Investigation and Implementation of Multi-Stereo Camera System Integration for Robust Localization in Urban Environments

Abhishek Rai [1], Eslam Mounier [2,3], Paulo Ricardo Marques de Araujo [2], Aboelmagd Noureldin [2,4], Kamal Jain [1]

[1] Department of Civil Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India – (abhsihek_r, kjainfce)@ce.iitr.ac.in
[2] Queen's University, Kingston, ON, Canada – (eslam.abdelmoneem, paulo.araujo)@queensu.ca
[3] Ain Shams University, Cairo, Egypt
[4] Royal Military College, Kingston, ON, Canada - nourelda@queensu.ca

Keywords: Stereo Camera, Localization, 3D Point Cloud, Multi-sensor, prior map.

Abstract

Urban environments are dynamic and complex, posing constant challenges for the localization and navigation of autonomous vehicles (AV). This demands more innovative sensor systems for effective autonomous navigation. Autonomous vehicles use sensors like LiDAR, cameras, and radar to traverse complicated urban environments with precision. These technologies have advantages in improving perception and localization, but they have their own shortcomings – LiDAR can be costly and falters under adverse weather conditions, cameras are sensitive to lighting conditions, and radars lack high-resolution details. Beyond these complexities, environmental conditions like signal-blocking skyscrapers, unpredictable obstacles, and the high costs of precision sensing add further convolution. A multi-sensor integrated solution can be a reliable option to overcome these challenges. Our work explores the use of a multi-stereo camera array that provides a 360° perception for localization in dense urban environments. We use computer vision algorithms to derive 3D point clouds from stereo-images and localize the cameras using a prior 3D map to balance cost and performance. We tested the system in Calgary's urban setting with various lighting conditions and GNSS-denied zones. Our approach provided accurate localization in 85% of the cases we tested. The results demonstrate that our multi-stereo camera system can help to achieve robust localization in challenging urban situations. This approach offers a cost-effective alternative to LiDAR-based systems while ensuring adequate accuracy.

1. Introduction

Autonomous Vehicles are the self-driven motorized means of transport with capability to perceive and navigate itself without the need of human intervention. Autonomous navigation is achieved by the integration and culmination of a group of sensors along with five functional systems which includes localization, perception, planning, control and system management (Pendleton et al., 2017). In the domain of autonomous navigation systems, precise and reliable environmental perception coupled with accurate localization capabilities have emerged as critical prerequisites for successful deployment in complex urban environments (Kutti et al., 2018; Jo et al., 2014). The conventional methodologies for localization of vehicle predominantly rely on sensing solutions such as Light Detection and Ranging (LiDAR) systems or monocular cameras, which present limitations in terms of either substantial cost implications or restricted coverage capabilities (El-Sheimy and Li, 2021).

Satellite-based navigation systems and inertial navigation systems or their fusion system are the most commonly used methods of localization in AVs. Advantage of using Global Satellite Navigation Systems (GNSS) is that it provides regular update on the global position of the vehicle. Its accuracy ranges between a few meters to a few millimetres depending on the signal strength, and the quality of the equipment used. Inertial navigation systems (INS), which uses accelerometer and gyroscope, to estimate the attitude of the vehicle, do not require external infrastructure, but it is highly prone to drifting as well as it need integration with GNSS to provide global positioning. GNSS provides good accuracy when reliable signal is recieved from the extra-terrestrial satellite constellation, but due to its dependency on the external satellites it also faces issues in areas such as indoor environments, underground tunnels and urban canyons with high-rise buildings (Gu et al., 2015). Researchers used road-matching algorithms alongside GNSS and INS to improve vehicle localization. These algorithms rely on a pre-existing road map to help guide the vehicle's estimated position. While these approaches enhance global localization accuracy, it still falls short of providing the precise positioning needed for fully autonomous driving (Najjar, 2005; Guivant and Katz, 2007).

Modern localization techniques utilize visual and LiDAR-based simultaneous localization and mapping (SLAM) to pinpoint a vehicle's location. Visual SLAM can be divided into two primary approaches: feature-based and direct methods, which differ in their error management strategies. The feature-based method performs well in static environments with abundant textures but faces challenges in dynamic settings with limited textures or when there is significant rotational movement. The direct method addresses some of these issues but is primarily effective in indoor environments. In contrast, LiDAR-based SLAM addresses the shortcomings of visual SLAM, especially in areas with low texture or variable lighting conditions. However, LiDAR sensors tend to be much more expensive and less prevalent than cameras. Another issue with SLAM is the potential for error accumulation over time, particularly if the system runs for extended periods without revisiting earlier locations or incorporating previous constraints. This becomes especially problematic in regions with weak GPS signals, complicating loop-closing optimization.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-G-2025-1255-2025 | © Author(s) 2025. CC BY 4.0 License.

Recently, improvements in computer vision and mobile mapping systems (MMS) have made high-definition (HD) maps more readily available (Nüchter, 2007; Bosse, 2012; Lin et al., 2021). Consequently, localization methods that depend on pre-existing maps have become more popular. These methods require aligning onboard sensor data with the established map, and the most efficient way to achieve this is by using the same type of sensor for both mapping and localization (Yoneda et al., 2014; Ruchti et al., 2015). Several studies have investigated the use of LiDAR for 3D mapping and localization. However, due to cost and hardware limitations, many researchers lean towards camera-based localization as a more feasible option (Stewart and Newman, 2012; Wolcott and Eustice, 2014; Neubert et al., 2017; Xu et al., 2017; Mounier et al., 2024).

