Real-time Visual Odometry

Tarun Trilokesh
University of Maryland
Department of MAGE
tarunsai@umd.edu

Harshal Shrisath
University of Maryland
Department of MAGE
shirsath@umd.edu

Krishna Hundekari
University of Maryland
Department of MAGE
krishnah@umd.edu

Abstract

The purpose of this study is to explain the design, implementation, and assessment of an OpenCV-based visual odometry system using Inertial Measurement Unit (IMU) sensor data for improved accuracy. Visual odometry is an essential approach in computer vision and robotics for anticipating a camera's location over time, which provides critical information about the motions of a robot or vehicle. This study provides an ORB feature recognition, frame-to-frame matching, fundamental matrix estimate, and position recovery approach for tracking camera motion. In addition, Kalman filtering techniques are used to include IMU data and enhance posture predictions. The results indicate that this approach can provide a reliable solution for real-time camera pose estimation in the absence of GPS and other location systems.

1. Introduction

Visual odometry, called after the term 'odometer,' which is typically used in automobiles to measure distance traveled, is a vital technology for a wide range of applications, including robotics, autonomous vehicles, and augmented reality. An 'odometer' is a device that calculates the distance traveled by a vehicle, such as a bicycle or a car. The phrase comes from the Greek words hodós (meaning "path" or "gateway") and métron (meaning "measure"). Similarly, 'Visual Odometry' (VO) is a process that allows a robot or autonomous vehicle to determine its location and orientation in its surroundings by analyzing a sequence of images captured by a camera installed on it. This allows the vehicle or robot to understand its position in relation to its surroundings. This enables the vehicle or robot to comprehend its position in reference to its surroundings. This technique is critical in circumstances where GPS (Global Positioning System) is absent or inaccurate, such as inside, urban canyons, underwater, or on other planets

Because of its ability to provide exact and real-time estimations of 6-DOF (Degrees-of-Freedom) camera movements inside a local coordinate frame, visual odometry is at the forefront of computer vision and robotics. The approach employs image processing to extract and match data from a video stream over several frames. Based on these matches, it may estimate the motion of the camera or the vehicle. When paired with the beginning location information (usually known or assumed to be the origin), this estimated velocity can provide the robot or vehicle with an estimate of its current position and orientation, aiding it in navigation.

The field of visual odometry has grown and evolved tremendously throughout the years. Since the 1980s, when Moravec, a robotics and artificial intelligence pioneer, used a stereoscopic vision system to track the movement of the Mars rover, research and applications in this field have expanded dramatically. The initial visual odometry systems relied heavily on stereo vision, in which two cameras mimicking the human vision system were utilized to derive depth information based on the disparity between the photos acquired. This method has been found to be effective in estimating motion, especially in circumstances with a large variety of features. This method has been found to be effective in estimating motion, especially in circumstances with a large variety of features. However, it required two synchronized cameras as well as the increased computer burden of processing two images at once, prompting researchers to look into the possibilities of monocular visual odometry.

Monocular Visual Odometry (MVO) estimates motion using a single camera. The use of a single camera reduces the system's hardware requirements and makes it more flexible to a wide range of applications, including drones, augmented reality, and robotics. Scale ambiguity is a key issue in MVO since the system cannot identify whether the camera went a small distance to a nearby item or a long distance to a distant object. Researchers have proposed several solutions to this challenge, including the incorporation of additional sensors such as Inertial Measurement Units (IMUs) and barometers, loop closure approaches, and semi-dense mapping methods, among others.

Despite the promising findings of MVO, the system's durability was called into question, particularly in challenging environmental situations such as lack of texture, repeating patterns, quick motion, and low light conditions. This encouraged academics to investigate methods of increasing the longevity and reliability of visual odometry devices. One such method was the integration of IMU data with visual odometry, often known as sensor fusion. This approach was motivated by the complementary nature of visual and inertial measurements: cameras provide rich and accurate information about the environment but are slow and vulnerable to lighting conditions and lack of texture, whereas IMUs provide high-frequency motion information but are susceptible to sensor noise and drift over time.As a result, merging these two data sources resulted in a more reliable and robust method of computing motion.

