ENPM808 Independent Study

Introspective Vision SLAM and Sensor Fusion for enhanced Localization of Robots.

Technical Report

Badrinarayanan Raghunathan Srikumar (119215418)

MEngg. Robotics
University of Maryland
College Park,
13th May 2024.

ACKNOWLEDGMENT

I would like to thank my professor Dr. Pratap Tokekar for his consistent support and guidance during the course of this project. Additionally, I would also like to extend my gratitude to Vishnu Dutt Sharma for his assistance in both doubt clearance and software assistance, without which successful completion of the project would not have been possible

SL NO	TOPIC
I	INTRODUCTION
II	RELATED WORK
III	FEATURE-BASED V SLAM
IV	INTROSPECTIVE VISION
V	INTROSPECTION FUNCTION
V.A	TRAINING DATA GENERATION
V.B	GUIDED FEATURE EXTRACTION
V.C	RELIABILITY-AWARE BUNDLE ADJUSTMENT
VI	EXPERIMENTAL SETUP AND RESULTS
VII	FUTURE WORK
VIII	REFERENCES

I. INTRODUCTION:

Visual Simultaneous Localization and Mapping (V-SLAM) involves extracting features from observed images and establishing correspondences between these features over consecutive frames and using that to estimate the pose of the camera and the 3D location of the features. Through joint optimization of feature reprojection error and motion data from odometry or IMUs, V-SLAM reconstructs a robot's trajectory and a sparse 3D map of feature locations in the environment. However, ensuring accurate tracking and mapping necessitates the selection of features from static objects and consistent identification of feature correspondences. Despite significant efforts to filter out erroneous features or reject improbable correspondences, V-SLAM solutions still suffer from errors arising from incorrect feature matches and features extracted from moving objects. Additionally, V-SLAM assumes that reprojection errors are independent and identically distributed (i.i.d), a flawed assumption as features from low-contrast or repetitive-textured regions exhibit wider error distributions compared to those from sharp, unique corners. Consequently, even state-of-the-art V-SLAM solutions encounter catastrophic failures in challenging real-world scenarios such as specular reflections, lens flare, and moving object shadows, underscoring the limitations of current methodologies.

Hence a novel approach, Introspective Vision for SLAM (IV-SLAM), was introduced that significantly varied from conventional V-SLAM solutions. The algorithm uses a function approximator to learn a predictive function that determines a reliability score of features, which aids in differentiating a good feature from a bad feature. Consequently, it enhances feature extraction by extracting features for further processing only from areas of the image with a high feature reliability score. It also enhances Bundle Adjustment, by building a context-aware total noise model that explicitly accounts for heteroscedastic noise, and learns to account for bad correspondences, moving objects, non-rigid objects, and other causes of errors.

By leveraging the capabilities of IV-SLAM, the objective of this project was to use the algorithm on scenarios where scene features vary across different angles, developing methods to intelligently merge disparate features, thereby enhancing overall feature extraction efficiency and consequently improving the localization of the robots.

II.RELATED WORK:

ORB-SLAM: ORB-SLAM is a feature-based monocular SLAM system that operates in real-time, in small and large, indoor and outdoor environments. The system is robust to severe motion clutter, allows wide baseline loop closing and relocalization, and includes full automatic initialization. It uses the same features for all SLAM tasks: tracking, mapping, relocalization, and loop closing. It selects the points and keyframes of the reconstruction, which leads to excellent robustness and generates a compact and trackable map that only grows if the scene content changes, allowing lifelong operation. One of the main design ideas in the system is that the same features used by the mapping and tracking are used for place recognition to perform frame-rate relocalization and loop detection. This makes the system efficient and avoids the need to interpolate the depth of the recognition features from near SLAM features. ORB was chosen, which are oriented multi-scale FAST corners with a 256-bit descriptor associated. They are extremely fast to compute and match, while they have good invariance to viewpoint.

IV-SLAM builds on top of ORB-SLAM, wherein it includes an introspection function that chooses the best set of features in a frame to be used for further processing.

1) It equips V-SLAM with an introspection function that learns to predict the reliability of image features.
2) It modifies the feature extraction step to reduce the number of features extracted from unreliable regions of the input image and extract more features from reliable regions.
3) It modifies the BA optimization to take into account the predicted reliability of extracted features.

III.FEATURE-BASED V-SLAM:

In Visual SLAM, the camera projection matrix P and the 3D scene point T are obtained based on the features extracted from each frame. Bundle Adjustment is a non-linear optimization performed to update P and T by minimizing reprojection error for every frame. This is given by:

$$\epsilon_{t,k} = z_{t,k} - \bar{z}_{t,k}$$

Where $z_{t,k}$ is the actual observation and $\bar{z}_{t,k}$ is the predicted observation. $\bar{z}_{t,k}$ is given by:

$$\bar{z}_{t,k} = K[R|t]z_{t,k}$$

Where K is the camera matrix, R is the rotational matrix and t is the translation vector. The reprojection error $\epsilon_{t,k}$ is minimized using Bundle Adjustment where:

$$P,T = argmin_{P,T} = \sum L(\epsilon_{t,k}\Sigma_{t,k}\epsilon_{t,k}^{T})$$

$\Sigma_{t,k}$ is the covariance matrix associated with the scale at which a feature has been extracted. The choice of loss function L plays a key role in the performance of the system.

