Enhancing Aerial Camera-LiDAR Registration through Combined LiDAR Feature Layers and Graph Neural Networks

Jennifer Leahy [1], Shabnam Jabari [1]

[1] University of New Brunswick, 3 Bailey Dr, Fredericton NB - (jennifer.leahy, sh.jabari)@unb.ca

Keywords: Camera-LiDAR Registration, LiDAR Feature Layers, Feature Matching, Graph Neural Networks, SuperGlue, Aerial Data.

Abstract

Integrating optical images with Light Detection and Ranging (LiDAR) data is an important advance in Photogrammetry, Geomatics and Computer Vision, registering the strengths of both modalities (height and spectral information). Most orthoimages and aerial LiDAR data are georeferenced to a common ground coordinate system; however, a registration gap remains, and achieving high-accuracy registration between these datasets is challenging due to their differing data formats and frames of reference. In this paper, we propose an approach to enhance camera-LiDAR registration through combined LiDAR feature layer generation and Deep Learning. Our method involves creating weighted combinations of feature layers from LiDAR data, leveraging intensity, elevation, and bearing angle attributes. Subsequently, a 2D-2D Graph Neural Network (GNN) pipeline serves as an intermediate step for feature detection and matching, followed by a 2D-3D affine transformation model to register optical images to point clouds. Experimental validation across aerial scenes demonstrates significant improvements in registration accuracy. Notably, in urban building areas, we achieved an RMSE of around 1.1 pixel, marking a reduction of 5 pixels compared to georeferenced baseline values. In rural road scenes, our method yielded a pixel RMSE of 1.3, with a 4-pixel reduction compared to baseline results. Additionally, in water scenes, which tend to be noisy in LiDAR data, we achieved a pixel RMSE of 1.8, representing a slight half-pixel reduction compared to the baseline. Therefore, by using weighted and combined LiDAR feature layer and GNN feature matching, this approach augments the number of key points and matches, directly correlating with the observed registration reduction in pixel RMSE across diverse aerial scene types.

1. Introduction

Combining optical images with Light Detection and Ranging (LiDAR) data is an important advancement in Photogrammetry, Geomatics and Computer Vision, registering color and texture from optical cameras with depth and geometric details from LiDAR. Establishing key point correspondences between data from the two sensors is central to this fusion. However, challenges such as diverse data formats, inconsistent frames of reference, and noise, contribute to the difficulty in achieving high accuracy, emphasizing the ongoing research in this field.

In the domain of 2D-3D LiDAR-image registration, Li et al. (2022) categorized developed techniques into four classic methods. Information Theory methods integrate statistical relationships between sensor data (Parmehr et al., 2012; Weese et al., 1997). Feature-based techniques leverage distinctive features in LiDAR and image data (Shu et al., 2022; Yan et al., 2023; Yao et al., 2010). Ego-motion approaches, seen in Odometry (Bai et al., 2022; Dellenbach et al., 2021; Taylor & Nieto, 2015) and Kalman Filtering (Das et al., 2022; Gao & Harris, 2002; Kunjumon & G S, 2021; Mu et al., 2020) aim to model the movement or motion of the sensor platform, providing valuable context for registration. Finally, Learning-Based methods use neural networks to learn mappings between LiDAR and image data (Chen et al., 2022; Jeon & Seo, 2022; J. Li & Lee, 2021). Despite notable progress in the field, current methods lack the simplicity and computational efficiency of 2D-2D image matching techniques (Karami et al., 2015), attributed to the complexities of handling 3D data in three-dimensional space. Recognizing this highlights the central challenge of this work: integrating a 2D-2D multi-modal feature matching into camera-LiDAR registration. This entails transforming LiDAR data into a 2D feature layer, aligning it with the intricacies of an optical camera image. The 2D-2D correspondences between the datasets are then employed in the 2D-3D registration model.

In this paper, Section 3.2, outlines our method for creating combined feature layers from LiDAR data. This involves leveraging various channel combinations, including intensity, elevation, and bearing angle. In Section 3.3, we present the implementation of a 2D-2D Deep Leaning pipeline for keypoint detection and image matching using a Graph Neural Network (GNN). In section 3.4, we employ Random Sample Consensus (RANSAC) to eliminate outliers and establish a 6-parameter affine transformation to register optical images to point clouds. We evaluate performance using metrics like Euclidean distance and Root Mean Square Error (RMSE), which measure the disparity between the GNN's camera image keypoint pixel coordinate location and the predicted value by the model. In section 4, our methodology is validated with 3 scenes of aerial LiDAR data and orthophotos across diverse conditions, including dense urban areas, rural road networks, and water bodies.

