Real Time Visual Localization and Mapping of Mobile Robot in Dynamic Environment

Nischal Maharjan, Rashik Shrestha, Sajil Awale, Shrey Niraula
Supervised By: Jitendra Kumar Manandhar

Department of Electronics and Computer Engineering
IOE Pulchowk Campus, Tribhuvan University
Pulchowk, Lalitpur

Abstract

Robot localization is an integral part in mobile robotics. It is the base for path planning and navigation tasks for robot and also for AR/VR applications. SLAM has been an well known method for mapping the unknown environment and localizing yourself in the map. Visual SLAM uses visual sensors such as camera to perform SLAM. It is one of the most researched topics in mobile robotics and visual sensors are too cheap nowadays. This project uses camera as its only sensor to build 3d map of entire room and localize yourself in the built map. The map can then be used for navigation purposes within the mapped environment. Various problems like dynamically changing environment, change in lighting conditions, lack of textured environment are the hindrances for visual SLAM. Some of these problems has been well tackled in this project. Dynamic portion of the environment has been masked to minimize its effect. Light invariant feature extraction has been used to tackle with difference in lightning conditions. Robot Operating System (ROS) has been used to communicate between various processes.

Objectives

1. Map unknown environment and localize a mobile robot using only monocular camera
2. Deal with moving people in environment

I. Introduction

Localizing is one of major task in robotics. Yet, today's system can not localize and navigate in 3d world as human does. Localization is finding pose in prebuilt map of the surrounding. But when you are lost, and unaware of where you in environment, you try to create the map by roaming around and at the same time an localizing yourself in that map. This is very well known Simultaneous Localization and Mapping (SLAM) problem. But, machines are not smart to figure out stationary and dynamic objects. A robust algorithm is needed to distinct between static and dynamic objects and store only those coming from permanent portion of environment in its map database.

II. Theory

Camera Projection

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = L(K\begin{bmatrix} R & t \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix})$$

L is Lens configuration of camera
K is Camera Matrix
[R t] is rotation and translation of camera
P = K[R t] in combination is termed as projection matrix.

Visual Features (ORB Extraction and Matching)

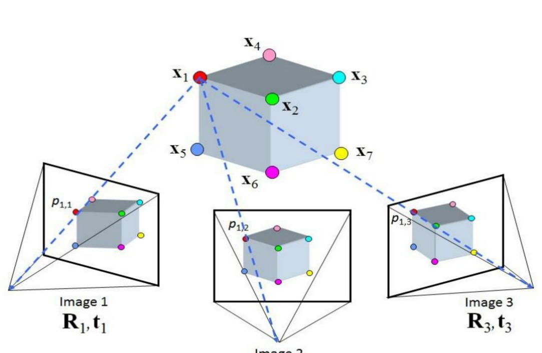
Visual Features or corner points are points in images that are invariant under change in view, different illumination and change in scale. ORB extractor uses FAST algorithm to estimate keypoints and BRIEF to compute descriptor on basis of intensity of pixels. The hamming distance between descriptor is use to find the 2D-2D correspondence points in two different images

Triangulation

Given 2D correspondences and the relative pose between 2 camera views the 3D point can be estimated by using technique known as Triangulation

$$\lambda \begin{bmatrix} x_1 \\ 1 \end{bmatrix} = P_1 \begin{bmatrix} X \\ 1 \end{bmatrix} \quad \lambda \begin{bmatrix} x_2 \\ 1 \end{bmatrix} = P_2 \begin{bmatrix} X \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} \begin{bmatrix} x_1 \\ 1 \end{bmatrix} \times P_1 \\ \begin{bmatrix} x_2 \\ 1 \end{bmatrix} \times P_2 \end{bmatrix} \begin{bmatrix} X \\ 1 \end{bmatrix} = 0$$



Linear PnP(Pose Estimation)

Given the 2D-3D correspondence the pose of camera w.r.t. world coordinate system can be computed by the process known as Linear PnP.