Recent advancements in computer vision algorithms and three-dimensional reconstruction techniques, particularly in areas such as disparity estimation and point cloud processing, have opened new avenues for exploring multi-stereo camera configurations. Despite these technological progressions, significant challenges persist in accurately aligning and fusing data streams from multiple stereo pairs while maintaining real-time processing capabilities. Previous research contributions like Mur-Artal et al. (2017) and Levinson et al. (2007) have explored various aspects of stereo vision-based mapping and localization. Xu et al. (2017), Heng et al. (2019), Kim et al. (2018), Yabuuchi et al. (2021), and others have utilized stereo cameras for localization utilizing prior 3D point cloud maps, but their approaches either rely on a single stereo camera or focus exclusively on forward-facing stereo camera systems.

However, the implementation of single stereo camera configurations faces substantial challenges in providing comprehensive environmental coverage. The inherent limitations in the field of view and depth range associated with individual stereo cameras frequently result in incomplete scene reconstruction and potential blind spots, which could significantly compromise navigation safety and positional accuracy (Häne et al., 2017; Heng et al., 2019). The complexity is further aggravated by the unreliability of complementary systems such as GNSS due to multipath errors and INS due to error accumulation in these environments (Gu et al., 2015). However, the potential of multiple synchronized stereo cameras for comprehensive environmental perception remains relatively unexplored in the existing literature ((Häne et al., 2017; Wan et al., 2018). Lee et al. (2025) and Häne et al. (2017) have suggested the use of multiple stereo cameras to provide a 360-degree view of the surrounding area.

In our work, we explore a system of stereo cameras capable of providing an almost 360-degree view that effectively captures points and offers comprehensive localization capabilities in dense urban environments. The proposed methodology aims to develop, implement, and assess a novel multi-stereo camera system comprising five synchronized stereo pairs strategically mounted to provide near-360-degree coverage around a vehicular platform. 3D point clouds generated from stereo images are used for localization using a prior 3D map of the environment, and the results are assessed.

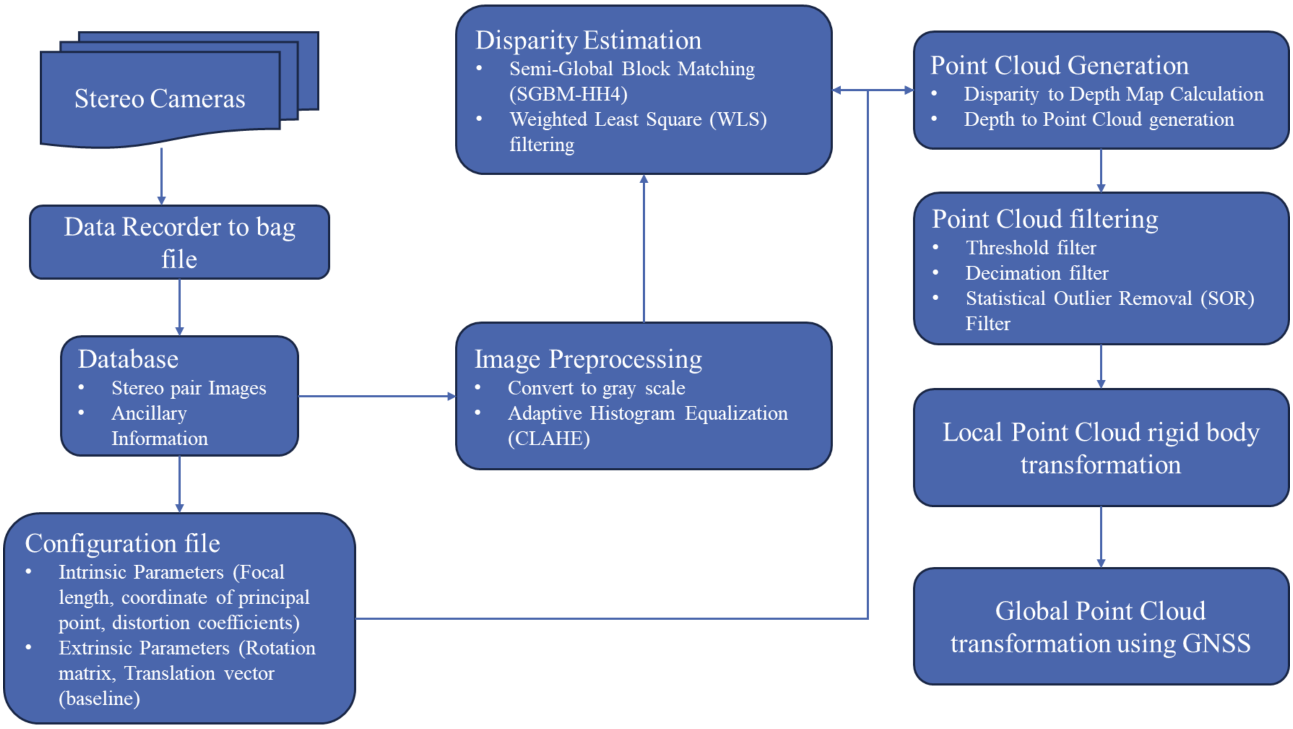

Stereo Cameras
Data Recorder to bag file
Database
Stereo pair Images
Ancillary Information
Configuration file
Intrinsic Parameters (Focal length, coordinate of principal point, distortion coefficients)
Extrinsic Parameters (Rotation matrix, Translation vector (baseline)
Disparity Estimation
Semi-Global Block Matching (SGBM-HH4)
Weighted Least Square (WLS) filtering
Image Preprocessing
Convert to gray scale
Adaptive Histogram Equalization (CLAHE)
Point Cloud Generation
Disparity to Depth Map Calculation
Depth to Point Cloud generation
Point Cloud filtering
Threshold filter
Decimation filter
Statistical Outlier Removal (SOR) Filter
Local Point Cloud rigid body transformation
Global Point Cloud transformation using GNSS