2. Related Work

Two-dimensional image matching has advanced beyond traditional methods like Scale Invariant Feature Transform (SIFT) and Speed Up Robust Feature (SURF). While these methods are effective at finding correspondences between 2D images regardless of orientation and scale (Karami et al., 2015), there has been a recent paradigm shift driven by the integration of machine learning. Transformers have played a pivotal role in this evolution, as demonstrated by Wang et al.'s (2023) Cross-modality Multi-scale Progressive Dense Registration (C-MPDR) scheme, optimizing coarse-to-fine registration in a single stage. Similarly, Sun et al. (2021) introduced Local Feature Matching with Transformers (LOFTR), which produces dense matches even in low-texture areas. Another notable advancement involves Graph Neural Networks (GNN), such as SuperGlue (Sarlin et al., 2020), which not only finds correspondence but also rejects non-matchable points, achieving state-of-the-art results in pose estimation for challenging real-world

environments. Its front-end keypoint detector, SuperPoint (DeTone et al., 2017), seamlessly pairs with its middle-end pipeline, which incorporates attention-based context aggregation.

The shift towards utilizing LiDAR in a 2D image format has emerged as a key focus in recent research, capitalizing on the superior performance of image-matching algorithms. Liu et al. (2017) investigated various 2D LiDAR image models, among which the Fast Optimal Bearing-Angle (FOBA) notably reduced scene segmentation time costs with minimal accuracy loss. Although FOBA's primary purpose was for scene segmentation, Lin et al. (2017) built upon this foundation and used it for registration, namely between 3D point clouds. They utilized Bearing Angle images with SURF for matching and achieved a significant reduction in computational costs. Zhuang et al. (2013), use a similar methodology but deal with the random disturbances caused by unexpected movements of people and other objects. Addressing increased outliers in these image pipelines, Li & Li (2021) incorporated Random Sample Consensus (RANSAC) into their method, leveraging 2D submap projection images to find dependable matched pairs for the pose estimation.

Subsequent studies have further expanded the application of creating 2D images from LiDAR data, to include the registration of camera images with LiDAR data. For example, Scaramuzza et al. (2007) focused on camera-LiDAR extrinsic calibration using range images, a concept extended by Koide et al. (2023) with Super Glue, a Graph Neural Network, on intensity images. While these methods make strides in advancing singular aspects of LiDAR data, such as range, intensity, or bearing angle, they may overlook the potential of exploring various combinations and weights of LiDAR layers to better represent the three bands of an optical camera image.

3. Proposed Method

3.1 Overview

This research employs a systematic approach leveraging high-resolution orthoimages and dense aerial LiDAR data to establish accurate correspondences for registration within a two-dimensional framework. We presume that essential information such as intensity, elevation, and scan angles can be extracted from the LiDAR dataset. This pipeline can be employed across diverse outdoor scenes. Figure 1 presents a general flowchart detailing the process.

3.2 LiDAR Feature Layer Generation

Each LiDAR data point contains information apart from its geometric coordinates. The returned LiDAR pulses may also record the intensity of the reflected signal off the surface, providing visual information similar to that of an optical camera. Elevation data, is also obtained from the laser beam's vertical position, indicating terrain or object height. Additionally, scan angles between pulses reveal the laser beam's direction relative to the LiDAR sensor and surface, aiding in capturing geometric detail. By utilizing the first returns of intensity, elevation, and scan angle information, LiDAR feature layers can be created by converting each 3D point into a 2D grid cell. This cell represents a specific area on the ground and contains the respective information derived from the LiDAR data.