IV.INTROSPECTIVE VISION:

IV-SLAM models the observation error distribution to be dependent on the observations. This is done by designing L to specifically be as follows:

$$L_{\delta(z)}(x) = \{x \quad if\ x < \delta(z)\}$$
$$L_{\delta(z)}(x) = \left\{2\delta(z)(\sqrt{x} - \delta(z)/2) \quad otherwise\right\}$$

$x \in [0,\infty)$ is the squared error value, and $\delta(z)$ is an observation-dependent parameter of the loss function and it quantifies the reliability of the observations. IV-SLAM uses the introspection function to estimate the reliability score, which can be used to calculate $\delta(z)$, such that the choice of loss function serves well for Bundle Adjustment.

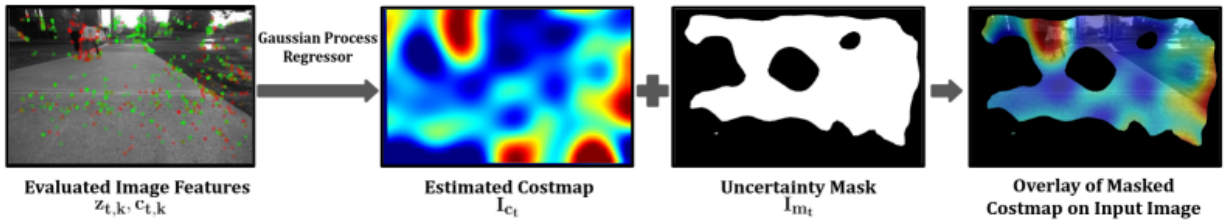
During the training phase, input images and estimated observation error values are used to learn to predict the reliability of image features at any point on an image. During the inference phase, $\delta(z)$ is estimated for each observation using the predicted reliability score, where a smaller value of $\delta(z)$ corresponds to an unreliable observation. Using x and $\delta(z)$, the loss function is chosen and consequently, this loss function is used to solve for Bundle Adjustment.

V.INTROSPECTION FUNCTION:

To apply a per-observation loss function $L_{\delta(z)}$, IV-SLAM learns an introspection function that given an input image I_t and a location (i, j) on the image, predicts a cost value $c_t(i,j) \in [0,1]$ that represents a measure of reliability for image features extracted at $I_t(i,j)$. Higher values of $c_t(i,j)$ indicate a less reliable feature. IV-SLAM requires a set of pairs of input images and their corresponding target image feature costmaps D = $\{I_t, I_{c_t}\}$ to train the introspection function. The training is performed offline and a 3D lidar-based SLAM solution is used to intelligently provide loose supervision in the form of estimates of the reference pose of the camera P. P is only used to flag the frames, i.e if the estimated Pose and reference pose provided by the lidar-based solution is way off, those frames are not used for training.

V.A.TRAINING DATA GENERATION: The V-SLAM algorithm is run on the images and at each frame K_t, the Mahalanobis distance (d^t) of the reference and estimated pose of the camera is calculated. A χ^2 test is performed on d^t and if it fails, the current frame will be flagged as unreliable and a training label will not be generated for it.

For each frame recognized as reliable for training data labeling, reprojection error values $\epsilon_{t,k}$ are calculated for all matched image features. A normalized cost value $c_{t,k} = \epsilon^T_{t,k} \Sigma^{-1}_{t,k} \epsilon_{t,k}$ is then computed for each image feature, where $\Sigma_{t,k}$ denotes the diagonal covariance matrix associated with the scale at which the feature has been extracted. The set of sparse cost values calculated for each frame is then converted to a costmap I_{c_t} the same size as the input image. This is achieved using a Gaussian Process regressor. The generated costmaps along with the input images are then used to train the introspection function using a stochastic gradient descent (SGD) optimizer and a mean squared error loss (MSE) that is only applied to the regions of the image where the uncertainty of the output of the GP is lower than a threshold.



During inference, input images are run through the introspection function I to obtain the estimated cost maps. IV-SLAM uses I_{c_t} to both guide the feature extraction process and adjust the contribution of extracted features when solving for P and T.

V.B.GUIDED FEATURE EXTRACTION: Each image is divided into equally sized cells and the maximum number of image features to be extracted from each cell is determined to be inversely proportional to the sum of the corresponding costs of each feature within that cell. This helps IV-SLAM prevent situations where the majority of extracted features in a frame are unreliable.