Figure 1. Overview of system workflow.

2. Methodology

2.1 System Overview

The system architecture incorporates several key components and processing stages: Initially, data from five stereo cameras are recorded into a bag file with synchronised time. This data base records stereo pair images along with the position information recorded using GNSS and INS. Once we have the bag files, we prepare a configuration file for the setup which consists of intrinsic and extrinsic camera parameters. We then make use of Robotic Operating System (ROS) to extract the images from the bag file. Disparity estimation is performed on each stereo pair utilizing the Semi Global Block Matching (SGBM) algorithm, followed by Weighted Least Squares (WLS) filtering to enhance the accuracy and density of the resulting point clouds. Implementing WLS filtering significantly reduces noise and outliers in the disparity maps, resulting in more reliable three-dimensional reconstructions. Subsequently, the generated point cloud is filtered and dynamically sub-sampled for fast processing and accurate results. All the point clouds at an epoch are aligned using a robust calibration and alignment procedure that ensures an accurate transformation of individual point clouds into a common reference frame. The aligned point clouds undergo further processing to generate a comprehensive three-dimensional representation of the surrounding environment. For localization purposes, the system utilizes an initial position estimate using the IMU, and the generated point cloud is matched with a pre-existing high-precision reference map using the Iterative Closest Point (ICP) matching algorithm. Results are compared with the trajectory information derived from tightly coupled GNSS/INS integration. Figure 1 shows the overview of the system, and we will discuss the steps in detail in the following sub-sections.

2.2 Depth Estimation

The data processing step involves extracting the data from bag files. We utilize the capabilities of ROS and Python to extract the images from five cameras & mechanical LiDAR and store them. Along with the images, we extract the ancillary information about the sensors, including intrinsic and extrinsic parameters that would later be used for disparity calculations and point cloud generation. The next step in the methodology is to create a configuration file for each camera and enrich it with intrinsic parameters (focal length (fx, fy), coordinate of principal point (cx, cy) and distortion coefficients) and extrinsic parameters (rotation matrix, translation vector and baseline between the stereo camera sensors). This arrangement is advantageous as it consolidates the camera parameters in a single file, streamlining the calibration process and allowing flexibility in adjustments and fine-tuning.

Images from the stereo cameras are pre-processed in two steps. Firstly, the images are converted to grayscale. Converting images to grayscale reduces computational complexity by processing only a single channel instead of three (RGB). We then apply Contrast Limited Adaptive Histogram Equalization (CLAHE) to grayscale images. This step improves the local contrast of the image, rendering features more distinguishable. It also reduces the noise that can occur with standard histogram equalization. CLAHE adapts to different regions of the images, making it robust to different lighting conditions. It is crucial for urban environments with variable lighting conditions, thus outperforming global histogram equalization by preserving local details important for accurate stereo matching.

To derive disparity from the stereo pairs, we utilize the Semi-Global Block Matching (SGBM) algorithm with the HH4 mode. SGBM-HH4 enhances depth accuracy and handles texture-less areas better than default SGBM but requires more tuning. We used SGBM as it is less sensitive to illumination changes and works well in texture-less regions prevalent in urban scenes, which makes it suitable for our case. The produced disparity map is further refined using a Weighted Least Squares (WLS) filter. WLS filtering, unlike other smoothening filters, smooths the disparity map while preserving the edges and filling small holes while improving the 3D depth accuracy. These steps ensure the depth estimates are highly accurate, forming a robust base for the localization algorithm. The combination of SGBM with WLS filtering outperforms simple stereo-matching techniques, granting the accuracy and reliability needed for navigating complex urban surroundings.

2.3 Point Cloud Generation

After we have the disparity map, the next step is to convert the disparity values to actual depth measurements. The depth value is calculated using the relationship between depth and disparity expressed as:

$$\text{depth (Z)} = \frac{Focal\ length \times baseline}{disparity} \tag{1}$$

The formula uses known camera parameters, i.e. focal length and baseline, to translate disparity value into absolute depth measurements, which signifies the three-dimensional representation of the scene. These depth values are combined with the RGB values from the colour image to construct a 3D point cloud. This involves projecting each pixel into the 3D space using the depth values.

$$X = (I_x - c_x) * {}^{Z}/_{f_x} \tag{2}$$

$$Y = (I_y - c_y) * {}^{Z}/_{f_y} \tag{3}$$

$$Z = depth \tag{4}$$

where f_x, f_y = focal length in x and y directions
I_x, I_y = Pixel Coordinates
c_x, c_y = coordinates of principal point
X, Y, Z = Coordinates of 3D point in object coordinate frame

Several point cloud filtering methods are employed to make the point cloud data manageable and filter out the noise. First, a threshold filter is applied to remove the points farther than 20 meters from the sensors. This is done because the depth accuracy decreases with the distance from the stereo camera, and points beyond the threshold become increasingly less reliable as you move away from the sensor. Second, a decimation filter is used to reduce the overall point count so that the computational efficiency can be maintained, and we can achieve near real-time processing speeds as it reduces the volume of data generated while still preserving the geometrical information of the scene. Third, we apply a statistical outlier removal (SOR) filter to detect and discard the noise points that vary from the local point group, improving the quality of the 3D representation. The SOR filter is particularly effective in addressing measurement errors and artifacts that may arise from the stereo-matching process or environmental factors.

