Visual-Inertial SLAM

Zhichen Zhang
Department of Electrical and Computer Engineering
University of California San Diego
San Diego, California
Zhz087@ucsd.edu

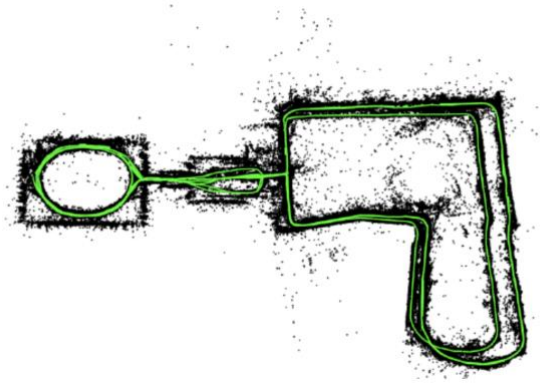
***Abstract*— This paper presented a method of implementing visual-inertial simultaneous localization and mapping (SLAM) using Extended Kalman Filter (EKF) [1]. We are provided with synchronized measurement from an IMU and a stereo camera. Intrinsic camera calibration and the extrinsic calibration between two sensors were used in order to specify the transformation from the IMU to the left camera frame.**

I. INTRODUCTION

Computer vision, classification, and object detection play significant rule in modern technology. People rely on these technologies to track object in pictures/videos to make decisions. In this paper, our goal was to implement visual-inertial simultaneous localization and mapping (SLAM).

In this project, we implemented visual-inertial SLAM using Extended Kalman Filter with IMU and stereo camera data [1].

We divided this project into three steps: IMU localization via EKF prediction, landmark mapping via EKF update, and visual-inertial SLAM.



II. PROBLEM FORMATION

A. *Data and Parameters*

t	Time stamps with shape 1xt.
features	Visual feature point coordinates in stereo images. 4xnxt, where n is the number of features.
Linear_velocity	IMU measurements in IMU frame with shape 3xt.
Rotational_velocity	IMU measuremens in the IMU frame with shape 3xt.
K	Left camera intrinsic matrix.
b	Stereo camera baseline
Cam_T_imu	Extrinsic matrix from IMU to left camera in SE(3) with shape 4x4.

AVI videos of landmarks:

We were provided with two training datasets with two different avi videos containing landmarks from the left and right cameras. The coordinates of these images were stored in *features*.

B. *IMU localization via EKF prediction.*

The prediction step predicted mu and sigma from the previous pose. Mu is a 4x4 matrix in SE(3) whereas sigma is a 6x6 matrix. In this case, we used the hat map for imu input.

The **hat map** $\hat{\cdot}: \mathbb{R}^3 \to \mathfrak{so}(3)$ transforms an $\mathbb{R}^3$ vector to a skew-symmetric matrix:

$$\hat{\omega} := \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}$$

Then we were able to update our mu and sigma using:

$$\mu_{t+1|t} = \exp(-\tau\hat{u}_t)\,\mu_{t|t} \qquad u_t := \begin{bmatrix} v_t \\ \omega_t \end{bmatrix}$$
$$\Sigma_{t+1|t} = \mathbb{E}[\xi_{t+1|t}\xi_{t+1|t}^T] = \exp\left(-\tau\hat{\hat{u}}_t\right)\Sigma_{t|t}\exp\left(-\tau\hat{\hat{u}}_t\right)^T + W$$

C. Landmark mapping via EKF update.

The update step update landmarks' world frame poses using visual feature point coordinates and transformation matrices. The update step is:

$$\hat{\mathbf{z}}_{t,i} := M\pi\left({}_O T_I\, T_t \mu_{t,j}\right) \in \mathbb{R}^4 \qquad \text{for } i = 1, \ldots, N_t$$

Jacobian matrix was calculated as:

$$H_{i,j,t} = \begin{cases} M\frac{d\pi}{d\mathbf{q}}\left({}_O T_I\, T_t \mu_{t,j}\right) {}_O T_I\, T_t D & \text{if observation } i \text{ corresponds to landmark } j \text{ at time } t \\ \mathbf{0} \in \mathbb{R}^{4\times 3} & \text{otherwise} \end{cases}$$

The update step is:

$$K_t = \Sigma_t H_t^T\left(H_t \Sigma_t H_t^T + I \otimes V\right)^{-1} \qquad I \otimes V := \begin{bmatrix} V & & \\ & \ddots & \\ & & V \end{bmatrix}$$
$$\mu_{t+1} = \mu_t + DK_t(\mathbf{z}_t - \hat{\mathbf{z}}_t)$$
$$\Sigma_{t+1} = (I - K_t H_t)\Sigma_t$$

