ECE 276A Project 3: Visual-Inertial SLAM

Winston Chou  
*PID: A17460970*

*Abstract*—This project focuses on implementing a visual-inertial simultaneous localization and mapping (SLAM) system using an extended Kalman filter (EKF). The system integrates measurements from an inertial measurement unit (IMU) and a stereo camera to estimate both the trajectory of the robot and the positions of static landmarks in the environment. The IMU provides linear and angular velocity data, while the stereo camera captures visual features with precomputed correspondences between left and right frames. The SLAM process involves two main steps: an EKF prediction step based on IMU kinematics to estimate the robot's pose and an EKF update step using visual observations to refine landmark positions. The project assumes known extrinsic and intrinsic calibration parameters for the sensors. Results demonstrate the effectiveness of the proposed approach in estimating accurate robot trajectories and mapping landmark positions, despite challenges such as noisy measurements and partial observability.

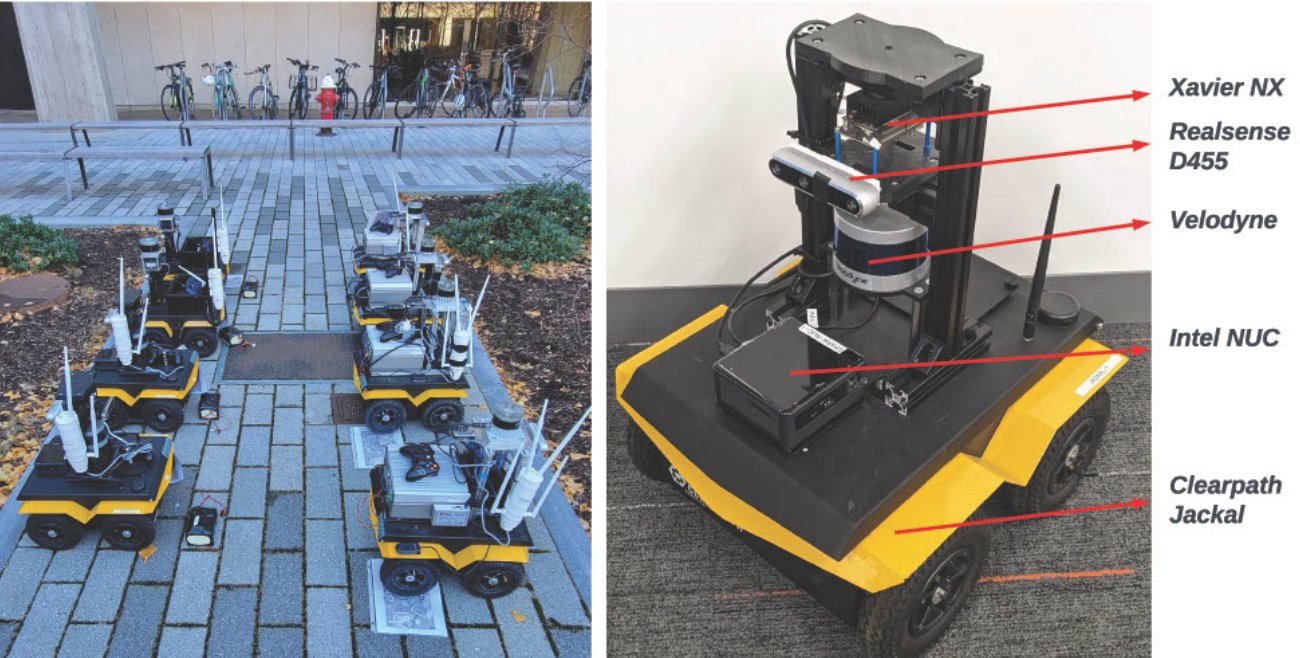
Keywords—SLAM, Markov Chain, pose estimation, multi-sensor fusion, IMU, RGBD camera

# Introduction

Simultaneous localization and mapping (SLAM) is a fundamental problem in robotics that involves estimating a robot's trajectory while simultaneously building a map of its environment. This capability is critical for autonomous navigation in various applications, including self-driving cars, drones, and robotic exploration. SLAM is particularly challenging due to uncertainties in sensor measurements, dynamic environments, and computational complexity.