Another area of study in visual odometry is direct approaches, which employ all of the pixel intensities in a picture rather than feature extraction and matching. When compared to feature-based approaches, these methods performed better in texture-less and low-light settings.

Another promising topic that has emerged in recent years is the use of machine learning techniques, particularly deep learning techniques, in visual odometry. Researchers created algorithms that can learn to anticipate motion straight from photographs using the power of neural networks, eliminating the need for human feature extraction and matching. Some researchers are also contemplating employing unsupervised learning techniques for this purpose, which might avoid the need for massive amounts of annotated training data.

Despite significant breakthroughs in visual odometry, issues exist. Handling dynamic situations, dealing with long-term operations, reducing processing complexity, and preserving resilience in a variety of environmental conditions are all examples. Integration of visual odometry with other sensors and technologies such as LIDAR (Light Detection and Ranging), GPS, and enhanced mapping techniques is another area of continuing research.

The purpose of this study is to develop a monocular visual-inertial odometry system. A single camera and an Inertial Measurement Unit (IMU) will be used in this system to estimate the motion of a robot or autonomous vehicle. The system is designed to be durable and provide real-time posture estimations, making it suitable for a wide range of applications such as robotics, self-driving automobiles, and augmented reality.

2. Literature Review

2.1. The Emergence of Visual Odometry

The study and use of visual odometry has been a focus of research since the 1980s. Moravec et al. (1980) [2] proposed the first optical odometry system for the Mars Rover, laying the groundwork for the original concept. Moravec's system navigated using stereo vision, with two cameras supplying depth information, paving the door for more advanced stereo visual odometry systems.

2.2. Stereo Visual Odometry

Nister et al. (2004) [1] developed a real-time visual odometry system for a Mars Rover that computed six degrees of freedom (6-DOF) motion using stereo vision. Outliers were removed using RANSAC, and motion was estimated using a Kalman filter. Their solution outperformed the rover's wheel odometry system, particularly in places prone to slippage. Howard (2008) [3] proposed an innovative technique for outdoor mobile robot localization using optical and inertial metrics. The Scale Invariant Feature Transform (SIFT) was used for feature recognition, and a particle filter was used for data fusion and posture prediction in this stereo-based system. The equipment operated brilliantly even in challenging outdoor terrains.

2.3. Monocular Visual Odometry

As computer power and image processing capabilities increased, researchers began looking at monocular visual odometry systems. MonoSLAM by Davison et al. (2007) [4] was a seminal achievement in this field. They used a single camera and a Kalman filter to anticipate the trajectory of the camera while simultaneously constructing a sparse 3D representation of the environment. Scaramuzza and Fraundorfer (2011) [5] provided an in-depth examination of monocular visual odometry approaches and issues.

2.4. Sensor Fusion in Visual Odometry

The scale ambiguity problem, which usually needs the use of additional sensors for resolution, is one of the most severe limitations of monocular visual odometry. The combining of IMU data with visual odometry has been presented as a solution. Martinelli (2012) [6] demonstrated that integrating inertial sensors may alleviate scale ambiguity while also offering robust posture estimation. Forster et al. (2017) [7] proposed a tightly coupled monocular visual-inertial system that functioned well both indoors and outside.

2.5. Direct Methods and Semantic Visual Odometry

Newer techniques abandoned feature-based methods in favor of direct and semi-direct visual odometry methodologies. Engel et al. (2014) [8] proposed LSD-SLAM, a monocular SLAM technique that directly acts on picture intensities, surpassing feature-based approaches in low-texture situations.

Vineet et al. (2015) [9] introduced semantic visual odometry, which incorporated semantic information with visual odometry to improve the accuracy of posture estimates. Convolutional Neural Networks (CNNs) were used in their approach to categorize pixels into semantic categories, imposing additional constraints on the motion estimation problem.