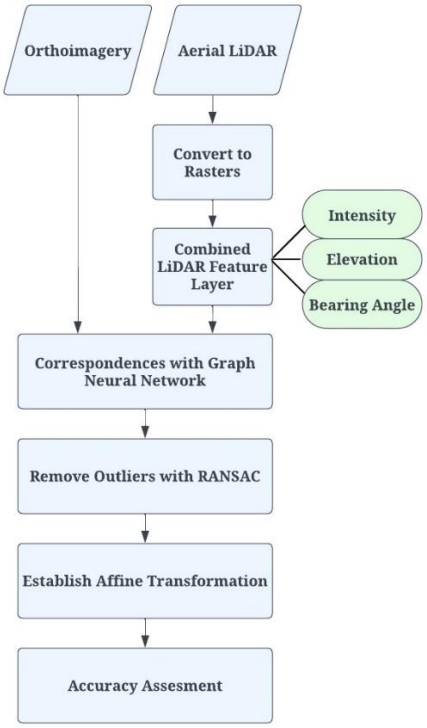

Figure 1. Workflow for Registration of Orthoimages and Aerial LiDAR Data

Additional steps are involved for creating a Bearing Angle image as it requires a relationship between the range and scan angle information. The range information is derived from the sensors' constant flight height (H), the surface elevation (h), and the cosine of the scan angle (A). Treating each input data as a raster, each pixel can be transformed into the range data according to Equation 1 below.

$$p = (H - h)/cosA \quad (1)$$

Building upon previous research (Liu et al., 2017), a function was created to compute the bearing angles for each pixel. The bearing angle is the angle between the laser beam and the line segment connecting two consecutive laser scanning points. The function takes the range and scan angle rasters as inputs and proceeds to iterate through each pixel applying Equation 2 below.

$$BA = (a - b * \cos \Phi) / (\sqrt{a^2 + b^2 - 2ab * \cos \Phi}), \quad (2)$$

where a = current height value in the height image
b = preceding range value in the height image
Φ = corresponding angle of increment

The rasterized LiDAR data undergo a two-step enhancement process. Firstly, their intensity values are stretched using Histogram Equalization, enhancing contrast. Subsequently, the images are cubically interpolated for smoother transitions and finer details. Finally, they are exported as 8-bit image files, resized to reduce their dimensions, and made compatible with subsequent processing steps.

Our investigation aims to combine these feature layers into a multilayer format, exploring various combinations and weight distributions to assess their impact on matching results and accuracy. Given that the GNN operates most effectively with

grayscale images, we merge the three layers into a single band to maintain control over band weighting. This process begins with normalization, scaling each layer's pixel values to fall within the range of 0 to 1. Subsequently, we merge and weigh the layers in different combinations, so that the combined contributions of all layers sum up to 1. Weight adjustments are fine-tuned with smaller increments allocated to areas yielding more confident matches, while larger increments are assigned to areas showing underperformance. By testing different combinations of weight increments (as shown in Table 1), including individual layer weighted at 100%, double layer combinations, and triple layer combinations, we aim to find the most effective input configuration for the GNN.

3.3 Correspondences with Graph Neural Networks

The Graph Neural Network architecture for this methodology consists of a front-end keypoint detector, SuperPoint (DeTone et al., 2017), and a middle-end keypoint matcher, SuperGlue (Sarlin et al., 2020). Illustrated in Figure 2, this GNN adaptation leverages combined LiDAR feature layers.

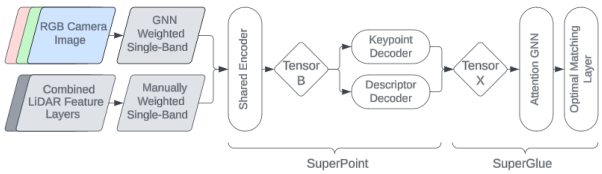

Figure 2. GNN Pipeline Adaptation using Combined LiDAR Feature Layers

In the front end, both the raw optical image and its corresponding LiDAR feature layer serve as inputs. They undergo processing through a shared encoder and two parallel interest and descriptor decoders. The shared encoder preprocesses the images, generating tensor B, reduced to 1/8th of its original size. Meanwhile, the interest and descriptor decoders evaluate each pixel of Tensor B, predicting its likelihood of being a key point and computing associated descriptors, resulting in the creation of Tensor X, with image features.

