Project 3: Visual-Inertial SLAM

Jaidev Shriram (University of California, San Diego)

I. Introduction

A fundamental problem in autonomous driving is the simultaneous localisation and mapping of a moving car. The fundamental difficulty with the problem is how coupled the two tasks, localisation and mapping, are, which is then made more difficult by the fact that the world around is highly complex and varied in nature, and that the car's sensors. As a result, trivially tracking the location of the car and building a map will not suffice, as sensor noise can compound significantly and lead to drift. In this project, we combine two sources of information - images captured from an onboard camera and the inertial information from the IMU sensor to simultaneously localise and build a map of the surroundings.

We focus on a probabilistic approach to SLAM, where we model the probability of the car's position and various points on the map as a Gaussian distribution. As the car moves, we utilise the incoming sensor data to update these beliefs based on an observation model. In this scenario, we build a 2D map of the world, indicating whether a grid in the real world is occupied or free. At a high level, performing these iterative predictions of the pose using a robot's motion model, and updating the pose based on sensor observations is representative of a bayes filter. In this project, we use the Extended Kalman Filter, a version of the Bayes filter, which applies to non-linear systems.

Our results show that this filtering of the pose using visual features is able to effectively estimate the pose of the robot, while building a high-quality map of the scene.

II. Problem Formulation

In our scenario, we must localise a car equipped with a stereo RGB camera and IMU. Formally, we are tasked with estimating the state $\mathbf{x}_{0:t}$ of the car till a time instant t, based on the corresponding control inputs $\mathbf{u}_{0:t-1}$ and observations $\mathbf{z}_{0:t}$ while building a map $\mathbf{m}$. An observation $\mathbf{z}_{ij}$ represents the pixel coordinates of the point $\mathbf{m}_j \in \mathbb{R}^3$ at the *i*th instant in the left and right camera of the stereo set-up.

A. SLAM

We assume that the robot trajectory is represented as a Markov chain, where the position of the robot x_t is only dependent on its previous state x_{t-1} and the previous input u_t, and the observation $\mathbf{z}_t$.

As a result, the motion model can be expressed as:

$$x_{t+1} = f(x_t, u_t, w_t) \tag{II.1}$$

,

where w_t is motion noise.

Similarly, the observation seen from the camera at time t can be seen as:

$$z_t = h(x_t, m, v_t), \tag{II.2}$$

where h is the observation model function and v_t is observation noise. Note that z_t is available at run time and that the goal of SLAM is to calculate $\mathbf{m}$ and $\mathbf{x}$, to maximise the joint probability $p(\mathbf{m}, \mathbf{x}|\mathbf{z}, \mathbf{u})$ given the robot's sensor information. Due to the markox assumption, we can say that we are maximising the following probability instead:

$$p(z_i|x_i, m) * p(x_i|x_{i-1}, u_{i-1}) \tag{II.3}$$

The first term is the observation model, and the second term is the motion model.

B. Visual Mapping

The mapping problem is concerned with estimating the coordinates $\mathbf{m} \in \mathbb{R}^{3M}$ of landmarks that are visible in observations $\mathbf{z}_t \in \mathbb{R}^{4N_t}$, where N_t is the number of visible landmarks. Here, we assume that there is a pre-computed mapping between the landmarks and the observations seen at any given time instant. Further, we assume that the landmarks are all static. Hence, given the robot's pose, landmark observations can be converted to 3D space, and the observation model can focus on maximising $p(z_i|x_i, \mathbf{m})$, where z_i will consist of observations of the same landmark across multiple timesteps.

C. Visual-Inertial Odometry

Assuming that we are given the linear and angular velocity of the moving car (v_t, ω_t), as well as the world-frame landmark coordinates $\mathbf{m} \in \mathbb{R}^{3M}$ and their corresponding feature observations z_t, we must estimate the pose of the car $T_t \in SE(3)$.