$$\lambda \begin{bmatrix} x \\ 1 \end{bmatrix} = P \begin{bmatrix} X \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \times \begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix} \tilde{X} = 0 \qquad \begin{bmatrix} 0 & -\tilde{X}^T & v\tilde{X}^T \\ \tilde{X}^T & 0 & -u\tilde{X}^T \\ -v\tilde{X}^T & u\tilde{X}^T & 0 \end{bmatrix}_{3\times12} \begin{bmatrix} P_1^T \\ P_2^T \\ P_3^T \end{bmatrix}_{12\times1} = 0$$

Graph Optimization

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} -log(\prod_{i.i} L(z_{ij}, P(x_i, x_j)))$$

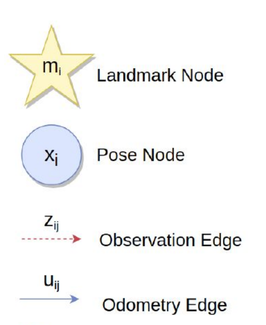
$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} -log(\prod_{i,j} \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp \frac{-(z_{ij} - P(x_i, x_j))^2}{2\sigma_{ij}^2})$$

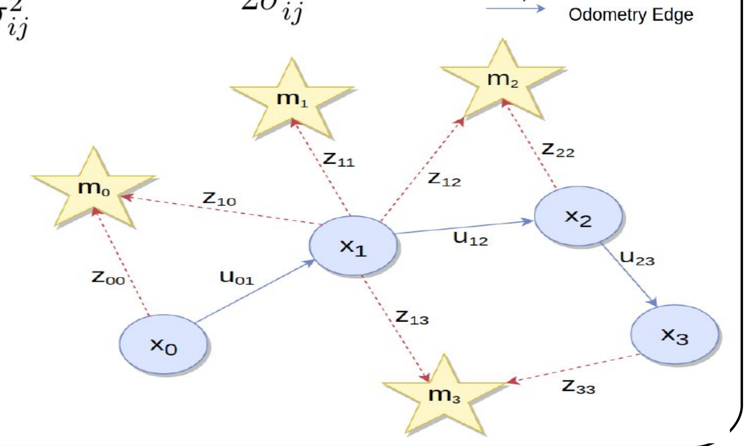
$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} F(\mathbf{x})$$

$$\Delta\hat{\mathbf{x}} = \arg\min_{\Delta\mathbf{x}} F(\check{\mathbf{x}} + \Delta\mathbf{x})$$

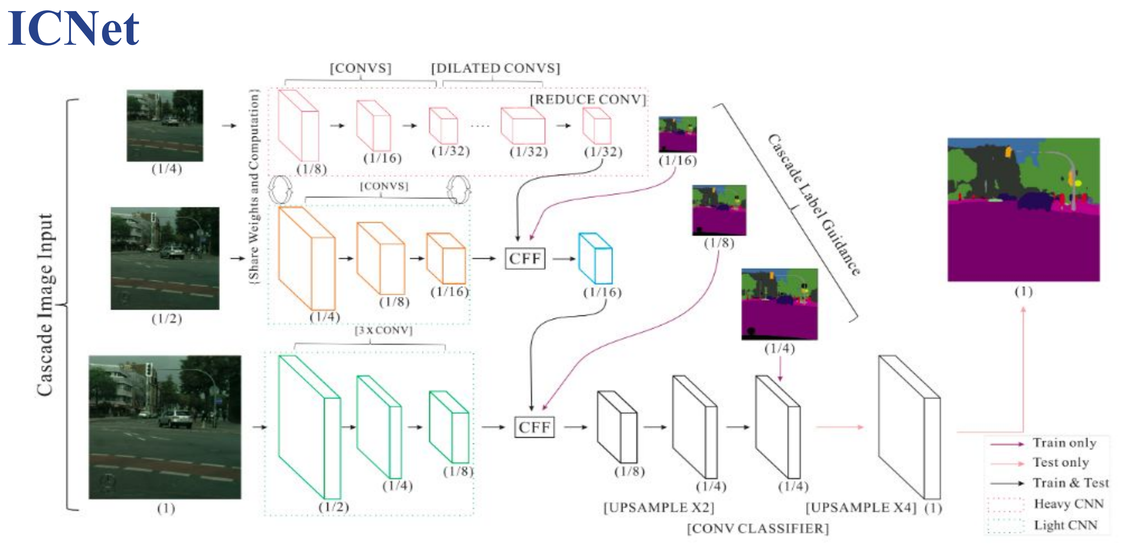
$$\left.\frac{\delta F(\check{\mathbf{x}} + \Delta\mathbf{x})}{\delta\Delta\mathbf{x}}\right|_{\Delta\mathbf{x}=\Delta\hat{\mathbf{x}}} = 0$$