Following feature detection, the middle-end utilizes key points from tensor X as interconnected nodes throughout and between image pairs. Messages about descriptors are transmitted along these edges, with attention mechanisms prioritizing relevant information. Finally, the Optimal Matching Layer determines the best match between key points based on their descriptors, by assigning a score to each potential match. The objective is to maximize the total score of all matches to find accurate and reliable features matching across both camera and LiDAR images.

To validate the correspondence results and proceed only with the high-performing LiDAR feature layers, the inputs undergo SuperGlue evaluation at a significantly higher set matching threshold of 0.95. The matches are assessed by calculating Euclidean distances and Root Mean Square Error (RMSE) between all predicted values and manually selected ground truth values.

3.4 Image Transformation

Establishing (x, y) pixel coordinate correspondences between the camera and LiDAR image data enables the use of a 6 parameters affine transformation to register the datasets, facilitating the creation of a colorized 3D point cloud model. This transformation model is suitable for this research because the LiDAR and camera images are rasters in the same coordinate system and relief displacement has already been rectified. It incorporates linear translation, rotation, and uniform scaling, to align the 2D image with the 3D LiDAR data.

The transformation equation, expressed in Equation 3 below, was implemented in Python using the RANSAC algorithm. RANSAC randomly selects a subset of data points from the input correspondence and iteratively fits the 6-parameter affine model to each subset, evaluating the residuals selecting the coefficients that minimize the residuals, and removing outliers from the data. The fitted coefficients, a to e, are calculated using least squares. Essentially, these coefficients determine how the pixel coordinates in the optical image (x, y) are transformed to match the corresponding coordinates in the LiDAR data (X, Y).

$$X = a * x + b * y + c \qquad (3)$$
$$Y = d * x + e * y + f$$

To evaluate the accuracy of the transformation, the matching correspondences from SuperGlue were randomly divided into 70% for training and 30% for testing. The accuracy was assessed by computing the RMSE between the original LiDAR point location and the transformed layer point to that location.

4. Experiment Validation

To validate our research experimentally, we utilized aerial LiDAR data from the RIEGL VQ-780i system, with the density of 6 points per square meter, alongside high-resolution ortho imagery acquired by the UltraCam Eagle at a resolution of 0.07 meters per pixel (GeoNB, 2015, 2020). Our investigation spans three distinct outdoor scenes—buildings, roads, and water—within the city of Fredericton, New Brunswick, as depicted in Figure 3. With the airborne sensors Operating at an approximate altitude of 1000 meters, our study areas, strategically sized to encompass a square kilometer, facilitate focused analysis. Throughout, we maintain a uniform coordinate system of NAD1983 CSRS datum, NB Stereographic map projection, and CGVD2013 vertical datum.

Figure 3. Study Area Encompassing 3 Scenes: Buildings, Roads, and Water

In each scene, specific Regions of Interest (ROIs) are deliberately shifted between the image and LiDAR data to exaggerate the proof of concept. The camera images, inherently in RGB format, remain unaltered throughout the process. The LiDAR data are subjected to preprocessing in ArcGIS Pro, utilizing first return lidar values to generate three raster outputs: elevation, intensity, and bearing angle, for each of the three scenes outlined in Section 3.2. Consequently, this yields nine distinct image outputs, as illustrated in Figure 4., with columns representing Buildings, Roads, and Water, and rows corresponding to intensity, elevation, and bearing angle.

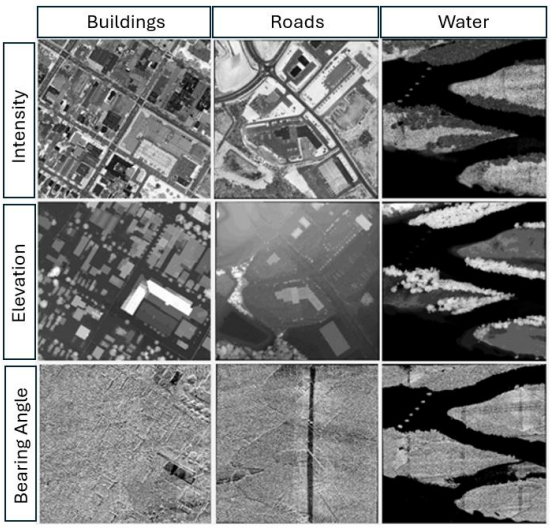

Figure 4. Single LiDAR Feature Layer Outputs

When creating the rasters, a sampling value of 0.5 was chosen, relative to the point spacing of our data 0.3, adhering to the guideline recommending twice the average point spacing. To fill the gaps in the created raster, bilinear interpolation was used to average values from neighboring data cells, eliminating small voids within the data. Subsequently, after raster creation, the intensity, elevation, and bearing angle values underwent cubic interpolation and stretching using Histogram Equalization, to improve visualization of the data. Finally, all outputs were exported as 8-bit Unsigned files.