D. Bayes Filter

As stated earlier, we will adopt a probabilistic inference technique, the Bayes Filter, to estimate the state of the car by combining visual evidence and the control readings from the IMU. The Bayes filter keeps track of:

1) Predicted PDF: $p_{t+1|t}(x_{t+1}) = p(x_{t+1}|z_t, u_t)$,
2) Updated PDF: $p_{t+1|t+1}(x_{t+1}) = p(x_{t+1}|z_{t+1}, u_t)$,

where we have combined the Markov assumption and rules in conditional probability. The prediction step incorporates information from the motion model to predict the pose of the car at instant t and the update step subsequently, refines this pose estimate by accounting for the observation model.

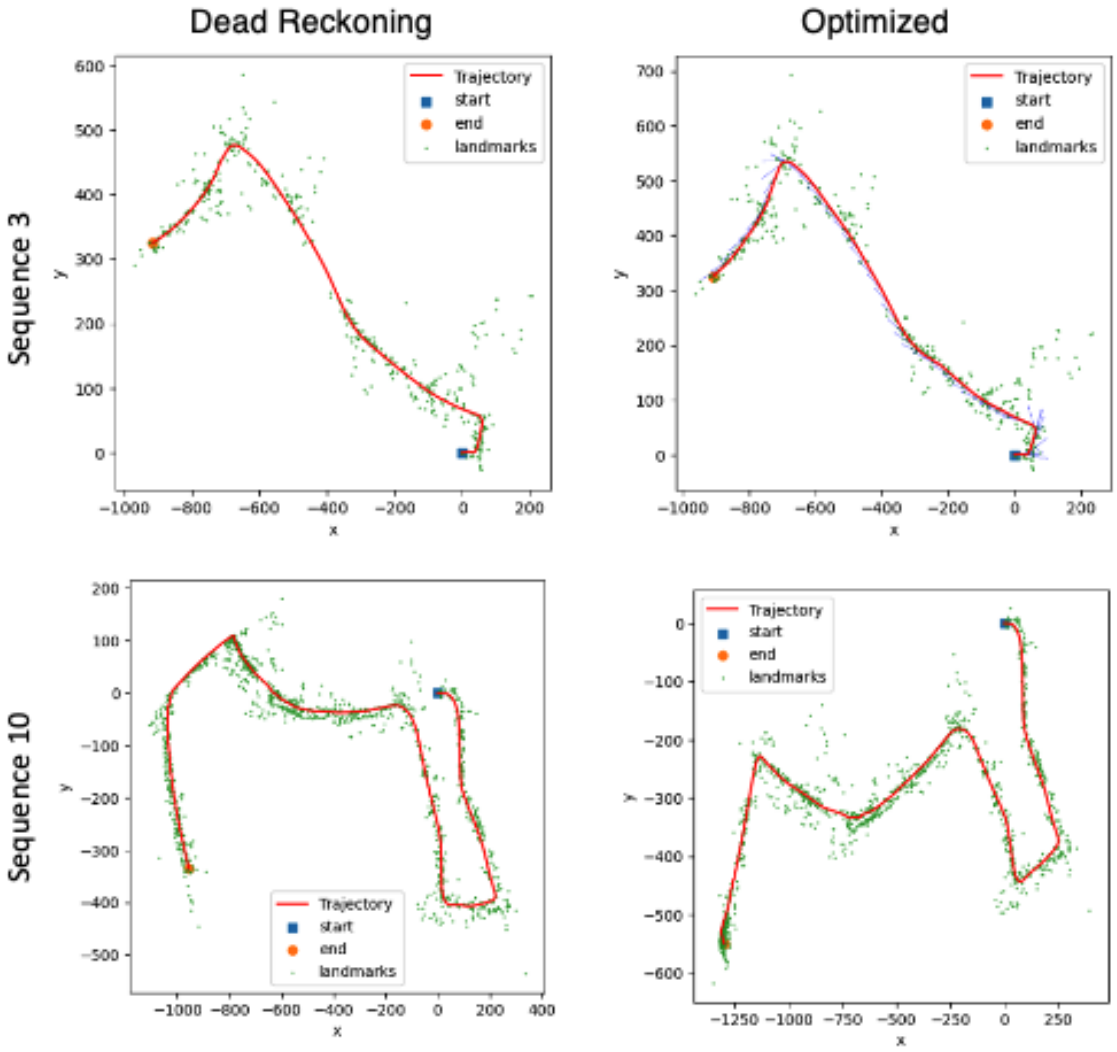

Fig. 1: Dead Reckoning and Optimised Trajectory for Sequence 3 and 10.