Once we have the optimized point cloud, we apply rigid body transformation to align it with the vehicle’s reference frame. This transformation is applied to all the point clouds generated from five stereo cameras so that they align with the vehicle. This step helps ensure that the point clouds are aligned with each other as well as the GNSS/INS system. Finally, to make the point clouds globally consistent and match the prior 3D map, we apply the global transformation to the locally consistent point clouds, using the trajectory information derived from the integrated GNSS/INS system. Finally, we have a set of point clouds from various cameras, all aligning to the global coordinate system.

2.4 Localization

Since we already have globally aligned point clouds from the multi-stereo cameras, we first implemented a testing methodology that simulates real-world scenarios where GNSS signals are compromised. This approach allows the evaluation of the system’s performance under controlled conditions while mimicking the effects of urban canyons on localization accuracy. The testing methodology follows several steps. First, we intentionally introduce errors in rotation and translation to the combined point cloud derived from the five stereo cameras at a single epoch. This simulates the drift that would typically occur when relying solely on inertial measurements in a GNSS-denied setting. This circumstance is especially challenging for traditional navigation systems and acts as an ideal test for a multi-stereo camera approach. To initialize the localization algorithm, we first use the last known GNSS position before the simulated GNSS-denied condition. This replicates the condition where the vehicle has entered an underpass or is between high-rise structures where GNSS signals are unreliable. We use the Iterative Closest Point (ICP) algorithm to align the error-induced combined point cloud with the prior high-precision 3D point cloud map. The corrections applied to the combined point cloud perform several iterations to align the point cloud to the closest match and then provide an RMSE error for the match and the new rotation and translation. Since the combined point cloud has 360-degree coverage, it has comprehensive information about the surrounding environment and can align with the 3D map quite accurately. Using the series of corrected transformations, we generate a trajectory that represents the vehicle path through the simulated scenario.

We compare the generated rotation and translation with the ground truth data to quantify the accuracy of the localization system. This methodology allows us to thoroughly assess the system’s ability to localize in challenging urban environments accurately.

3. Experiments

3.1 Sensors Platform

In this research, we utilize the navigation and instrumentation (NavINST) research laboratory multi-sensor system platform. The setup incorporates advanced hardware components to develop, test, and implement robust localization and navigation algorithms for AVs. The sensor platform consists of five Zed2i stereo cameras, each capable of recording high-resolution images at 1280 x 720 pixels. These stereo cameras come with an in-built IMU sensor. Four of these cameras have a 2.12 mm focal length with an ultrawide field of view (FOV) (110°(H) x 70°(V) x 120°(D)), and these cameras are positioned at the top corners of the vehicle, rotated 45 degrees outward to expand the peripheral vision. Meanwhile, the fifth camera is mounted at the front center on top of the vehicle with a focal length of 4 mm (FOV: 72°(H) x 44°(V) x 81°(D)), facing forward to capture the broad frontal imagery. The illustrated configuration of the multi-stereo camera system is strategically designed to provide comprehensive 360-degree coverage around a vehicular platform, ensuring comprehensive environmental perception.

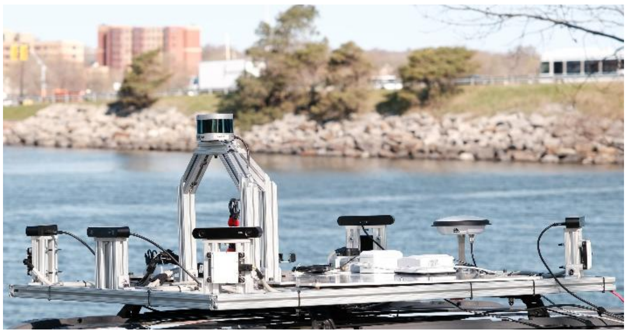


Figure 2. Multi-sensor platform – 4 corner stereo cameras, one forward-looking stereo camera, one mechanical LiDAR, and a GNSS receiver integrated with high-end (reference).