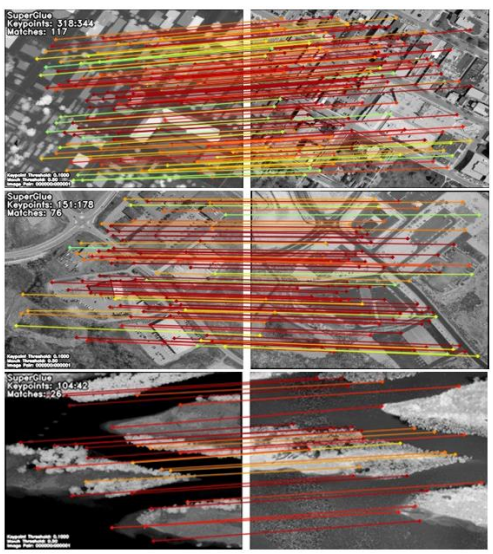

Figure 5. Visualization of SuperGlue Matches for Buildings (Top), Roads (Middle), and Water (Bottom).

The individual intensity, elevation and bearing angle rasters are combined into a single layer within each respective scene, using the normalization and weighting process described in Section 3.2. Various combinations and weights are tested, as detailed in the first three columns of Table 1, to find the most effective input configuration for the GNN. These multi-layer images are input to Super Glue's Graph Neural Network (GNN) for feature detection and matching. Initial optimal results are attained by utilizing the 'outdoor' Super Glue configuration, with a keypoint threshold of 0.1 and a match threshold of 0.5, alongside a Non-Maximum Suppression (NMS) radius of 5 and maintaining a consistent image size of 320x240. This configuration provides low run-time and high confidence, yielding numerous matches during the initial examination. Visual representations in Figure 5 showcase the match strengths, color-coded from weakest (blue) to strongest (red).

The matching results are compiled in Table 1 to categorize layers based on Bearing Angle, Intensity, and Elevation, showcasing their respective weighted influences. Weight values of 'NA' and 100 indicate exclusion and sole representation, respectively. Bolded red values denote superior combinations compared to the more traditional single-layer approach and red arrows point in the direction of increasing matches, showing the trends in each scene.

	Feature Layers			Test Areas		
	Bearing Angle	Intensity	Elevation	Buildings	Roads	Water
	Weights (%)			Matching Threshold @ 0.5		
Single layer	100	NA	NA	2	0	18
	NA	100	NA	89	**69**	19
	NA	NA	100	**105**	49	**25**
Double layer	NA	5	95	**110**	41	**26**
	NA	10	90	**117**	33	**25**
	NA	20	80	77	45	**26**
	NA	25	75	73	25	**25**
	NA	34	66	38	47	**26**
	NA	50	50	82	47	22
	NA	66	34	97	**76**	16
	NA	75	25	92	**71**	19
	NA	80	20	96	**73**	20
	NA	90	10	100	**72**	22
	NA	95	5	102	**73**	22
Triple layer	60	30	10	3	0	18
	60	10	30	1	0	19
	50	25	25	2	0	21
	40	30	30	3	0	20
	33.33	33.33	33.33	6	20	15
	30	60	10	57	61	18
	30	10	60	58	0	23
	20	30	50	2	41	19
	20	50	30	67	61	19
	20	40	40	16	58	17
	10	30	60	16	52	19
	10	60	30	89	**78**	22
	10	70	20	98	**76**	21
	10	80	10	102	**75**	18
	10	85	5	92	**78**	17
	5	85	10	104	**75**	19
	5	90	5	104	**75**	21
	5	10	85	**108**	23	22
	5	5	90	104	40	22

Table 1. Number of Correspondences based on Layer Combination Weights

The Building scenes show the most matches with 10% Intensity + 90% Elevation, while the Road scene performs best with 10% BA + 60% Intensity + 30% Elevation. The Water scene exhibits relatively consistent performance across layers, with enhanced results observed at 5% Intensity + 95% Elevation. Further trends reveal a preference for heavier weighting on the layer that best represents the scene's key features: Buildings and Water benefit from heavier Elevation weighting, capturing taller features, while Roads benefit from heavier Intensity weighting,

highlighting flatter areas and road features. It's also noteworthy that the Bearing Angle layer underperformed, lacking clear feature definition, and offering minimal benefit to the multi-layer format. Weighting this layer over 30% led to suboptimal outcomes.