III. Approach

A. Robot Configuration

In this project, we work with a moving car equipped with a stereo RGB camera and IMU.

IMU: The IMU measures the linear and angular velocity of the data, v_t and ω_t of the robot at time t.

Stereo Camera: A stereo RGB camera is placed on the the car, with its position with respect to the IMU given by the transformation matrix ${}_OT_I \in SE(3)$. The stereo camera is pre-calibrated and the calibration matrix can be expressed as:

$$K_s = \begin{bmatrix} f_u & 0 & c_u & 0 \\ 0 & f_v & c_v & 0 \\ f_u & 0 & c_u & -f_u * b \\ 0 & f_v & c_v & 0 \end{bmatrix}, \tag{III.1}$$

where f_u, f_v are the focal length and c_u, c_v are the coordinates of the camera center, and b is the baseline of the stereo camera. For the purposes of this project, we will assume that the transformation matrix ${}_OT_I$ converts the axes from the world frame order to optical frame.

B. Extended Kalman Filter

Due to the non-linear nature of the car's motion and observation model, we adopt an extended kalman filter to improve our pose estimates.

We assume a gaussian prior for the map such that $p(\mathbf{m}|\mathbf{z}_{0:t}) \sim N(\mu_t, \Sigma_t)$ with $\mu_t \in \mathbb{R}^{3M}$ and $\Sigma_t \in \mathbb{R}^{3M \times 3M}$. Similarly, we assume that the robot's pose is a gaussian distribution such that $T_t|\mathbf{z}_{0:t}, \mathbf{u}_{0:t-1} \sim N(\mu_{t|t}, \Sigma_{t|t})$, where $\mu_{t|t} \in SE(3)$ and $\Sigma_{t|t} \in \mathbb{R}^{6\times 6}$. There are two components to this problem: Visual Mapping and Visual-Inertial Odometry. We will describe each separately but note that they are done simultaneously.

C. Visual Mapping

Visual mapping assumes a gaussian prior on the landmark coordiantes, conditioned on the observations, $\mathbf{m}|\mathbf{z}_{0:t} \sim$

$N(\mu_{t+1|t}, \Sigma_{t+1|t}), \mu_{t+1|t} \in \mathbb{R}^{3M}, \Sigma_{t+1|t} \in \mathbb{R}^{3M\times 3M}$. Since we assume that the landmarks are all static, the landmarks only have an observation model and subsequently, an EKF update step. Let us define the observation model as:

$$z_{t,i} = h(T_t, \mathbf{m}_j) + v_{t,i} = K_s \pi({}_O T_I T_t^{-1} \underline{\mathbf{m}_j}) + v_{t,i}, \quad \text{(III.2)}$$

where $\underline{\mathbf{m}_j} = \begin{bmatrix} \mathbf{m}_j \\ 1 \end{bmatrix}$ and π is the projection function such that:

$$\pi(\mathbf{q}) = \frac{1}{q_3}\mathbf{q} \in \mathbb{R}^4 \quad \text{(III.3)}$$

Note that $z_{t,i}$ refers to the the ith observation at time t of a landmark j. If we stack these observations, we obtain:

$$z_t = K_s \pi({}_O T_I T_t^{-1} \underline{\mathbf{m}}) + v_t, \quad \text{(III.4)}$$

where v_t is $I*$ observation_noise. Putting it all together, we can say the update step given a new observation $\mathbf{z}_{t+1} \in \mathbb{R}^{4N_{t+1}}$.