This project aims to solve the SLAM problem using data from an IMU and a stereo camera mounted on a vehicle. The IMU provides linear and angular velocity measurements, while the stereo camera captures visual features used for landmark mapping. By leveraging an extended Kalman filter (EKF), the system integrates these sensor inputs to estimate both the robot's pose and the positions of static landmarks in 3D space.

The proposed approach consists of two key components: an EKF prediction step that uses IMU kinematics to estimate the robot's trajectory over time, and an EKF update step that incorporates stereo camera observations to refine landmark positions. Known extrinsic calibration between the IMU and camera frames, as well as intrinsic camera parameters, are utilized to ensure accurate sensor fusion.

This report details the implementation of the visual-inertial SLAM system, including problem formulation, technical approach, results, and analysis. The findings highlight the potential of combining visual and inertial data to achieve robust localization and mapping in real-world scenarios.

1. Sensor Setup. Clearpath Jackal robots on MIT’s campus, equipped withIMU, 3-D LIDAR scanner, and an RGBD camera, Realsense D455. [Source: ECE276A\_PR3.pdf]

# problem formulation

## Simultaneous Localization And Mapping

SLAM can be described as a probabilistic process modeled as a Markov Chain. At discrete time steps , the robot's pose  evolves based on the control input  and motion noise . The pose at the next time step  is determined by a probabilistic function:

Where represents motion noise. This relationship is commonly referred to as the **motion model**.

To map the environment, the robot collects observations at each time step. Let the observation at time  be denoted as  and the environment as . The sensor observations follow a probabilistic relationship:

Where represents observation noise. This relationship is commonly referred to as the **observation model**.

The SLAM problem involves estimating the environment  and the robot's poses  using observations and control inputs at each time step . The goal is to compute the joint probability distribution of the environment and robot poses conditioned on the observations and control inputs. This probabilistic formulation captures the uncertainty inherent in SLAM:

The relationship between the environment and robot poses is challenging to determine directly. However, leveraging the Markov assumptions allows for decomposing the joint probability density function into manageable components. This decomposition includes terms for the initial state of the robot and environment, observation likelihoods based on sensor readings, and motion probabilities derived from control inputs and previous poses.

In practical SLAM implementations, maximum likelihood estimation (MLE) is often used to find optimal values for the robot's trajectory  and the environment . The optimization process involves maximizing the sum of log-likelihoods for sensor observations and motion probabilities over all time steps. This approach helps determine both the robot's poses and a map of the environment effectively.

## Bayesian Filtering

Bayes filtering is a probabilistic method used to estimate the state of dynamic systems, such as a robot's pose, by combining information from control inputs and observations. This technique relies on the Markov assumptions and Bayes' rule to infer the system's state over time. The Bayes filter operates in two main steps:

1. *Prediction Step:*  
   In this step, the prior probability distribution of the system state at time , along with the control input, is used to predict the state at the next time step . The motion model governs this process, and the predicted probability distribution is computed by integrating over all possible states at time :
2. *Update Step:*  
   Once a new observation is received at time , the predicted probability distribution is updated using the observation model. This step incorporates measurement information to refine the estimate of the system's state at . The posterior distribution is normalized by dividing by a constant factor (marginal likelihood) to ensure it remains a valid probability distribution:

## Landmark-based Mapping

Landmark-based mapping focuses on creating a map of the environment using noisy and uncertain sensor observations, assuming the robot's poses are known. The environment is modeled as a set of static landmarks, with each landmark represented by its location in 3D space. A landmark's position is denoted as , where  ranges from 1 to , the total number of landmarks. Collectively, the landmarks are represented as a matrix , with each landmark specified by three numerical values corresponding to its coordinates.

The robot can detect landmarks at each time step , and the observations are denoted as . Since multiple landmarks may be sensed simultaneously,  represents a composite observation encompassing all detected landmarks at that time.

The goal of this mapping process is to estimate the locations of the landmarks based on the robot's pose  and the observations . This estimation relies on an observation model, which defines the probabilistic relationship between the observations, robot pose, environment landmarks, and measurement noise.