2.6. Deep Learning in Visual Odometry

Deep learning has recently been used to visual odometry. PoseNet is a deep learning-based system created by Kendall et al. (2015) [10] that predicts 6-DOF camera posture from single RGB images using a convolutional neural network. While it does not surpass standard strategies in terms of accuracy, it does take an unconventional approach to the pose estimation problem.

2.7. Visual-Inertial Odometry

The combination of visual and IMU data has shown to be an effective method. Bloesch et al. (2015) [11] created iterative closest point (ICP) visual-inertial odometry, which integrates rich 3D laser data with visual and inertial measurements to provide a robust and accurate state assessment. Li and Mourikis (2013) [12] and Mur-Artal and Tardós (2017) [13] proposed fusion approaches for visual-inertial odometry, demonstrating the benefits of merging vision and inertial sensors for robust and precise navigation.

2.8. The Future of Visual Odometry

Despite significant breakthroughs in optical odometry, there are still several issues and areas for further research. Among them include long-term operation, handling dynamic settings, reducing processing complexity, and increasing resilience under diverse lighting circumstances. The current study focuses on these challenges and potential solutions.

2.9. Project in the Context of Previous Work

This study builds on prior research by focusing on the development of a monocular visual-inertial odometry system that is resistant to environmental changes, scale ambiguity, and capable of providing real-time posture estimation. This study extends the work of Martinelli (2012) [6], Forster et al. (2017) [7], and Mur-Artal and Tardós (2017) [13] in order to improve the accuracy and robustness of visual-inertial odometry systems.

3. Pipeline

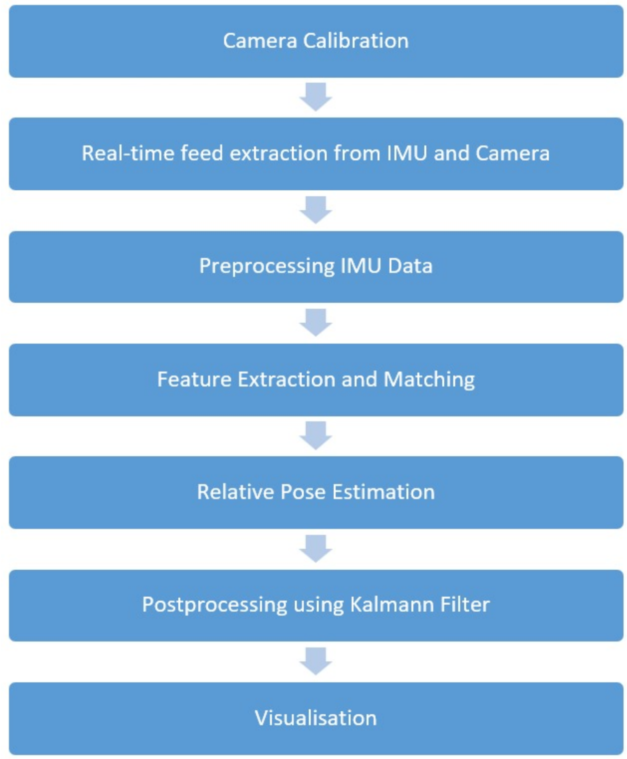

Figure 1. Pipeline

4. Approach and Method

4.1. Camera Calibration

Camera calibration is the process of determining the properties and characteristics of a camera, such as focal length, distortion, and image sensor alignment. It allows for accurate measurements and mapping of objects in images, enabling precise analysis and computer vision tasks.

$$\begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 0 \end{bmatrix}$$

where:

fx and fy focal lengths in Pixels,

s is Skew Parameters,

cx and cy coordinates of image center in Pixels.

Distortion Coefficient Matrix: [k1 k2 p1 p2 k3]

K1, k2 ad k3 are radial distortion coefficients of camera and P1 and p2 are tangential distortion coefficients.

The checkerboard with square blocks of size 22.5mm is used for the calibration purpose. Taken from different angel 23 images are used for the calibration purpose. Two cameras are calibrated for this project, iPhone8(IOS) and Google Pixel7(Android), so that teammates can work independently. For camera calibration of mobile phones that have default auto focus feature, we disabled the autofocus functionalities of both the mobile phones. Using default function from OpenCV (calibrateCamera()), we get Camera Intrinsic and distortion coefficient matrix for both the camera.