$$\boxed{\hat{\mathbf{x}} = \check{\mathbf{x}} + \Delta\hat{\mathbf{x}}}$$





ICNet



ICNet is semantic segmentation model that consists of three branches of different resolution or scales (¼,1/2, original scale). The low and other branch mask results are cascaded together to achieve final mask result. ICNet achieves high speed performance by combining coarse map from low branch with finer details from medium and high resolution branch.

III. Methodology

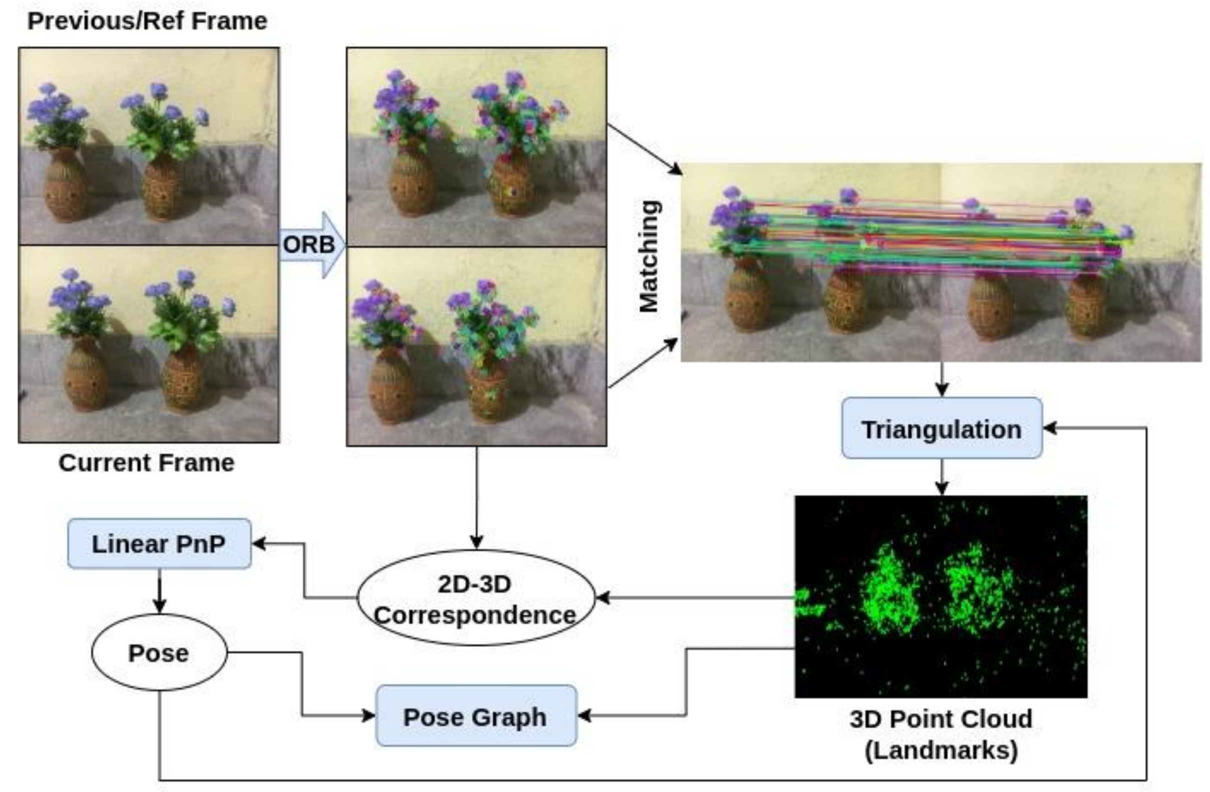


Fig: Structure from motion paradigm