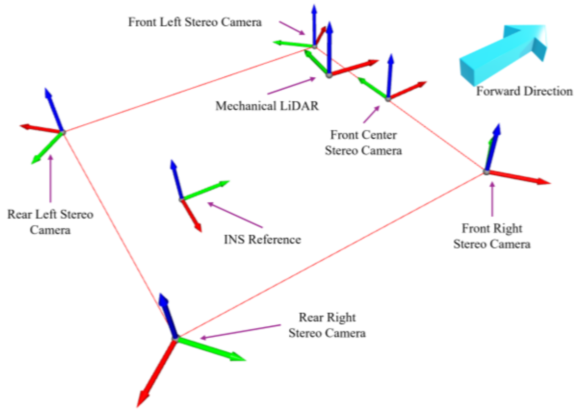


Figure 3. Position and Orientation of Stereo Cameras and Mechanical LiDAR with respect to Reference

Accompanying the cameras, the sensor platform includes a Velodyne Puck LITE mechanical LiDAR with 16 laser channels, with FOV (360° (H) and 30° (V)), and a 100m range. Its point cloud is used to assess and compare the point clouds generated by the multi-stereo camera system. Additionally, the sensor platform also comprises dual-frequency Global Navigation Satellite Systems (GNSS) receiver - Novatel PwrPak7-E1 and a high-grade Inertial navigation System (INS) - tactical grade KVH1750 IMU. We had a base station located within 15km from the trajectory site collecting data as well. This extra data was used inside the Novatel Inertial Explorer software to postprocess the reference data to correct for multiple GNSS-related errors and achieve a Real-time Kinematic (RTK)-like solution with higher positioning accuracy. These sensors assist in providing the initial position estimates to seed the localization algorithm and provide the ground truth for system performance assessment. Data acquisition is handled through multiple computers managed using the Robotic Operating System (ROS). This framework ensures time-synchronized data collection across all sensors stored in bag files that can be used for testing and processing.

3.2 Data

Using the NavINST system, data was collected from real road tests in normal land vehicle driving conditions in downtown Calgary, AB, Canada, capturing measurements from all sensors. The recorded data is stored in ROS bag files and processed on a high-performance system equipped with an Intel Core i7 processor and an Nvidia RTX 360 GPU.Additionally, we utilize a prior 3D map of Calgary, generated using a mobile mapping system, which provides a georeferenced 3D point cloud in the UTM Cartesian coordinate system, covering the downtown area. This prior map serves as a reference for both localization and algorithm assessment.

4. Results and Discussions

3D point clouds generated from the five Zed2i stereo cameras produced a dense and accurate representation of the urban scenarios. Table 1 shows the sample stereo pairs images from the stereo cameras at an epoch. Conversion of RGB images to grey and using the Adaptive Histogram equalization methods are crucial in reducing the input data size as well as ensuring that the features in images are highlighted while making them more robust to different lighting conditions. The implementation of the SGBM algorithm, together with WLS filtering, significantly enhanced the quality of the disparity maps. It resulted in producing a depth accuracy of 0.15 meters at a range of 20 meters under optimal conditions. WLS filtering provided particularly effective results by reducing noise as compared to the traditional stereo-matching results. It added to the overall computation cost but the resultant improvement in disparity map is a trade-off worth the results that are generated. Additional filtering of the point cloud made sure that for further processing at the localization step the point cloud can be processed efficiently while maintaining the accuracy.

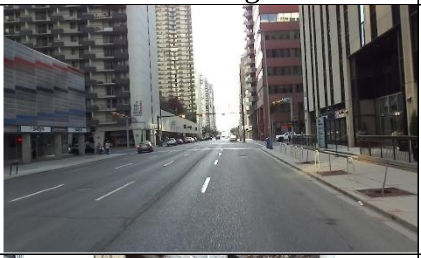
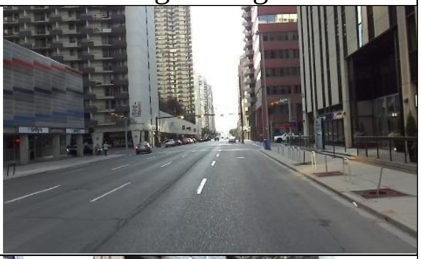
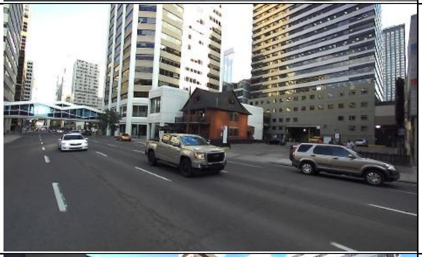
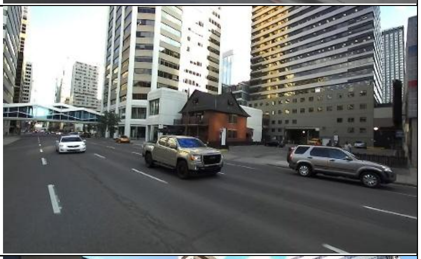
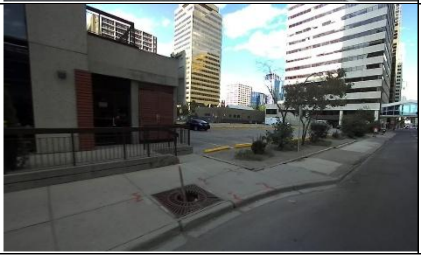
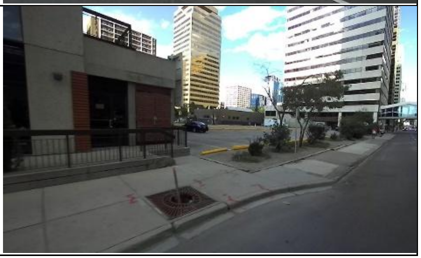
	Left Image	Right Image
Front Center Camera		
Front Left Camera		
Front Right Camera		
Rear Left Camera		
Rear Right Camera		

