{R}^{(6+3M)\times(6+3M)}$$

where μ_m is the estimated landmark position in Eq. 23. and μ_p is the estimated six degrees of freedom of inverse IMU pose in Eq. 36.

The Extended Kalman Filter predict step on the joint estimated state and covariance is derived from the predict step of visual-inertial odometry only because all the landmarks are assumed static. The equations given the IMU measurement u_t are listed as follows:

$$\mu_{t+1|t} = \begin{bmatrix}\mu_{m,t+1|t} \\ \mu_{p,t+1|t}\end{bmatrix} = \begin{bmatrix}\mu_{m,t+1|t} \\ \exp(-\tau\widehat{u_t})\,\mu_{p,t|t}\end{bmatrix} \tag{45}$$

$$\Sigma_{t|t} = F_t\Sigma_{t|t}F_t^T + \mathrm{W}$$

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

$$\mathrm{W} = \begin{bmatrix}0 & 0 \\ 0 & \mathrm{W_p}\end{bmatrix}$$

The update step is formed by combining the update step of both visual mapping and visual-inertial odometry. The update step equations are listed as follows:

Predicted observations:

$$\tilde{z}_{t,i} = M\pi\left({}_oT_i\mathrm{U_t}, \underline{\mu}_{m,t,j}\right) \in \mathbb{R}^4 \tag{46}$$

Observation matrix:

$$H_{t+1|t} = [H_{m,t+1|t} \quad H_{p,t+1|t}] \in \mathbb{R}^{4N_t\times(3M+6)} \tag{47}$$

EKF update:

$$\mathrm{K}_{t+1|t} = \Sigma_{t+1|t}H_{t+1|t}^T(H_{t+1|t}\Sigma_{t+1|t}H_{t+1|t}^T + \mathrm{I}\otimes\mathrm{V})^{-1} \tag{48}$$

$$\mu_{t+1|t} = \begin{bmatrix}\mu_{m,t+1|t+1} \\ \mu_{p,t+1|t+1}\end{bmatrix} = \begin{bmatrix}\mu_{m,t+1|t} + \mathrm{K}_{t+1|t}(z_t - \tilde{z}_t) \\ \exp\left(K_{t+1|t}(\widehat{z_{t+1}} - \tilde{z}_{t+1})\right)\mu_{p,t+1|t}\end{bmatrix} \tag{49}$$

IV. Results

The proposed EKF based SLAM algorithm is tested with 2 data set, and all of them are collected in real driving scenarios. The hyper-parameters used in our algorithm are same for all the data set except for the number of landmarks. The results are shown in Fig.2., which show the estimated trajectory and the 2D position of the visual features. The red line in the figure is the estimate robot trajectory, and the green dots are the position of landmarks. The estimate position of the visual features also distributes in a reasonable way.

On the other hand, the IMU-based landmark mapping does not provide excellent results. The reason is that the trajectory is not corrected based on the observations of visual features. The quality of the estimate position of the visual features also relies on the correct pose of the robot. Thus, the landmark estimates are also not accurate.

Also, one important observance is that noise covariances W are quite small, and it would lead poor results if such hyperparameters getting to large. In comparison, another noise V was quite robust to change, with maps changed only slightly between 1000 to 4000.

One failure in this project review is the EKF SLAM part. The problem researcher of this paper faced is singular matrix when computing matrix inversion in observation matrix. Since the whole part of $H_{t+1|t}\Sigma_{t+1|t}H_{t+1|t}^T$in Eq. 48. Is positive semi-definite, this mistake should not happen. Another reason related to this is inappropriate variance value, which too large may cause instability in the algorithm. The expected result of visual-inertial SLAM should be similar to the one in Fig. 2. with little bias.

References

[1] A. Filipescu, V. Minzu, B. Dumitrascu, A. Filipescu and E. Minca, "Trajectory-tracking and discrete-time sliding-mode control of wheeled mobile robots," 2011 IEEE International Conference on Information and Automation, 2011, pp. 27-32, doi: 10.1109/ICINFA.2011.5948958.

[2] https://docs.opencv.org/3.4/d9/dba/classcv_1_1StereoBM.html

[3] UCSD ECE276A: Sensing & Estimation in Robotics (Winter 2022) https://natanaso.github.io/ece276a

Parameter	Description	Value
Σ_m	Prior landmark estimate covariance	$0.001 * \mathrm{I} \in \mathbb{R}^{3\times3}$
Σ_p	Prior pose estimate covariance	$0.001 * \mathrm{I} \in \mathbb{R}^{6\times6}$
V	Observation	$1000 * \mathrm{I} \in \mathbb{R}^{4\times4}$
W_p	Process noise covariance	$0.001 * \mathrm{I} \in \mathbb{R}^{6\times6}$

Table 1. Hyperparameters for EKF visual inertial SLAM

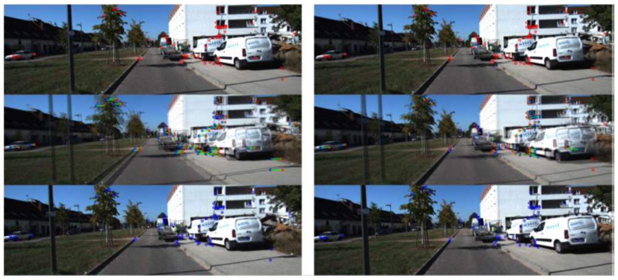

Fig.1. Visual features matched across the left-right camera frames (left) and across time (right).

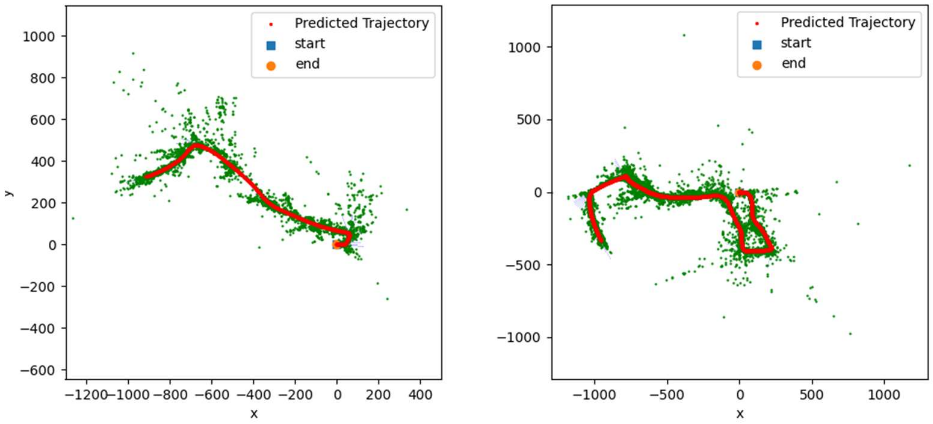

Fig. 2. Results of Experiments (a) 10.npz (b) 03.npz