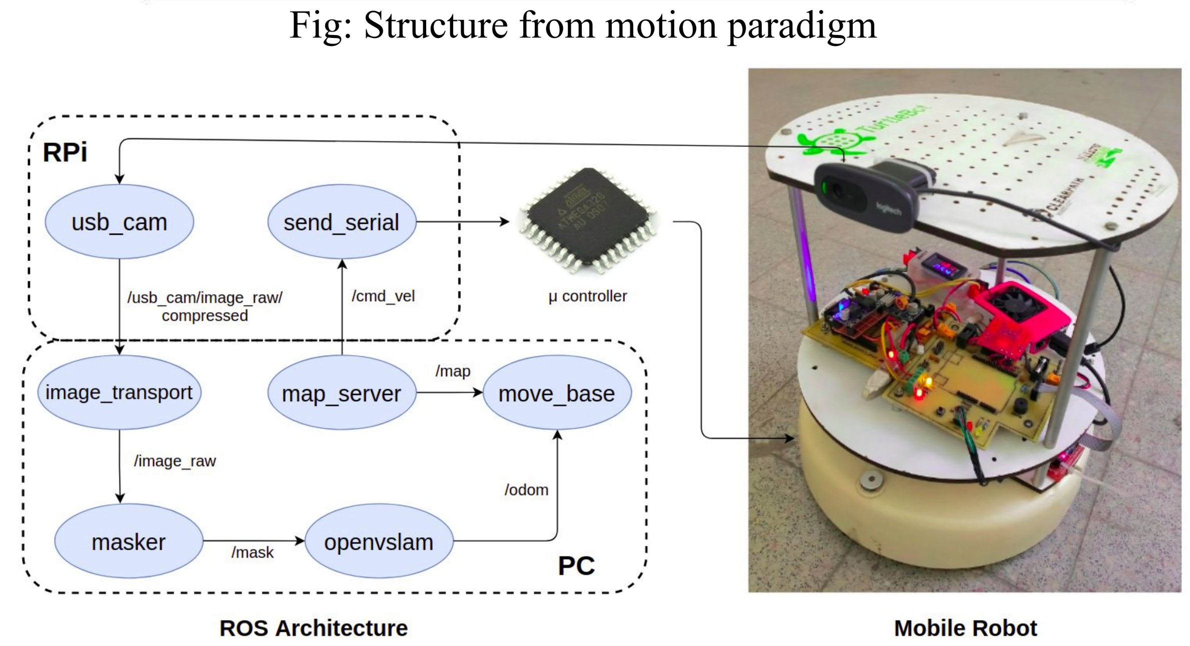


Fig: ROS Communication

Mask Generation

Model Comparison: Model selection was performed based on the mIOU vs inference speed trade off as seen in result.

Custom Dataset Generation: Custom dataset of walking was created. Multi Environment walking dataset (1435 images) and Locus Office Walking dataset (1350 images) were used for training and validation respectively.

Fig: Generation of Dataset

ICNet Training and Freezing of Layers: Resnet backbone frozen to reduce 11.4M to 400k trainable parameters