Table 1. Stereo images from five stereo cameras at an epoch

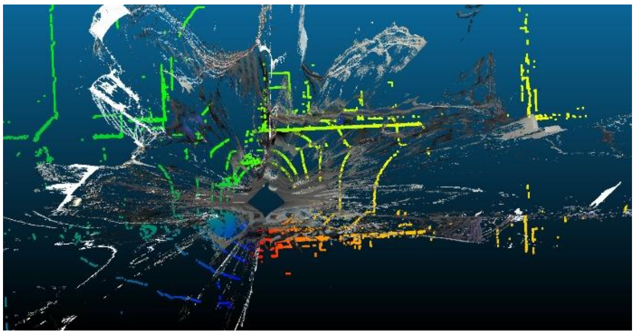


Figure 4. Combined Point Clouds from 5 Stereo Cameras aligned with the mechanical LiDAR point cloud.

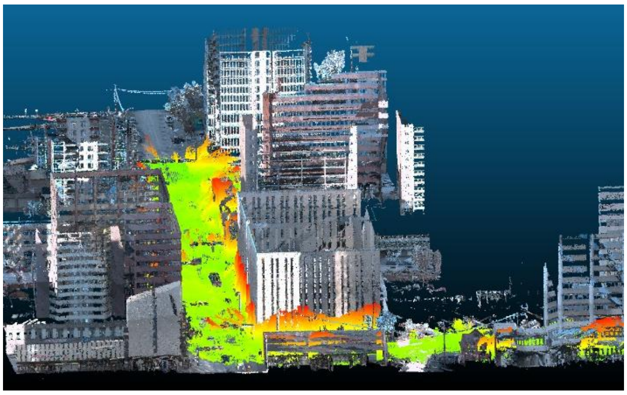


Figure 5. Point Clouds from Stereo Cameras aligned with the prior 3D Point Cloud Map for the trajectory of the Car in Downtown Calgary.

Figure 4 shows the combination of five point clouds from stereo cameras aligned with the mechanical LiDAR point cloud. Figure 5, shows the alignment of all the point clouds with the 3D map of Calgary. Performance of the proposed multi-stereo camera system for localization in urban environment was evaluated and the results thus generated demonstrated that the developed sensor system along with the processing pipeline has capability to provide accurate and reliable localization in challenging urban scenarios. Using ICP matching algorithm to align the combined point cloud from five stereo cameras with high-precession reference map, we achieved a convergence rate in 85% test cases within 2 meters of translation and 8 degrees of rotation.

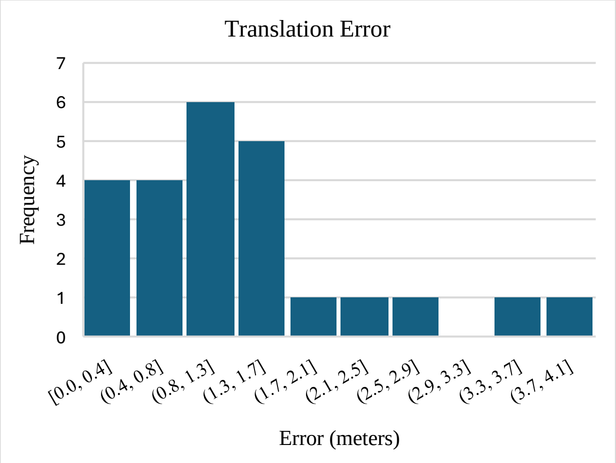


Figure 6. Histograms of translation errors derived from ground truth comparisons.

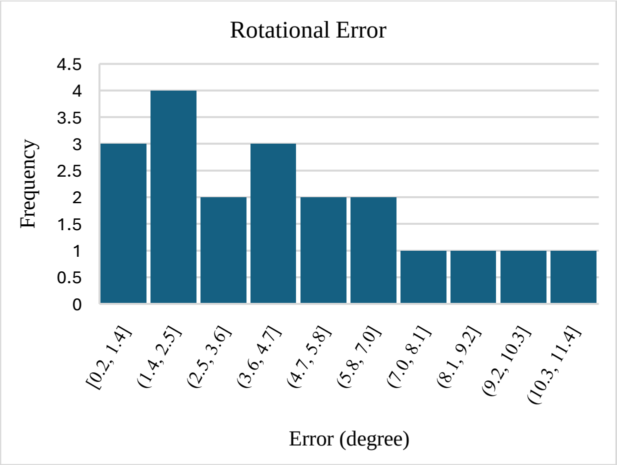


Figure 7. Histograms of rotational errors derived from ground truth comparisons.

The multi-stereo camera configuration showed promising results and addressed the limited field of view constraint inherent in single stereo pair systems. It provided near-360-degree coverage around the vehicular platform, our system achieved comprehensive environmental perception, crucial for navigation in complex urban environments characterized by wide roads and high-rise structures. The combined point cloud generated using the synchronized cameras provides dual benefits. In addition to 360-degree awareness, it also allows additional information for 3D point cloud matching which proves to be helpful in alignment of point cloud with the prior map. The limitation of stereo camera’s shorter baseline limits the reliability of depth, but this also creates an opportunity to use the information provided by cameras which are aligned 45 degrees looking outward. This allows the visibility of sidewalk and buildings that has lot of features which improves the probability of getting those features which can be matched with the point cloud map. By leveraging these additional viewpoints, we can identify the most reliable side for localization. For instance, on wide roads, we can assign greater weight to features on the side closer to the sidewalk, improving alignment and robustness in positioning.