$$\begin{aligned} K_{t+1} &= \Sigma_{t+1|t} H_{t+1}^T (H_{t+1} \Sigma_{t+1|t} H_{t+1|t}^T + I)^{-1} \\ \mu_{t+1|t+1} &= \mu_{t+1|t} + K_{t+1}(\mathbf{z}_{t+1} - K_s \pi({}_O T_I T_{t+1}^{-1} \mu_{t+1|t})) \\ \Sigma_{t+1|t+1} &= (I - K_{t+1} H_{t+1}) \Sigma_{t+1|t} \end{aligned} \quad \text{(III.5)}$$

Here $H_{t+1} \in \mathbb{R}^{4N_t \times 3M}$ is the observation model jacobian evaluated at μ_t, such that:

$$H_{t+1,i,j} = \frac{\partial h(T_{t+1|t}, \mathbf{m}_j)}{\partial \mathbf{m}_j}, \quad \text{(III.6)}$$

evaluated at $\mathbf{m}_j = \mu_{t,j}$. Note that since there is no prediction step $\mu_{t+1|t} = \mu_{t|t}$ for all map points. This derivative can be calculated ussing chain rule to give:

$$H_{t+1,i,j} = K_s \frac{d\pi}{d\mathbf{q}} ({}_O T_I T_{t+1}^{-1} \underline{\mu_{t,j}}) {}_O T_I T_{t+1}^{-1} P^T, \quad \text{(III.7)}$$

where the ith observation corresponds to the jth landmark point and P is $[I_{3\times 3}, 0_{3\times 1}]$.

D. Visual-Intertial Odometry

With visual odometry, we intend to represent the pose as a gaussian distribution, however since $T \in SE(3)$, we do this by considering a perturbation on this pose. Specifically, we assume that $T_t | \mathbf{z}_{0:t}, \mathbf{u}_{0:t-1} \sim N(\mu_{t|t}, \Sigma_{t|t})$ with $\mu_{t|t} \in SE(3)$ and $\Sigma_{t|t} \in \mathbb{R}^{6\times 6}$. Hence, $T_t = \mu_{t|t} \exp(\delta \hat{\mu_{t|t}})$ with $\delta\mu_{t|t} \sim N(0, \Sigma_{t|t})$.

The motion model of the car and the predict step of the car assuming control input $\mathbf{u}_t = \begin{bmatrix} v_t \\ \omega_t \end{bmatrix} \in \mathbb{R}^6$ is:

$$\begin{aligned} \mu_{t+1|t} &= \mu_{t|t} \exp(\tau_t \hat{\mathbf{u}}_t) \\ \Sigma_{t+1|t} &= \exp(-\tau \mathbf{u}_t^\curlywedge) \Sigma_{t|t} \exp(-\tau \mathbf{u}_t{}^\curlywedge)^T + W, \end{aligned} \quad \text{(III.8)}$$

where W is the covariance of the noise.

The observation model is identical to the one used in the visual mapping phase, except here the pose of the car is represented by $\mu_{t+1|t}$. Hence,

$$\tilde{\mathbf{z}}_{t+1,i} = K_s \pi({}_O T_I \mu_{t+1|t}^{-1} \underline{\mathbf{m}_j}) \quad \text{(III.9)}$$

Similar to the visual mapping update step, the update equations for the pose are:

$$\begin{aligned} K_{t+1} &= \Sigma_{t+1|t} H_{t+1}^T (H_{t+1} \Sigma_{t+1|t} H_{t+1|t}^T + I)^{-1} \\ \mu_{t+1|t+1} &= \mu_{t+1|t} \exp((K_{t+1}(z_{t+1} - \tilde{z}_{t+1}))\hat{}) \\ \Sigma_{t+1|t+1} &= (I - K_{t+1} H_{t+1}) \Sigma_{t+1|t} \end{aligned} \quad \text{(III.10)}$$