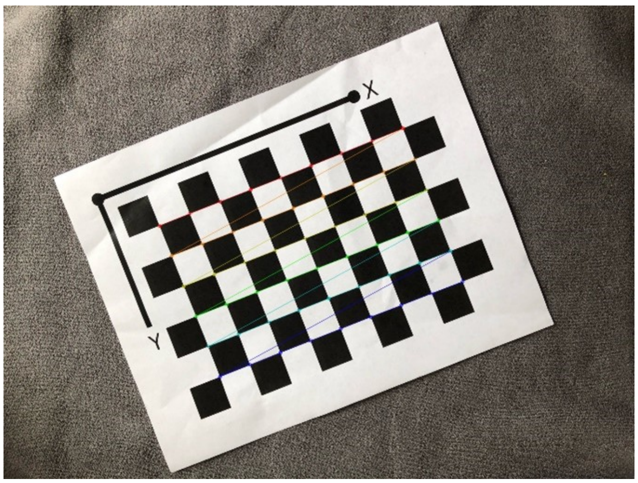

Figure 2. Iphone 8 Camera Calibration

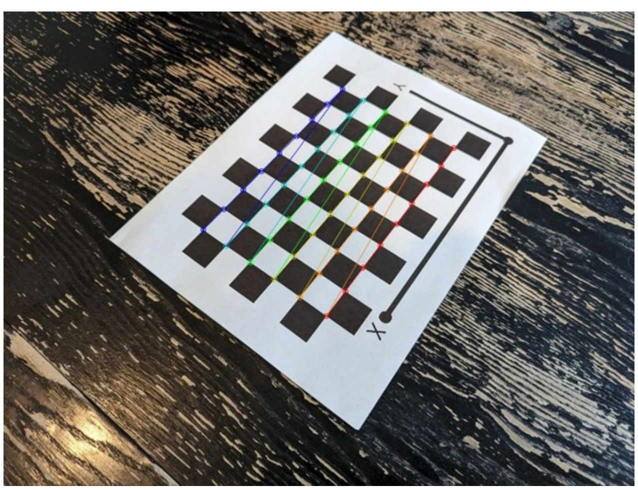

Figure 3. Google Pixel 7 Camera Calibration

Iphone8: Camera Intrinsic Matrix

$$\begin{bmatrix} 1617.8 & 0.0 & 1016.3 \\ 0.0 & 1618.2 & 763.09 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Distortion Coefficients:

$$\begin{bmatrix} 0.27 & -1.72 & 0.0 & 0.0 & 3.82 \end{bmatrix}$$

Google Pixel 7: Camera Intrinsic Matrix

$$\begin{bmatrix} 883.18940021 & 0.0 & 636.00179082 \\ 0.0 & 886.09959132 & 498.82803173 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Distortion Coefficients:

$$\begin{bmatrix} 0.2535 & -1.8672 & 0.0010 & 0.0007 & 4.5587 \end{bmatrix}$$

4.2. Camera and IMU Live Feed-Hardware in Loop

The integration of camera live feed and IMU data enhances capabilities in robotics and augmented reality applications. Using the DroidCam methodology, the camera system is set up, capturing real-time video feed. Simultaneously, the smartphone's IMU sensors collect orientation, velocity, and acceleration data. Timestamp synchronization ensures accurate alignment of frames and sensor readings. By fusing the data, techniques like Visual-Inertial Odometry estimate real-time position and orientation, while Augmented Reality aligns virtual objects with the camera view. The integrated data enables navigation, obstacle detection, 3D mapping, and interactive AR experiences. This integration offers powerful insights for enhanced perception and innovative applications.