Building and other tall structures casts shadows which creates dynamic lighting conditions that hampers the localization accuracy of the vehicle. It also gets tricky with dynamic objects like cars, pedestrians etc. moving throughout the scene. In addition to these complexities, we also face GNSS signal drop due to urban canyon and multipath effect. All these conditions posed various challenges in the localization of vehicle, but several methods used in our system proved to be useful in countering their effects. Utilizing Adaptive Histogram equalization to increase robustness to the changing lighting conditions, using WLS filtering to reduce the effect of low feature or texture-less surface resulting in less accurate disparity map and several point cloud filtering techniques used to reduce the adverse effects of moving objects resulted in improving the accuracy of the system. By positioning cameras at the corners of the vehicle’s roof, angled 45 degrees looking outward, we enhance localization by leveraging features present on the sides of the road. The traditional methods approach relies on either forward-facing or strictly side-facing cameras that may struggle in environments with limited distinguishing features directly ahead or to the sides, whereas for our sensor arrangement the angled cameras provide features for road as well as the sidewalk or buildings increasing the availability of reliable reference points. On wide roads, where features on one side may be more structured (e.g., near the sidewalk), the system can be designed in future to assign higher confidence to those features for better alignment and localization accuracy. Integrating these advantages, the system can dynamically adjust feature weighting based on environmental context, ensuring more robust and precise localization.

Compare with the LiDAR based solutions, performance of our system suggests that it offers competitive results. The accuracy and comprehensive coverage indicate that the multi-stereo camera approach for camera localization in urban environments has great potential at a lower cost that the high-end sensing solutions. Although processing time for the localization is fast, but it needs to be improved by further optimizing the solution to incorporate real-time processing. For future research we will focus on optimization of the processing pipeline to enable it to work in real time by incorporating parallel processing. We will also focus on fusing perception algorithms to semantically understand the environment and integrate inertial measurements so make the system more efficient and robust.

5. Conclusion

In this research, we presented a multi-stereo camera system designed for accurate and reliable vehicle localization in urban environments. By utilizing five ZED2i stereo cameras with a strategically designed configuration, we achieved near 360-degree environmental perception, addressing the limited field of view inherent in single stereo pairs. Our approach demonstrated the effectiveness of combining stereo depth estimation with advanced filtering techniques such as Adaptive Histogram Equalization, WLS filtering, and point cloud processing to enhance localization accuracy. The disparity maps generated with SGBM and WLS filtering achieved a depth accuracy of 0.15 meters at a 20-meter range, significantly improving the reliability of depth estimation.

The proposed system was evaluated using ICP-based point cloud alignment against a high-precision reference map, achieving an 85% convergence rate within 2 meters of translation and 8 degrees of rotation. The unique positioning of cameras at 45-degree outward angles provided additional visual features from sidewalks and building facades, which enhanced localization performance, particularly in wide-road scenarios where traditional forward-facing or side-facing cameras may struggle. By doing so, we were able to track more reliable features based on environmental context; our system demonstrated robustness in complex urban settings, including those with GNSS signal dropouts, dynamic lighting conditions, and moving objects.

When compared to LiDAR-based solutions, our multi-stereo camera system exhibited competitive localization performance at a significantly lower cost. While the system effectively processes localization data in near real-time, further optimization is necessary to achieve fully real-time operation. Future work will focus on optimizing the processing pipeline through parallel computing, integrating inertial measurements for improved robustness, and incorporating semantic perception algorithms to enhance scene understanding. With these advancements, the proposed system has the potential to become a cost-effective and efficient alternative for urban vehicle localization, contributing to the broader adoption of vision-based navigation solutions in autonomous driving and intelligent transportation systems.

Acknowledgments

This research is supported by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant numbers: RGPIN-2020-03900 and ALLRP-560898-20. The authors acknowledge Micro Engineering Tech Inc. (METI) for providing the 3D maps used in this research.

References

Bosse, M., Zlot, R., & Flick, P., 2012: Zebedee: Design of a spring-mounted 3-d range sensor with application to mobile mapping. *IEEE Transactions on Robotics*, 28(5), 1104-1119. doi.org/10.1109/TRO.2012.2200990.

El-Sheimy, N., Li, Y., 2021: Indoor navigation: state of the art and future trends. *Satellite Navigation* 2, 7. doi.org/10.1186/s43020-021-00041-3.

Guivant, J., Katz, R., 2007: Global urban localization based on road maps. *In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, San Diego, CA, USA. doi.org/10.1109/IROS.2007.4399178.

Gu, Y., Hsu, L. T., and Kamijo, S., 2015: GNSS/onboard inertial sensor integration with the aid of 3-D building map for lane-level vehicle self-localization in urban canyon. *IEEE Transactions on Vehicular Technology*, 65(6), 4274-4287. doi.org/10.1109/TVT.2015.2497001.

Häne, C., Heng, L., Lee, G. H., Fraundorfer, F., Furgale, P., Sattler, T., & Pollefeys, M., 2017: 3D visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68, 14-27. doi.org/10.1016/j.imavis.2017.07.003.

Heng, L., Gao, Z., Yeo, K., Sun, J., Gehrig, S., Lee, G.H., 2019. Project AutoVision: Localization and 3D scene perception for an autonomous vehicle with a multi-camera system. 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, pp. 4695-4702. doi.org/10.1109/ICRA.2019.8793949.