V.C.RELIABILITY-AWARE BUNDLE ADJUSTMENT: For each reliable feature extracted from the image, the $\delta(z)$ value is calculated as:

$$\delta(z) = \frac{1 - c_{t,k}}{1 + c_{t,k}} \delta_{max}$$

Where δ_{max} is a threshold value which has been set as χ^2 distribution's 95th percentile, i.e. 7.82 for a stereo observation. Using $\delta(z)$, the feature-specific loss function can be chosen using the equation mentioned earlier and a reliability aware Bundle Adjustment can be performed for each frame.

VI.EXPERIMENTAL SETUP AND RESULTS:

The primary focus of this project revolves around situations where the features extracted from a scene in one field of view are superior to those extracted from the same scene in a different field of view. This discrepancy commonly arises when employing multiple robots to survey an area. Factors such as lighting conditions can lead to certain frames captured by one robot lacking reliable features compared to those captured by another robot. Consequently, crucial information may be lost, potentially hindering the accurate localization and mapping capabilities of the robots. To address this challenge, communication of frames between the robots is proposed. By sharing data, particularly when one robot captures a scene with high-quality features while another captures the same scene from a different perspective with poor feature quality, the robots can leverage each other's strengths. This collaborative approach enables more accurate localization and mapping, as the combined feature points enhance the robots' ability to precisely determine their positions relative to the environment.

To identify frames suitable for this collaborative process, videos of the lab were recorded from various starting positions using a ZED stereo camera. The reliability of features was assessed using the introspection function, which assigned scores to the extracted features. Frames were selected where the features exhibited high reliability in one field of view but received low scores when observed from the other field of view. These frames serve as optimal candidates for data sharing between the robots, maximizing the effectiveness of their collaborative localization and mapping efforts.

In the image provided, features having low scores are depicted in varying shades of red. Upon examination of this image, it becomes evident that features surrounding the white table exhibit notably low scores. This observation indicates that these features are unreliable and potentially serve as bad reference points for localization and mapping tasks. It can also be noted that almost 50% of the feature points detected in this frame are unreliable.

Contrastingly, in this image, features have high scores are represented by brighter shades of green. Upon examination of this image, it becomes evident that features surrounding the white table exhibit notably high scores. This observation indicates that these features are reliably detected and can potentially serve as valuable reference points for localization and mapping tasks.

The contrast between the red and green shades allows for a clear distinction between features of differing reliability levels. This visual representation enables researchers or operators to quickly identify areas of interest where features are robustly detected, as indicated by the bright green hues. Conversely, regions with poorer feature scores are easily discernible through the presence of deeper red tones. By leveraging this visual feedback, decision-making processes related to feature selection and data utilization can be significantly enhanced. Specifically, areas highlighted in green can be prioritized for further analysis or utilized as reference points for guiding robot navigation or mapping activities. Conversely, areas depicted in red may prompt adjustments in robot positioning or sensor configurations to improve feature detection and reliability in those regions. Ultimately, the detailed color-coded representation facilitates more informed and efficient decision-making in the context of robot perception and navigation tasks.

VII.FUTURE WORK:

The next phase of this research will involve leveraging the identified feature disparity between different fields of view to facilitate communication and fusion of data between the robots, ultimately enhancing their localization and mapping capabilities. This will entail developing algorithms and communication protocols that enable the exchange of feature information between robots capturing the same scene from different perspectives.

Firstly, methods will be devised to establish reliable communication channels between the robots, allowing them to share feature data efficiently and securely. This may involve designing communication protocols tailored to the specific requirements of feature data transmission, considering factors such as bandwidth constraints, latency, and data integrity.

Subsequently, fusion algorithms will be developed to integrate the shared feature data from multiple robots into a unified representation of the environment. These algorithms will need to account for variations in feature quality and reliability across different fields of view, ensuring that the fused data accurately reflects the true characteristics of the environment.

Moreover, techniques for effectively combining feature information from multiple sources will be explored, taking into consideration factors such as spatial alignment, feature correspondence, and uncertainty estimation. This will involve developing robust fusion algorithms capable of handling discrepancies and inconsistencies in the feature data while preserving the overall accuracy of the localization and mapping estimates.

Refer to IV-SLAM GitHub repository to download and install the necessary libraries and dependencies.

The custom data that has been used for this work can be downloaded from this link.

The calibration configuration for the ZED camera can be found in this link.

VIII.REFERENCES:

[1] Sadegh Rabiee and Joydeep biswas. IV-SLAM: Introspective Vision for Simultaneous Localization and Mapping. In *IEEE Computer Society Conference On Computer Vision and Pattern Recognition*, 2020

[2] Raul Mur-Artal, J. M. M. Montiel and Juan D. Tardos. ORB-SLAM: a Versatile and Accurate Monocular SLAM System. In *IEEE Computer Society Conference On Computer Vision and Pattern Recognition*, 2015

[3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *International Conference on Computer Vision,* 2011

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision,* 2004

[5] Hao Wu, Xinxiang Zhang, Brett Story, and Dinesh Rajan. Accurate Vehicle Detection Using Multi-camera Data Fusion and Machine Learning. In *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2019