To validate results, a select few LiDAR feature layers from each scene undergo SuperGlue evaluation at a higher matching threshold of 0.85 at a resolution of 770 × 680. Subsequently, with these strong matches, the Python scikit-learn library's RANSACRegressor is employed to fit the 6-parameter affine transformation. This process utilizes 70% of matches as training data and 30% as testing data, effectively eliminating outliers. Through this registration of camera and LiDAR datasets, the generation of a 3D RGB point cloud model is facilitated for each scene, as depicted in Figure 5.

Figure 6. Snapshots of registered RGB Point Clouds of Buildings (Top), Roads (Middle), and Water (Bottom).

Model performance is assessed post-prediction with the testing data using Euclidean distances and RMSE between predicted and ground truth values, with results detailed in Table 2. The baseline was established from the original georeferenced rasters, with a manual selection of matching pixel locations across the image. It was divided into 16 sections, with 16 points chosen, including ground points, vegetation, and building corners, to create a diverse baseline evaluation.

Our analysis yielded significant improvements in registration accuracy across various environmental settings. In the Building scene, the baseline RMSE of 6.3 pixels was substantially reduced to 1.1pixels using a combination of 66% Intensity and 34% Elevation, marking a 5-pixel improvement in accuracy. Similarly, in the Road scene, the initial baseline RMSE of 5.3 pixels saw a reduction by 4 pixels to 1.300 when employing the intensity image. Furthermore, in the Water scene, the baseline RMSE of 2.3 pixels was improved by a half pixel to 1.8 pixels RMSE with the use of the Elevation image. Notably, all chosen combinations of layers yielded better accuracy than the baseline, demonstrating that utilizing a weighted and combined LiDAR feature layer and 2D image GNN approach to register these modalities provides meaningful results.

		GNN Match Confidence Threshold	Total Number of Matches	Baseline RSME	Our Registration Method RMSE
B	100% E	0.85	74	6.255	1.388
	66%I + 34%E		105		1.109
	10%BA + 80%I + 10%E		100		1.412
R	100% I	0.85	74	5.327	1.300
	90%I + 10%E		58		1.482
	10%BA + 60%I + 30%E		72		1.452
W	100% E	0.85	37	2.333	1.827
	5% I 95% E		32		1.964
	10%BA + 20%I + 70%E		28		2.253

Table 2. Registration Parameters, Accuracy, and Baseline RMSE Values in Various Scenes

This research provides insights for those aiming to register aerial optical images with LiDAR data. The findings suggest that using a combination of intensity, elevation, and bearing angle rasters, weighted appropriately, yields accurate results when matched with a GNN. For flat areas, a heavier emphasis on intensity weights is recommended, while varied intensity and elevation combinations are best for tall structures, like buildings and trees. But these combinations should be fine-tuned based on specific scene characteristics.

5. Conclusion

In this study, we investigated an approach to enhance optical camera-LiDAR registration through combined LiDAR feature layers and Graph Neural Networks. By generating combined feature layers from LiDAR data, employing a 2D-2D GNN pipeline for feature detection and matching, and establishing a 6-parameters affine transformation, we achieved significant improvements in registration accuracy across diverse outdoor scenes. For instance, our results indicated a 5-pixel improvement in registration accuracy compared to the baseline for Building scenes, a 4-pixel improvement in Road scenes, and a half pixel improvement in Water scenes.

Our experiments confirm our hypothesis that using a Graph Neural Network for feature matching between aerial orthoimages, and combined LiDAR feature layer combinations produces strong, viable matches, leading to accurate registration results and improving upon traditional baselines. The GNN's ability to focus on geometric features, rather than relying solely on spectral and intensity gradients like traditional methods, allows it to handle multi-modal matching effectively.