D. Visual SLAM.

Combining the IMU prediction step with the landmark update step and an IMU update step based on the stereo camera observation model and obtained a complete visual-inertial SLAM algorithm.

$$\begin{aligned} \mathbf{z}_{t+1,i} &= M\pi\left({}_O T_I \exp\left(\hat{\xi}_{t+1|t+1}\right)\mu_{t+1|t}\mathbf{m}_j\right) \\ &\approx M\pi\left({}_O T_I\left(I + \hat{\xi}_{t+1|t+1}\right)\mu_{t+1|t}\mathbf{m}_j\right) \\ &= M\pi\left({}_O T_I \mu_{t+1|t}\mathbf{m}_j + {}_O T_I\left(\mu_{t+1|t}\mathbf{m}_j\right)^{\odot}\xi_{t+1|t+1}\right) \\ &\approx \underbrace{M\pi\left({}_O T_I \mu_{t+1|t}\mathbf{m}_j\right)}_{\hat{\mathbf{z}}_{t+1,i}} + \underbrace{M\frac{d\pi}{d\mathbf{q}}\left({}_O T_I \mu_{t+1|t}\mathbf{m}_j\right){}_O T_I\left(\mu_{t+1|t}\mathbf{m}_j\right)^{\odot}}_{H_{i,t+1|t}}\xi_{t+1|t+1} \end{aligned}$$

where for homogeneous coordiantes $r \in \mathbb{R}^4$ and $\hat{\xi} \in \mathfrak{se}(3)$:

$$\hat{\xi} r = r^{\odot}\xi \qquad \begin{bmatrix} s \\ \lambda \end{bmatrix}^{\odot} = \begin{bmatrix} \lambda I & -\hat{s} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 6}$$

After computing the predicted observation based on $\mu_{t+1|t}$ and known correspondance and computing Jacobian, we were able to perform the EKF update:

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

III. Technical Approach

A. IMU localization via EKF prediction.

The prediction step predicted mu and sigma from the previous pose. Mu is a 4x4 matrix in SE(3) whereas sigma is a 6x6 matrix. First, we stacked linear_velocity and rotational_velocity to build our control input ut. The hat map transforms an R3 matrix to a skew-symmetric matrix. Then we built u_hat and omega_hat using the hat map. The curvy hat or perturbation has the size of 6x6. We used the hat map for our imu input in this case:

The **hat map** $\hat{\cdot}: \mathbb{R}^3 \to \mathfrak{so}(3)$ transforms an $\mathbb{R}^3$ vector to a skew-symmetric matrix:

$$\hat{\omega} := \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}$$

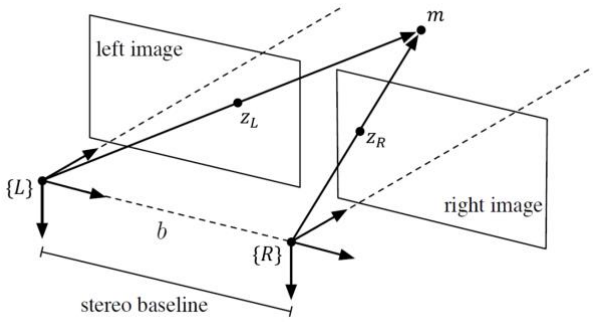
$$\hat{\hat{u}}_t := \begin{bmatrix} \hat{\omega}_t & \hat{v}_t \\ 0 & \hat{\omega}_t \end{bmatrix} \in \mathbb{R}^{6\times 6}$$

After having the hat map for u and w, we were then able to do prediction for both mu and sigma based on the previous mu and sigma. We first initialized mu to an identity matrix with size 4x4 and sigma to an identity matrix with size 6x6. Then we performed EKF prediction step shown below with W, motion model noise covariance.

$$\mu_{t+1|t} = \exp(-\tau\hat{u}_t)\,\mu_{t|t} \qquad u_t := \begin{bmatrix} v_t \\ \omega_t \end{bmatrix}$$
$$\Sigma_{t+1|t} = \mathbb{E}[\xi_{t+1|t}\xi_{t+1|t}^T] = \exp\left(-\tau\hat{\hat{u}}_t\right)\Sigma_{t|t}\exp\left(-\tau\hat{\hat{u}}_t\right)^T + W$$

B. Landmark mapping via EKF update.

The update step update landmarks' world frame poses using visual feature point coordinates and transformation matrices. In this step, we used the stereo camera model with provided baseline b and features from left and right cameras:



In this case, we have feature points from left and right camera and stereo baseline was provided. Therefore, we were able to get the stereo camera model:

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

The projection function is defined as:

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

And its derivative is defined as:

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

Now we have stereo camera model M, rotational matrix, and projection function. Then we were able to compute the predicted observation based on mu and known correspondence:

$$\hat{\mathbf{z}}_{t,i} := M\pi\left({}_O T_I \, T_t \boldsymbol{\mu}_{t,j}\right) \in \mathbb{R}^4 \qquad \text{for } i = 1, \ldots, N_t$$

Jacobian matrix was calculated as:

$$H_{i,j,t} = \begin{cases} M\frac{d\pi}{d\mathbf{q}}\left({}_O T_I \, T_t \boldsymbol{\mu}_{t,j}\right) {}_O T_I \, T_t D & \text{if observation } i \text{ corresponds t} \\ & \text{landmark } j \text{ at time } t \\ \mathbf{0} \in \mathbb{R}^{4\times 3} & \text{otherwise} \end{cases}$$

The update step is:

$$\begin{aligned} K_t &= \Sigma_t H_t^T \left(H_t \Sigma_t H_t^T + I \otimes V\right)^{-1} \\ \boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t + D K_t \left(\mathbf{z}_t - \hat{\mathbf{z}}_t\right) \\ \Sigma_{t+1} &= (I - K_t H_t)\Sigma_t \end{aligned} \qquad I \otimes V := \begin{bmatrix} V & & \\ & \ddots & \\ & & V \end{bmatrix}$$

In this update step. Kt is the Kalman gain which indicate a probability of which Gaussian should be trusted more. It is an indicator of weights of two gaussians. V is the observation model noise covariance. Since the covariance for our landmarks can get to 5000 sometimes. Therefore, we gave our V a large number.

The optical to world matrix can be calculated using M, IMU to Optical transformation, and World to IMU transformation. Therefore, we were able to transform a pose in the camera/optical frame into world frame (for landmarks).

We first initialized a flag matrix indicating which landmark has been seen before. If a landmark has not been observed, then we initialize the landmark using optical to world matrix and features, then set the flag of this landmark to True indicating that the landmark will be updated next time we see it. If a landmark has been observed, then we perform the update step for this landmark and update m, its pose.

C. Visual SLAM

Combine the IMU prediction step with the landmark update step and an IMU update step based on the stereo camera observation model to obtain a complete visual-inertial SLAM algorithm.

$$\begin{aligned} \mathbf{z}_{t+1,i} &= M\pi\left({}_O T_I \exp\left(\hat{\xi}_{t+1|t+1}\right) \boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right) \\ &\approx M\pi\left({}_O T_I \left(I + \hat{\xi}_{t+1|t+1}\right) \boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right) \\ &= M\pi\left({}_O T_I \boldsymbol{\mu}_{t+1|t} \mathbf{m}_j + {}_O T_I \left(\boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right)^{\odot} \xi_{t+1|t+1}\right) \\ &\approx \underbrace{M\pi\left({}_O T_I \boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right)}_{\hat{\mathbf{z}}_{t+1,i}} + \underbrace{M\frac{d\pi}{d\mathbf{q}}\left({}_O T_I \boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right) {}_O T_I \left(\boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right)^{\odot}}_{H_{i,t+1|t}} \xi_{t+1|t+1} \end{aligned}$$

where for homogeneous coordiantes $r \in \mathbb{R}^4$ and $\hat{\xi} \in \mathfrak{se}(3)$:

$$\hat{\xi} r = r^{\odot} \xi \qquad \begin{bmatrix} s \\ \lambda \end{bmatrix}^{\odot} = \begin{bmatrix} \lambda I & -\hat{s} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 6}$$

Next, we computed the predicted observation based on $\boldsymbol{\mu}_{t+1|t}$ and known correspondance:

$$\hat{\mathbf{z}}_{t+1,i} := M\pi\left({}_O T_I \boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right) \qquad \text{for } i = 1, \ldots, N_t$$

Then computed Jacobian:

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

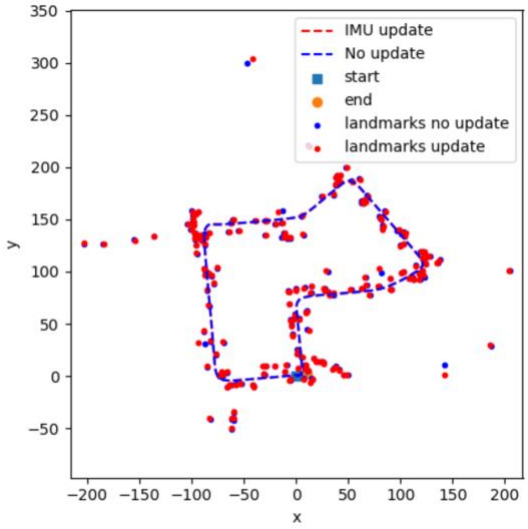
$$H_{i,t+1|t} = M\frac{d\pi}{d\mathbf{q}}\left({}_O T_I \boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right) {}_O T_I \left(\boldsymbol{\mu}_{t+1|t} \mathbf{m}_j\right)^{\odot} \in \mathbb{R}^{4\times 6}$$

we were able to perform the EKF update:

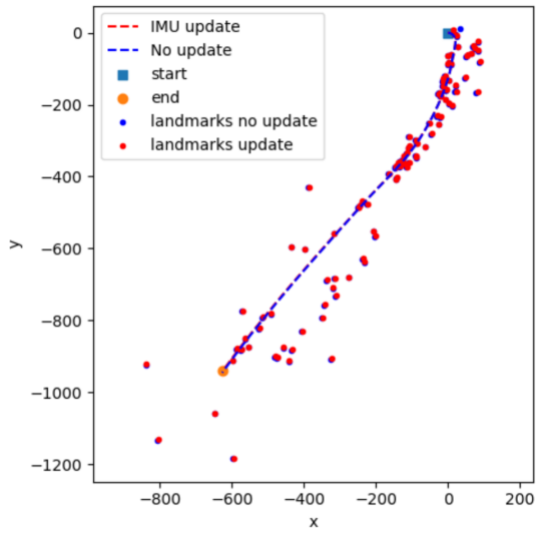
$$\begin{aligned} K_{t+1|t} &= \Sigma_{t+1|t} H_{t+1|t}^T \left(H_{t+1|t} \Sigma_{t+1|t} H_{t+1|t}^T + I \otimes V\right)^{-1} \\ \boldsymbol{\mu}_{t+1|t+1} &= \exp\left(\left(K_{t+1|t}(\mathbf{z}_{t+1} - \hat{\mathbf{z}}_{t+1})\right)^{\wedge}\right) \boldsymbol{\mu}_{t+1|t} \\ \Sigma_{t+1|t+1} &= (I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

IV. Results and Discussion

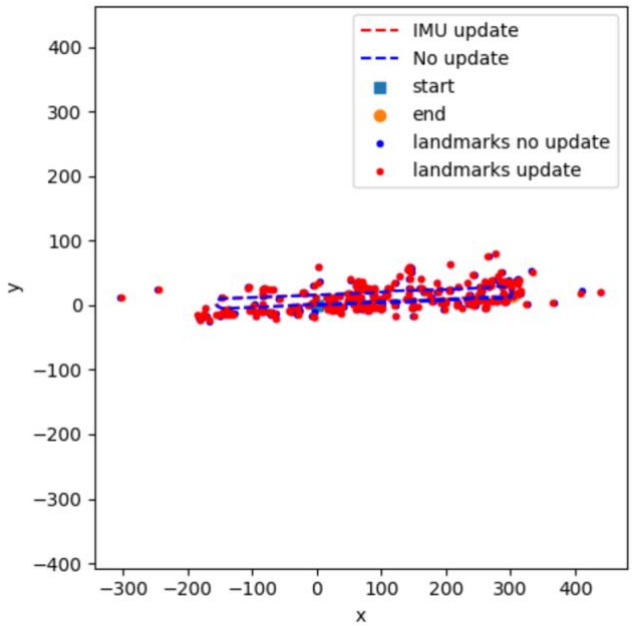
The result for set 27 is shown below:



The result for set 42 is shown below:



The result for testset 20 is shown below:



As we can see, there is not much difference (almost overlapped) between the dead-reckoning trajectories and the updated trajectories, which means that our IMU readings are reliable and to some extent, accurate.

V. Conclusion

Generally speaking, our model did a good job on trainset 27, transet 42, as well as testset 20 from my perspective. The only way to check whether it is a good result or not is to check the trajectories and landmarks against the video. As we can see, there is not much difference (almost overlapped) between the dead-reckoning trajectories and the updated trajectories, which means that our IMU readings are reliable and to some extent, accurate. We realized that sometimes the landmarks came from a moving object. For example, at 46s in testset 20, there are two landmarks came from two wheels of a moving car, this might cause some inaccuracy of our model.

References

[1] https://en.wikipedia.org/wiki/Extended_Kalman_filter