Jo, K., Kim, J., Kim, D., Jang, C. and Sunwoo, M., 2014: Development of Autonomous Car—Part I: Distributed System Architecture and Development Process, *IEEE Transactions on Industrial Electronics*, vol. 61, no. 12, pp. 7131-7140. doi.org/10.1109/TIE.2014.2321342.

Kim, Y., Jeong, J., Kim, A., 2018: Stereo camera localization in 3D LiDAR maps. 2018 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, pp. 1–9. doi.org/10.1109/IROS.2018.8594362.

Kuutti, S., Fallah, S., Katsaros, K., Dianati, M., Mccullough, F., & Mouzakitis, A. , 2018: A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications. *IEEE Internet of Things Journal*, 5(2), 829-846. doi.org/10.1109/JIOT.2018.2812300.

Lee, J. H., Ko, T. H., & Lee, D. W., 2025: Spatiotemporal Calibration for Autonomous Driving Multi-Camera Perception. *IEEE Sensors Journal*, vol. 25, no. 4, pp. 7227-7241. doi.org/10.1109/JSEN.2024.3523569.

Levinson, J., Montemerlo, M., Thrun, S., 2007: Map-based precision vehicle localization in urban environments. *Robotics: science and systems*, Vol. 4, pp. 121-128. doi.org/10.7551/mitpress/7830.001.0001.

Lin, X., Wang, F., Yang, B., & Zhang, W., 2021: Autonomous vehicle localization with prior visual point cloud map constraints in GNSS-challenged environments. *Remote Sensing*, 13(3), 506. doi.org/10.3390/rs13030506.

Mur-Artal, R., Juan D. T., 2017: Orb-slam2: An open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255-1262. doi.org/10.1109/TRO.2017.2705103.

Mounier, E., Elhabiby, M., Korenberg, M., & Noureldin, A., 2024: LiDAR-Based Multi-Sensor Fusion with 3D Digital Maps for High-Precision Positioning. *IEEE Internet of Things Journal*. doi.org/10.1109/JIOT.2024.3492913.

Najjar, M.E.E., 2005: A Road-Matching Method for Precise Vehicle Localization Using Belief Theory and Kalman Filtering. *Auton Robots*, 19, 173–191. doi.org/10.1007/s10514-005-0609-1.

Neubert, P., Schubert, S., & Protzel, P., 2017: Sampling-based methods for visual navigation in 3D maps by synthesizing depth images. *In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2492-2498. doi.org/10.1109/IROS.2017.8206067.

Nüchter, A., Lingemann, K., Hertzberg, J., & Surmann, H., 2007: 6D SLAM—3D mapping outdoor environments. *Journal of Field Robotics*, 24(8-9), 699-722. doi.org/10.1002/rob.20209.

Pendleton, S.D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y.H., Rus, D., Ang, M.H., 2017: Perception, Planning, Control, and Coordination for Autonomous Vehicles. *Machines*. 5(1):6. doi.org/10.3390/machines5010006.

Ruchti, P., Steder, B., Ruhnke, M., & Burgard, W., 2015: Localization on openstreetmap data using a 3d laser scanner. *In 2015 IEEE international conference on robotics and automation (ICRA),* Seattle, WA, USA, pp. 5260-5265. doi.org/10.1109/ICRA.2015.7139932.

Stewart, A. D., & Newman, P., 2012: Laps-localisation using appearance of prior structure: 6-dof monocular camera localisation using prior pointclouds. *In 2012 IEEE International Conference on Robotics and Automation,* Saint Paul, MN, USA, pp. 2625-2632. doi.org/10.1109/ICRA.2012.6224750.

Wan, G., Yang, X., Cai, R., Li, H., Zhou, Y., Wang, H., & Song, S., 2018: Robust and precise vehicle localization based on multi-sensor fusion in diverse city scenes. *In 2018 IEEE international conference on robotics and automation (ICRA),* Brisbane, QLD, Australia, 4670-4677. Doi.org/10.1109/ICRA.2018.8461224.

Wolcott, R. W., & Eustice, R. M., 2014: Visual localization within lidar maps for automated urban driving. *In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Chicago, IL, USA, pp. 176-183. doi.org/10.1109/IROS.2014.6942558.

Xu, Y., John, V., Mita, S., Tehrani, H., Ishimaru, K., Nishino, S., 2017: 3D point cloud map based vehicle localization using stereo camera. 2017 *IEEE Intelligent Vehicles Symposium (IV)*, Los Angeles, CA, USA, pp. 487–492. doi.org/10.1109/IVS.2017.7995765.

Yabuuchi, K., Wong, D.R., Ishita, T., Kitsukawa, Y., Kato, S., 2021: Visual localization for autonomous driving using pre-built point cloud maps. 2021 *IEEE Intelligent Vehicles Symposium (IV)*, Nagoya, Japan, pp. 913-919. doi.org/10.1109/IV48863.2021.9575569.

Yoneda, K., Tehrani, H., Ogawa, T., Hukuyama, N., & Mita, S., 2014: Lidar scan feature for localization with highly precise 3-D map. *In 2014 IEEE Intelligent Vehicles Symposium Proceedings*, Dearborn, MI, USA, pp. 1345-1350. doi.org/10.1109/IVS.2014.6856596.