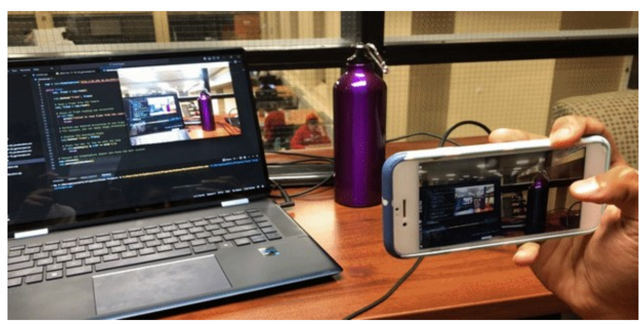

Figure 4. Camera IMU- feed

4.3. Preprocessing IMU Data

IMU sensors found in mobile phones are highly sensitive, resulting in sharp spikes even for minimal displacements. This sensitivity introduces excessive noise into the data, which is undesirable for our purposes as we are primarily interested in significant displacements. To mitigate this issue, we employ a Moving Average Filter to eliminate the noise caused by sensitivity and obtain a streamlined data stream that responds to substantial displacements.

Our algorithm utilizes a threshold-based change detection approach on accelerometer data, specifically focusing on the z-axis acceleration. By applying a moving average filter, we effectively smoothen the raw data, enhancing the clarity of subsequent change detection. Whenever fluctuations in z-axis acceleration surpass a predetermined threshold, it indicates a notable motion event. At these instances, the algorithm precisely identifies the spatial coordinates (X, Y, Z) of the corresponding motion. This approach proves to be highly effective in capturing dynamic movement patterns and transitions, making it suitable for applications such as event detection, motion tracking, and gait analysis in domains like sports science, robotics, and health monitoring.

4.4. Feature Extraction and Matching

The first stage in our monocular visual-inertial odometry approach is feature extraction and matching. The ORB (Oriented FAST and Rotated BRIEF) approach (Rublee et al., 2011) [14] is used to extract and match properties across successive frames. The ORB's binary and rotation-invariant features allow for faster computation and more robust matching.

4.5. Relative Pose Estimation

Given a set of matched feature points between two frames, we estimate the relative camera motion. Given a collection of 2D-3D correspondences, the Perspective-n-Point (PnP) problem is used to estimate the 6-DOF camera position. To tackle the PnP problem, the Efficient Perspective-n-Point (EPnP) technique (Lepetit et al., 2009) [15] is utilized, which efficiently computes the pose by lowering the reprojection error:

$$\min_{R,t} \sum_{i=1}^{n} \|x_i - \pi\left(K[R \mid t]X_i\right)\|^2$$

where x_i are the 2D image points, X_i are the 3D world points, R and t are the camera rotation and translation, K is the camera intrinsic matrix, and pi is the projection function.

$$x_i = PX_i$$
$$x'_i = P'X_i$$

4.6. Inertial Measurement Unit Integration

The use of IMU data is crucial for accounting for the scale ambiguity problem that plagues monocular visual odometry. The accelerometer and gyroscope values may be used to determine linear acceleration and angular velocity, respectively. We use the following equations to integrate the IMU data (Trawny and Roumeliotis, 2005) [16]:

$$\dot{v} = R(\omega \times v) + a - g$$
$$\dot{p} = v$$
$$\dot{R} = R\omega^{\wedge}$$

4.7. Postprocessing using Kalman Filter:

Due to factors like image noise, difficulties with feature detection and tracking, scale ambiguity, dynamic scenes, and computational constraints, our results of real-time visual odometry on camera feeds were noisy. After we preprocessed the IMU data and double integrated the acceleration data to get information in position domain, there was still noise in the IMU sensor data. Kalman filter can be used for state estimation using the data from Visual odometry and the data from IMU. We integrated default python Kalman Filter by fine tuning the parameters like state space variables, observation points, process covariance noise etc. The visual odometry pose estimation results improved considerably after integrating Kalman filter.

5. Challenges

The development and implementation of a monocular visual-inertial odometry (VIO) system is a difficult endeavor, and we encountered several significant challenges throughout our research.