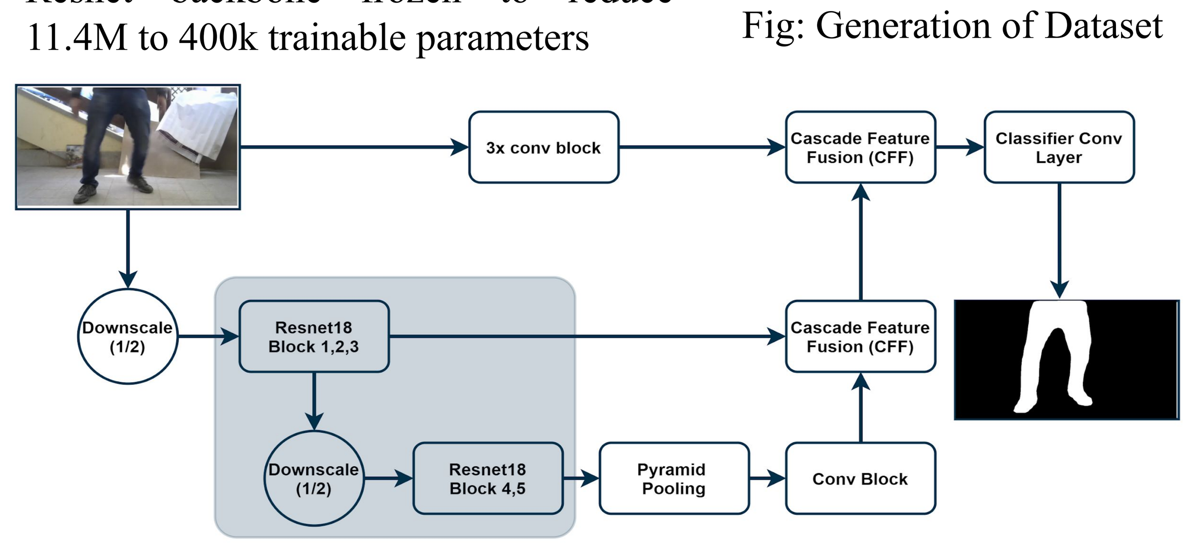


Fig: ICNet fine tuning

IV. Result and Analysis

Results in Real Environment

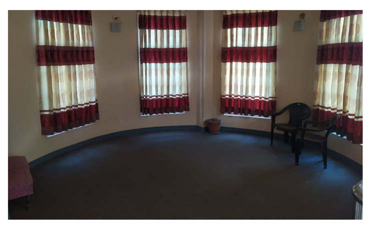


Fig: Room that is being mapped

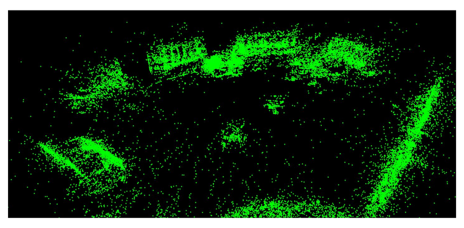


Fig: 3D Point Cloud Map Generated

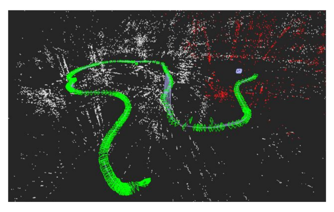
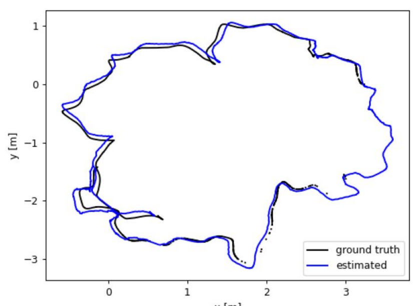


Fig: Trajectory of Camera

Results in Standard TUM Dataset



Metric	fr2_desk	fr2_slam3
Average RMS error(m)	0.097710	0.101926
Relative Translational error(m)	0.129474	0.028162
Relative Rotational error(deg)	14.376855	1.059033

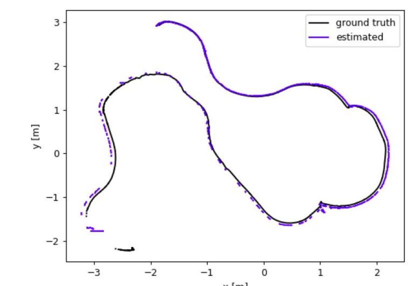
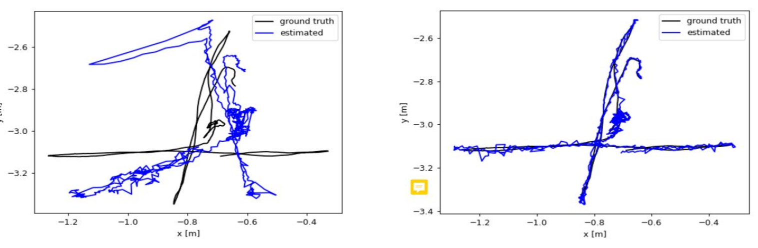


Fig: Estimated Trajectory and Ground truth of Static Environment



(a) Without Mask (b) With Mask

Fig: Effect of use of mask in trajectory estimation of Dynamic Environment.