An index map  is used to track which landmarks correspond to specific observations. At each time step , the robot observes  landmarks, with each observation denoted as  (Homogeneous Coordinates), where . The data association map specifies that the  landmark corresponds to the observation indexed by .

## Sensors Setup

The proposed solution addresses the SLAM problem using data from an IMU and a stereo camera mounted on a vehicle. The IMU provides measurements of linear velocity () and angular velocity (), both expressed in the IMU's frame of reference. The stereo camera captures visual data, with precomputed visual features and correspondences established between the left and right camera frames as well as across time steps (data association).

At each time step , the visual features are represented as , where each column corresponds to a landmark. Specifically, the  column contains the pixel coordinates of the  landmark in both the left and right camera images. If a landmark is not observable at time , its corresponding column in  is set to .

The system assumes that both followings are known:

1. *The transformation from the IMU frame to the stereo camera's optical frame* ( ) *is known (extrinsic calibration)*
2. *The stereo camera calibration matrix () is also known (intrinsic calibration). The calibration matrix is defined as:*

Where is the focal length, , are pixel scaling, and are the principal points, and is the stereo baseline.

# Technical approach

The implementation of the project is presented as follows:

## Extended Kalman Filter

Extended Kalman Filter (EKF) is a nonlinear extension of the Kalman Filter designed to handle systems with nonlinear dynamics and observations. It operates by linearizing the system around the current estimate of the mean and covariance using a moment-matching approach. The EKF is fundamentally a Bayes filter, relying on several key assumptions:

* The prior probability density function (pdf), , is Gaussian.
* The state transition model is affected by Gaussian noise, expressed as:
* The observation model is also influenced by Gaussian noise:
* The process noise and measurement noise are independent of each other, the state , and across time steps.
* The posterior pdf is approximated as Gaussian using moment matching.

The primary challenge in applying the EKF to nonlinear systems lies in the fact that the predicted and updated pdfs are not inherently Gaussian and cannot be computed in closed form. To address this, moment matching is employed to approximate these pdfs as Gaussians by evaluating their first and second moments (mean and covariance). This ensures that the EKF maintains a consistent probabilistic framework while handling nonlinearity effectively.

The Extended Kalman Filter (EKF) uses a first-order Taylor series expansion to approximate the integrals required for implementing a nonlinear Kalman Filter. The motion model is approximated as follows:

Where:

Similarly, the observation model is approximated as:

Where:

Based on these approximations, the EKF models are defined as follows:

**Prior:**

**Motion Model:**

The state transition includes Gaussian process noise:

Linearization parameters:

**Observation Model:**

Observations are affected by Gaussian measurement noise:

Linearization parameters:

**Prediction Step:**

Predicted Mean and Covariance

**Update Step:**

Kalman Gain

Posterior Mean & Covariance

## Visual Mapping: Landmark Mapping via EKF update

In the context of the landmark-based visual mapping problem, it is assumed that the inverse IMU pose  is known. Additionally, the landmarks are assumed to be static, and the data association —which specifies which landmarks are observed at each time step — is pre-computed by an external algorithm.

The observation model incorporates a Gaussian prior and observation noise and is expressed as follows:

**Prior:**

**Observation Model:**

Homogeneous Coordinates:

Where:

* represents the expected locations of all landmarks.
* is the covariance matrix of the estimate.
* is the calibration matrix mentioned in Problem Formulation

The projection function  and its derivative are defined as:

All observations at time step  can be stacked into a single vector:

where is the Kronecker product.

For the Extended Kalman Filter (EKF), the derivative of the observation model with respect to the landmark positions must be computed. Considering a small perturbation  for the landmark :

Using a first-order Taylor series approximation, the observation model for feature  at time step  becomes:

Where is defined as .

**EKF Update Steps:**

The EKF update steps for landmark mapping are as follows:

Predicted Observations:

for all observed features .

Observation Matrix:

The full matrix is stacked as: .

Kalman Gain:

State Update:

Since landmarks are assumed to be static, there is no prediction step for their positions. Each landmark's position is initialized when it is first observed using:

## EKF-Based Visual-Inertial Odometry

The localization problem aims to estimate the inverse IMU pose of the robot, , using IMU measurements , visual feature observations , and landmark coordinates  in the world frame. Similar to the visual mapping problem, data association between observations and landmarks is assumed to be pre-computed by an external algorithm.