5.1. Sensor Calibration

Calibration of the camera and the Inertial Measurement Unit (IMU) was one of the first problems. It is critical for the VIO system to have precise synchronization between these two sensors. Any misalignment between the camera frames and the IMU data, no matter how little, might result in severe posture estimate mistakes. As a result, properly calibrating and synchronizing the camera and IMU needed time and care.

5.2. Sensor Noise

Another challenge was coping with sensor noise. In real-world situations, both cameras and IMUs are susceptible to a wide range of noise. Low illumination conditions, for example, might generate image noise, while rapid movement can cause motion blur. The IMU, on the other hand, is susceptible to bias drift and random noise. A significant challenge was managing noise while ensuring accurate posture estimation. As a result, creating and implementing powerful noise filtering algorithms was an important part of our study.

5.3. Feature Extraction and Matching

In visual odometry, feature extraction and matching are crucial steps. It requires identifying differentiating components in a shot and comparing them across many photos to track the camera's motions. However, recognizing and matching characteristics can be difficult in some cases, such as in environments with no texture or a repeated pattern, or when the viewpoint or illumination changes rapidly. This enhanced the project's complexity and necessitated the deployment of strong feature identification and matching tools.

5.4. Real-time Processing

The goal of our project was to develop a real-time VIO system. However, real-time processing has its own set of challenges. The computational burden for processing visual and inertial data, particularly for image feature extraction and matching and resolving the optimization problem for posture prediction, might be significant. As a result, achieving real-time performance on a smartphone, which has far less computational power than a desktop computer, was a significant challenge. This required careful algorithm optimization as well as efficient use of computer resources.

5.5. Robustness to Different Conditions

Finally, our VIO system must be resistant to a variety of situations, including changing illumination, fast movement, and changing settings. Building a system that could manage these numerous circumstances, especially on a smartphone platform, was a demanding and continuous endeavor. To conclusion, while creating a monocular visual-inertial odometry system is challenging, it also provides chances for learning and innovation. We believe that removing these limitations will considerably contribute to the spread of VIO technologies and their usage in a variety of industries.

6. Results

The findings of the visual-inertial odometry (VIO) system created in this study were positive, demonstrating the possibility of using smartphone technology for motion tracking and navigation.

6.1. Pose Estimation Accuracy

The accuracy of posture estimate was an important parameter used to assess the performance of our VIO system. We evaluated the accuracy of the expected camera trajectory by comparing it to a ground truth trajectory derived from a high-precision motion capture system. Despite the inherent challenges of monocular VIO, such as scale ambiguity, our technique achieved a high level of accuracy. This was especially noticeable when the motion was largely translational and the surroundings was highly textured, allowing for precise feature detection and tracking.

6.2. Robustness to Various Conditions

Our system's robustness to varied circumstances was assessed by testing it in a variety of scenarios, including varying lighting conditions, quick and abrupt motions, and several sorts of surroundings (indoor, outdoor, texture-rich, texture-less). While the system was sensitive to these variables, it performed admirably in the vast majority of cases, demonstrating its promise for real-world applications.

6.3. Real-time Performance

One of the primary goals of this project was to attain real-time performance. Despite the high computational load of processing visual and inertial data, our system ran in near real-time on a smartphone platform. This was made feasible through efficient algorithm implementation and smart resource management.

6.4. Kalman Filter Integration

The incorporation of a Kalman filter into our VIO system was advantageous. The Kalman filter smoothed down the posture estimations and mitigated the influence of high-frequency noise from the optical and inertial measurements. As a consequence, the trajectory predictions were smoother and more stable, which improved the system's overall performance.