Here, H is the observation model jacobian calculated with respect to the pose $\mu_{t+1|t}$, which is expressed as:

$$H_{t+1,i} = -K_s \frac{d\pi}{d\mathbf{q}} \left({}_O T_I \mu_{t+1|t}^{-1} \underline{\mathbf{m}_j} \right) {}_O T_I \left(\mu_{t+1|t}^{-1} \underline{\mathbf{m}_j} \right)^\odot \quad \text{(III.11)}$$

The $^\odot$ operator can be expressed as:

$$\begin{bmatrix} s \\ 1 \end{bmatrix}^\odot = \begin{bmatrix} I & -\hat{s} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times 6} \quad \text{(III.12)}$$

E. Visual-Inertial SLAM

Visual Mapping and Visual-Inertial Mapping are done simultaneously, such that we have a joint state and covariance matrix:

$$\begin{aligned} \mu &= \begin{bmatrix} \mu_{map} \\ \mu_{pose} \end{bmatrix} \in \mathbb{R}^{3M+6} \\ \Sigma &= \begin{bmatrix} \Sigma_{map} & \Sigma_{map-pose} \\ \Sigma_{pose-map} & \Sigma_{pose} \end{bmatrix} \in \mathbb{R}^{3M+6 \times 3M+6} \end{aligned} \quad \text{(III.13)}$$

As a result, the algorithm described previously largely remains true, with some noteable modifications - we use the full covariance matrix for both mapping and odometry. We also update the covariance of the landmarks during the predict step to account for correlations between the landmark and pose. Hence, the predict step is now:

$$\Sigma_{t+1|t} = F_t \Sigma_{t|t} F_t^T + W, \quad \text{(III.14)}$$

where $F = \begin{bmatrix} I & 0 \\ 0 & \exp(-\curlywedge_t) \end{bmatrix}$

F. Hyperparameters

We initialise the covariance of landmarks $\Sigma_m = 2$, the covariance of the pose $\Sigma_x = 0.0005$, as pose estimates are more reliable. The observation noise is set to be 5 and the motion model noise is set to be 0.001 for both the linear and angular parts of the motion.

IV. Results

The result of the EKF SLAM can be seen in figure 1 which compares the trajectory obtained without any update step (dead reckoning), and the trajectory obtained using the observation model.

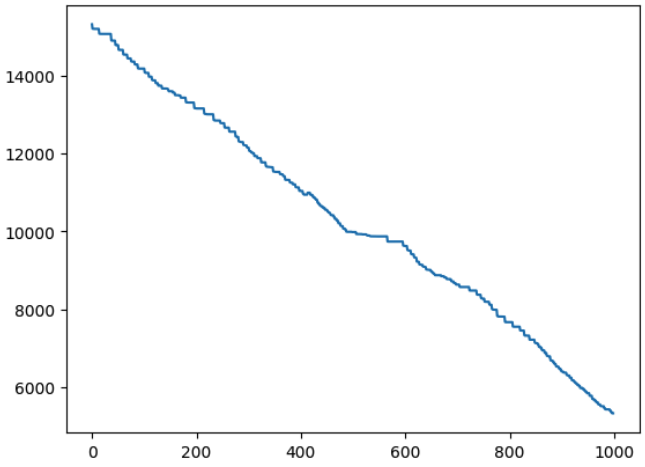

Fig. 2: The trace of the covariance matrix over many timesteps

V. Discussion

The Update Step Removes Uncertainty. As seen in figure 2, we see that as the SLAM algorithm runs for many timesteps, the overall uncertainity in the system drops significantly. This also helps confirm the implementation of the algorithm.

Outlier Rejection is Essential. We observe that without rejecting landmarks that are too far away or residual terms that are very large, the trajectory can be quite erratic, as these points have an undue influence on the trajectory. Another strategy may be to use an information prior on the points, where we weigh points near the robot more than ones far away.

Low Noise for Landmarks. Since the landmark estimates are noisy, initialising the covariance with small values causes the Kalman gain to be quite large. As a result, the trajectory is quite erratic. Careful hyper-parameter tuning was essential to get the algorithm working.