The motion model incorporates a Gaussian prior and process noise:

**Prior:**

**Motion Model:**

where:

* combines linear and angular velocity.
* is the time difference between two consecutive frames.
* represents Gaussian process noise.

To separate the deterministic motion from noise, discrete-time perturbation techniques are used to rewrite the motion model in terms of nominal kinematics and zero-mean perturbations:

Where

With the motion model defined, the Extended Kalman Filter prediction steps are listed as follows:

**Observation Model:**

The observation model is consistent with the visual mapping problem. For each observation at time step , a first-order Taylor series approximation is applied to account for perturbations in the inverse IMU pose:

Where for any homogeneous coordinates

**EKF Update Steps:**

The EKF update steps for visual-inertial odometry are as follows:

Predicted Observations:

for all observed features .

Observation Matrix:

The full matrix is stacked as:

Kalman Gain:

State Update:

## EKF-Based Visual-Inertial SLAM

To simultaneously estimate the robot's pose and the positions of landmarks, the proposed approach merges the prediction and update steps from EKF-based visual mapping and visual-inertial odometry. The joint state and covariance are defined under the Gaussian assumption as follows:

where is the estimated landmark position and is the estimated six degrees of freedom of IMU robot pose.

**Prediction Step:**

Since landmarks are assumed static, the prediction step only applies to the IMU pose. The joint state and covariance are updated using the IMU measurements  as follows:

where Wp is the process noise covariance.

**Update Step:**

The update step combines the visual mapping update with visual-inertial odometry. The equations are as follows:

Predicted Observations:

for all observed features .

Observation Matrices:

The full matrix is stacked as:

Kalman Gain:

State Update:

This combined EKF approach leverages both IMU measurements for pose prediction and stereo camera observations for landmark updates, enabling simultaneous estimation of robot trajectory and environmental mapping.

# results

***Dataset 20***

*Motion Model:* The motion model provided a solid baseline trajectory (Fig. 1a). While it is not perfect, it serves as a reliable starting point for further optimization.

*Observation Model:* The observation model improved upon the motion model by refining details.

*Factor Graph Without Loop Closure:* Not Implemented

*Factor Graph With Loop Closure:* Somehow did not work will with dataset 20, could be poor loop closure detections due to dense overlap between scans.

***Dataset 21***

*Motion Model:* The motion model struggled in areas such as the vertical hallway, introducing errors in the trajectory (Fig. 2a).

*Observation Model:* The observation model failed to correct these errors and instead converged to a local minimum, worsening the trajectory (Fig. 2b). This highlights the sensitivity of ICP to poor initial guesses.

*Factor Graph Without Loop Closure:* Not Implemented

*Factor Graph With Loop Closure:* Loop closure constraints helped correct some errors in the x-coordinate but introduced orientation inconsistencies, leading to unreliable mapping results (Fig. 2d).

***Key Observations***

*ICP Sensitivity:*

Both the standard ICP algorithm and its modified version are prone to converging to local minima. A good initial guess for translation and rotation is critical for achieving accurate results. Using multiple initial guesses with varying yaw angles, as done in this project, helps mitigate this issue.

***Potential Improvements:***

Equipping the robot with a 3D LiDAR or utilizing a full 3D point cloud from an RGB-D camera could enhance ICP performance by providing more data points and an additional dimension for alignment.

These enhancements would reduce reliance on precise initial guesses and improve robustness against local minima.

***Conclusion***

The results demonstrate that while motion models provide a reliable starting point, observation models and factor graphs with loop closure constraints are essential for refining trajectories and correcting drift. However, ICP's sensitivity to initial guesses remains a limitation, particularly in challenging environments like Dataset 21. Future work should explore integrating richer sensor data (e.g., 3D LiDAR) to enhance robustness and accuracy in scan matching and trajectory estimation.

1. Dataset 20
2. Dataset 21
3. ICP warm up images

##### References

1. N. Atanasov, UCSD ECE276A: Sensing & Estimation in Robotics (Winter 2025), https://natanaso.github.io/ece276a/schedule.html.