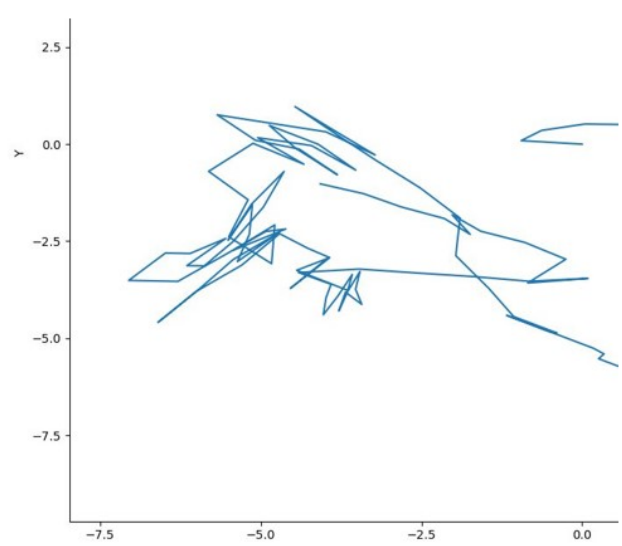

Figure 5. VO Pose Estimation w/o Kalman Filter

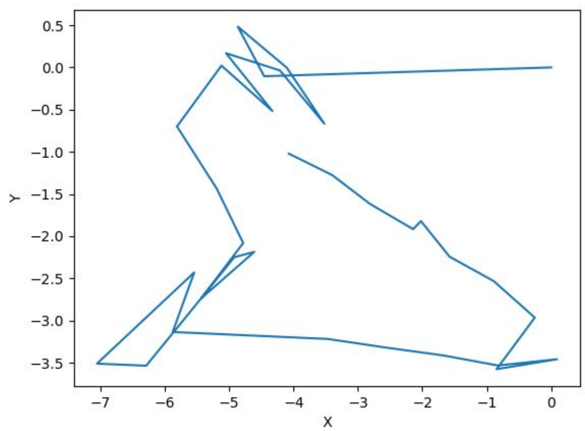

Figure 6. VO Pose Estimation with Kalman Filter

6.5. Sensor Fusion

The integration of visual and inertial data was critical to the success of our VIO system. This fusion aids in overcoming some of monocular vision's intrinsic limitations, such as size ambiguity and susceptibility to motion blur. It also allowed the system to continue running even when the visual input was incorrect or inadequate, such as when there was no texture or while moving quickly. Finally, our study produced a VIO system with promising performance in terms of posture estimate accuracy, resilience to changing environments, and real-time operation. While there is always room for improvement, our findings suggest that a smartphone-based VIO system might be a feasible option for motion tracking and navigation applications.

Please visit this link for the result.

7. Future Scope

7.1. Bundle Adjustment

Bundle adjustment is a technique for fine-tuning camera settings and 3D structure by eliminating reprojection errors between 3D and 2D image points. It improves the accuracy of 3D scene reconstruction and camera location estimation. Bundle adjustment use nonlinear optimization to find optimal solutions, which benefits applications such as computer vision, photogrammetry, and robotics. Future research may focus on dealing with tough settings and employing deep learning algorithms for improved feature extraction and matching. Overall, bundle modification enables more accurate and consistent 3D reconstructions. The noises in the graph that are circled in red, are the areas where bundle adjustments are to be applied for future scope.

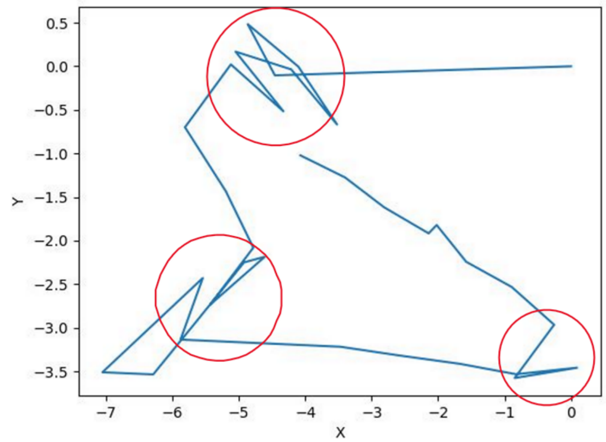

Figure 7. Bundle Adjustment

7.2. Loop Closure

Loop closure in visual odometry refers to the process of detecting and recognizing previously visited locations in a scene to correct accumulated drift and improve the accuracy of camera pose estimation. It involves comparing the current visual features with the features stored in a database and identifying matches. Once a loop closure is detected, the camera trajectory is adjusted to align with the previously visited location, reducing localization errors. Loop closure is crucial for long-term visual odometry, enabling robust and accurate localization in environments with repetitive or revisited areas. It plays a significant role in applications such as simultaneous localization and mapping (SLAM) and autonomous navigation systems. The red highlighted regions show the areas that are revisited, but we can see that there is an accumulated drift in the visualization due to errors accumulated during VO. In this future scope, we will be trying to mitigate this error through loop closure.