Dataset & Methods	Validated on Locus Office Dataset	
	mIOU(%)	FPS
ICNet	80.08	26.51525
BiSeNetv1	84.09	13.71467
DeepLabV3plus	88.77	7.28928
UNetPlus	82.59	5.58920
ICnet fine-tuned(ours)	83.27	24.03161

Table: FPS vs mIOU of segmentation models

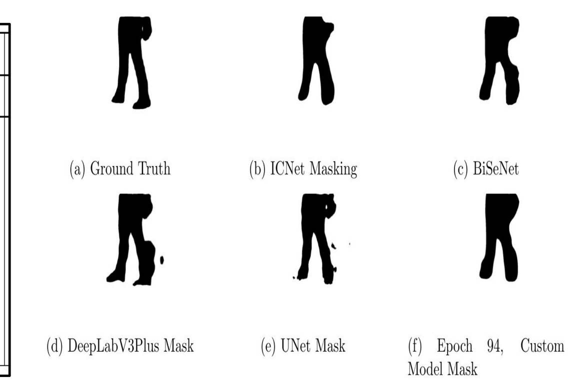
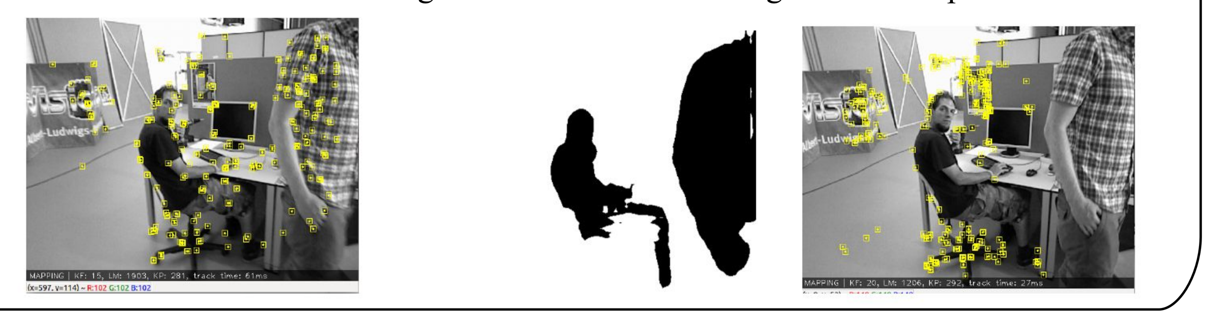


Fig: Mask Comparison



IV. Conclusion

Cheaper the price of sensors, larger the processing power needed to achieve the same accuracy. In this project an effort have been made to increase the accuracy of mapping using cheap sensor (i.e camera) and with only CPU. The accuracy might have improved drastically if expensive sensors such as 3D lidars and RADARS was used on highend GPU. But the goal of this project was not to achieve best localization accuracy, rather to develop an best algorithm which can perform well with cheap sensors and low processing power. So, it can be concluded that the project is in right track to achieve its objectives

V. References

- Shinya Sumikura, Mikiya Shibuya, and Ken Sakurada. "OpenVSLAM: A Versa-tile Visual SLAM Framework".
- Hengshuang Zhao et al. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. 2018. arXiv: 1704.08545 [cs.CV].
- Richard Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision. 2nd ed. USA: Cambridge University Press, 2003. isbn: 0521540518.

VI. Acknowledgment

We would like to profoundly thank and show gratitude to our supervisor **Mr. Jitendra Kumar Manandhar** for essential and continuous guidance, and all the teachers for their support. It has been privilege to work under his supervision. We would like to thank **Department of Electronics and Computer Engineering** for incorporating major project as syllabus for realization of our knowledge. We are grateful towards **Robotics Club** Pulchowk Campus for providing us with hardware requirements