Despite these successes, our study identified limitations, particularly regarding the performance of bearing angle images, potentially due to the nature of the height and scan angle data obtained at high altitudes. Future work could focus on optimizing BA image processing techniques to enhance their suitability for

registration tasks. Additionally, our future research will involve implementing this methodology with a wider range of data such as terrestrial LiDAR point clouds, hand-held camera imagery, and low-cost multi-sensor Mobile Mapping System (MMS) datasets. Accurate registration between optical cameras and LiDAR data holds significant potential to advance autonomous systems, urban planning, environmental management, and non-invasive mapping practices. Thus, our study contributes valuable insights and sets the stage for further research and innovation in this field.

References

Bai, C., Fu, R., & Gao, X. (2023). Colmap-PCD: An Open-source tool for fine image-to-point cloud registration. *ArXiv*, *abs/2310.05504*. https://api.semanticscholar.org/CorpusID:263830267

Bai, C., Xiao, T., Chen, Y., Wang, H., Zhang, F., & Gao, X. (2022). Faster-LIO: Lightweight Tightly Coupled Lidar-Inertial Odometry Using Parallel Sparse Incremental Voxels. *IEEE Robotics and Automation Letters*, *7*(2), 4861–4868. https://doi.org/10.1109/LRA.2022.3152830

Chen, H., Wei, Z., Xu, Y., Wei, M., & Wang, J. (2022). *ImLoveNet: Misaligned Image-supported Registration Network for Low-overlap Point Cloud Pairs*. 1–9. https://doi.org/10.1145/3528233.3530744

Das, D., Adhikary, N., & Chaudhury, S. (2022). Sensor fusion in autonomous vehicle using LiDAR and camera Sensor. *2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC)*, 336–341. https://api.semanticscholar.org/CorpusID:253271132

Dellenbach, P., Deschaud, J.-E., Jacquet, B., & Goulette, F. (2021). CT-ICP: Real-time Elastic LiDAR Odometry with Loop Closure. *2022 International Conference on Robotics and Automation (ICRA)*, 5580–5586. https://api.semanticscholar.org/CorpusID:237941150

DeTone, D., Malisiewicz, T., & Rabinovich, A. (2017). *SuperPoint: Self-Supervised Interest Point Detection and Description*.
Esri. (2022). ArcGIS Pro (Version 3.0.0) [Software]

Gao, J., & Harris, C. J. (2002). Some remarks on Kalmam filters for the Multi-sensor fusion. *Information Fusion*, *3*. https://doi.org/10.1016/S1566-2535(02)00070-2

GeoNB. (2015). *Orthoimagery*. Https://Geonb.Snb.ca/Nbimagery/Index.Html.

GeoNB. (2020). *LiDAR Data*. Https://Geonb.Snb.ca/Li/.

Jeon, Y., & Seo, S.-W. (2022). EFGHNet: A Versatile Image-to-Point Cloud Registration Network for Extreme Outdoor Environment. *IEEE Robotics and Automation Letters*, *7*(3), 7511–7517. https://doi.org/10.1109/LRA.2022.3183899

Karami, E., Prasad, S., & Shehata, M. (2015, January). *Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images*.

Kitware Inc. (2023). *LiDARView* (4.3.0).

Koide, K., Oishi, S., Yokozuka, M., & Banno, A. (2023). General, Single-shot, Target-less, and Automatic LiDAR-Camera Extrinsic Calibration Toolbox. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 11301–11307. https://doi.org/10.1109/ICRA48891.2023.10160691

Kunjumon, R., & G S, S. (2021). *Sensor Fusion of Camera and Lidar Using Kalman Filter* (pp. 327–343). https://doi.org/10.1007/978-981-16-2248-9_32

Li, J., & Lee, G. (2021). *DeepI2P: Image-to-Point Cloud Registration via Deep Classification*. 15955–15964. https://doi.org/10.1109/CVPR46437.2021.01570

Li, X., Xiao, Y., Wang, B., Ren, H., Zhang, Y., & Ji, J. (2022). Automatic targetless LiDAR–camera calibration: a survey. *Artificial Intelligence Review*, *56*, 1–39. https://doi.org/10.1007/s10462-022-10317-ylin