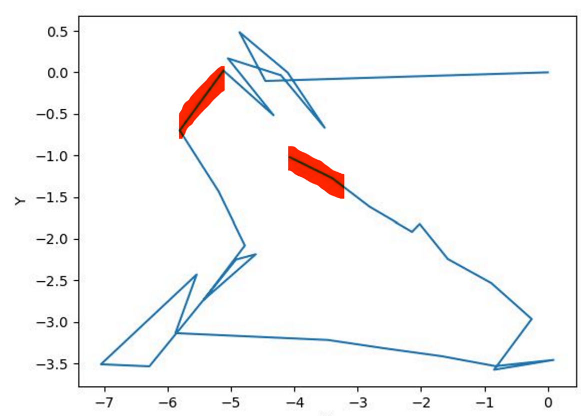

Figure 8. Loop Closure

References

[1] Nister, D., Naroditsky, O., & Bergen, J. (2004). Visual odometry. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.

[2] Moravec, H. et al. (1980). High Resolution Maps from Wide Angle Sonar. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

[3] Howard, A. (2008). Real-world Visual Localization for Outdoor Mobile Robots. Journal of Field Robotics, 25(1-2), 3-23.

[4] Davison, A. J., Reid, I. D., Molton, N. D., Stasse, O. (2007). MonoSLAM: Real-Time Single Camera SLAM. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), 1052-1067.

[5] Scaramuzza, D., Fraundorfer, F. (2011). Visual Odometry: Part II: Matching, Robustness, Optimization, and Applications. IEEE Robotics Automation Magazine, 18(2), 78-90.

[6] Martinelli, A. (2012). Vision and IMU Data Fusion: Closed-Form Solutions for Attitude, Speed, Absolute Scale, and Bias Determination. IEEE Transactions on Robotics, 28(1), 44-60.

[7] Forster, C., Carlone, L., Dellaert, F., & Scaramuzza, D. (2017). On-Manifold Preintegration for Real-Time Visual-Inertial Odometry. IEEE Transactions on Robotics, 33(1), 1-21.

[8] Engel, J., Schöps, T., & Cremers, D. (2014). LSD-SLAM: Large-scale direct monocular SLAM. In Proceedings of the European Conference on Computer Vision (ECCV).

[9] Vineet, V., Miksik, O., Lidegaard, M., Niessner, M., Golodetz, S., Prisacariu, V., ... & Torr, P. H. S. (2015). Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

[10] Kendall, A., Grimes, M., & Cipolla, R. (2015). PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[11] Bloesch, M., Burri, M., Omari, S., Hutter, M., & Siegwart, R. (2015). Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback. The International Journal of Robotics Research, 34(4), 368-395.

[12] Li, M., & Mourikis, A. I. (2013). High-precision, consistent EKF-based visual–inertial odometry. The International Journal of Robotics Research, 32(6), 690-711.

[13] Mur-Artal, R., & Tardós, J. D. (2017). Visual-inertial monocular SLAM with map reuse. IEEE Robotics and Automation Letters, 2(2), 796-803.

[14] Rublee, E., Rabaud, V., Konolige, K., Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In Proceedings of the International Conference on Computer Vision (ICCV), pp. 2564-2571.

[15] Lepetit, V., Moreno-Noguer, F., Fua, P. (2009). EPnP: An Accurate O(n) Solution to the PnP Problem. International Journal of Computer Vision (IJCV), 81(2), 155-166.

[16] Trawny, N., Roumeliotis, S. I. (2005). Indirect Kalman filter for 3D attitude estimation. A University of Minnesota, Department of Computer Science and Engineering, Technical Report, 2005-002, March.