Li, Y., & Li, H. (2021). LiDAR-Based Initial Global Localization Using Two-Dimensional (2D) Submap Projection Image (SPI). *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 5063–5068. https://doi.org/10.1109/ICRA48506.2021.9560740

Lin, C.-C., Tai, Y.-C., Lee, J.-J., & Chen, Y.-S. (2017). A novel point cloud registration using 2D image features. *EURASIP Journal on Advances in Signal Processing*, *2017*. https://doi.org/10.1186/s13634-016-0435-y

Liu, Y., Wang, F., Abdullah, D., He, G., & Zhuang, Y. (2017). Comparison of 2D Image Models in Segmentation Performance for 3D Laser Point Clouds. *Neurocomputing*, *251*. https://doi.org/10.1016/j.neucom.2017.04.030

Liu, Z., van Oosterom, P., Balado, J., Swart, A., & Beers, B. (2023). Data frame aware optimized Octomap-based dynamic object detection and removal in Mobile Laser Scanning data. *Alexandria Engineering Journal*, *74*, 327–344. https://doi.org/https://doi.org/10.1016/j.aej.2023.05.014

Mu, L., Yao, P., Zheng, Y., Chen, K., Wang, F., & Qi, N. (2020). Research on SLAM Algorithm of Mobile Robot based on the fusion of 2D LiDAR and Depth Camera. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2020.3019659

Parmehr, E. G., Fraser, C. S., Zhang, C., & Leach, J. (2012). Automatic Registration of Aerial Images with 3D LiDAR Data Using a Hybrid Intensity-Based Method. *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*, 1–7. https://doi.org/10.1109/DICTA.2012.6411697

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). SuperGlue: Learning Feature Matching With Graph Neural Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4937–4946. https://doi.org/10.1109/CVPR42600.2020.00499

Scaramuzza, D., Harati, A., & Siegwart, R. (2007). Extrinsic self calibration of a camera and a 3D laser range finder from natural scenes. *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4164–4169. https://doi.org/10.1109/IROS.2007.4399276

Shu, F., Wang, J., Pagani, A., & Stricker, D. (2022). Structure PLP-SLAM: Efficient Sparse Mapping and Localization using Point, Line and Plane for Monocular, RGB-D and Stereo Cameras. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2105–2112. https://api.semanticscholar.org/CorpusID:250492735

Sun, J., Shen, Z., Wang, Y., Bao, H., & Zhou, X. (2021). LoFTR: Detector-Free Local Feature Matching with Transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8918–8927. https://doi.org/10.1109/CVPR46437.2021.00881

Taylor, Z., & Nieto, J. (2015). Motion-Based Calibration of Multimodal Sensor Arrays. *Proceedings - IEEE International Conference on Robotics and Automation*, *2015*. https://doi.org/10.1109/ICRA.2015.7139872

Wang, D., Liu, J., Ma, L., Liu, R., & Fan, X. (2023). *Improving Misaligned Multi-modality Image Fusion with One-stage Progressive Dense Registration*.

Weese, J., Penney, G. P., Desmedt, P., Buzug, T. M., Hill, D. L. G., & Hawkes, D. J. (1997). Voxel-based 2-D/3-D registration of fluoroscopy images and CT scans for image-guided surgery. *IEEE Transactions on Information Technology in Biomedicine*, *1*(4), 284–293. https://doi.org/10.1109/4233.681173

Yan, G., He, F., Shi, C., Wei, P., Cai, X., & Li, Y. (2023). Joint Camera Intrinsic and LiDAR-Camera Extrinsic Calibration. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 11446–11452. https://doi.org/10.1109/ICRA48891.2023.10160542

Yao, J., Ruggeri, M., Taddei, P., & Sequeira, V. (2010). Automatic scan registration using 3D linear and planar features. *3D Research*, *1*, 1–18. https://doi.org/10.1007/3DRes.03(2010)06

Zhuang, Y., Jiang, N., Hu, H., & Yan, F. (2013). 3-D-Laser-Based Scene Measurement and Place Recognition for Mobile Robots in Dynamic Indoor Environments. *IEEE Transactions on Instrumentation and Measurement*, *62*, 438–450. https://api.semanticscholar.org/CorpusID:14904326