Efficient Camera Pose Estimation Approach for Infrastructure Inspection

Anas Alsharo[1], Max Midwinter[1], and C.M. Yeum[1]

[1]Civil and Environmental Engineering Department, University of Waterloo, Canda
aaalshar@uwaterloo.ca, mxxmidwi@uwaterloo.ca, cmyeum@uwaterloo.ca

Abstract

Visual inspection of civil infrastructure assisted by Unmanned Aerial Vehicles (UAVs) witnessed significant improvements due to the rapid development of drone-mounted cameras and sensors. In visual inspection, accurate pose estimation of the collected images is a pivotal task that enables registering images to pre-existing 3D scenes of the structure to identify the geometric relationship between the scene and the image. This allows for localizing regions of interest or defects in these images. In this work a novel image pose estimation technique is proposed to improve images pose estimation and registration of drone-collected images. The proposed method utilizes a 3D base Structure from Motion (SfM) model and pre-calibrated poses of base images to facilitate the registration and pose estimation of new query images via feature-based pose estimation framework. The method leverages deep-based feature matching to generate dense 2D correspondences to simultaneously generate the 2D and 3D correspondences needed to execute the perspective-n-point solver and accurately estimate the pose. The proposed approach was tested against drone-collected images of cell tower and the image registration efficiency was evaluated through ROI localization in the registered images.

Keywords –

Visual Inspection; Asset Management; Image Localization; Telecommunication Cell Tower

1 Introduction

Condition assessment of large and hard-to-access infrastructure is a complex task and often requires trained technicians to employ time- and resource-extensive techniques for physical inspection [1]. The rapid development in UAVs and their associated sensors has led to a wide range of applications in condition assessment of civil infrastructures [2]. However, UAVs (e.g., drones) can be considered a multifunctional data acquisition system. Hence, robust inspection algorithms are required to leverage and fuse the various data types that drones can acquire to make sure of the acquired data.

A widely adopted application of drones in condition assessment of civil infrastructure involves utilizing them to capture high-resolution images, which are then used to construct SfM models of the structure. For example, in bridge inspection, SfM models enhanced damage quantification [3], crack identification and visualization [4], and risk assessment and management [5]. Xiao et. al. utilized a deep learning-based point cloud segmentation algorithm to assist ROI extraction and crack identification in bridges [6]. For large highway truss structures, Yeum et. al. proposed autonomous ROI localization algorithm, and ROI classification algorithm [7]. Visual inspection using drone images was also implemented for crack and rust identification within ROIs extracted from monopole tower images [8] and produce a health index of power transmission tower [9].

In visual inspection, determining the 6 degrees of freedom (DoF) pose of the image (3 for translation and 3 for rotation) enables establishing a geometric relationship between a 3D scene of the structure and the images. In the SfM framework, the simultaneous 3D reconstruction of a scene and pose estimation of images are core processes. This is typically achieved by extracting correspondences between 2D keypoints across multiple images and solving for camera poses and 3D points simultaneously. This implies that in case of visual inspection that involves constructing SfM model using large volume of images, the images are registered automatically as a natural product of the 3D reconstruction.

On the other hand, the inspection process for civil infrastructure typically does not involve constructing an SfM model; instead, it primarily relies on visually analyzing high-resolution images of the structure, collected from specific targeted regions of interest (TRIs). Hence, SfM-enabled pose estimation is not possible and in case the inspector needs to estimate the pose of these images, then a pose estimation technique should be employed that works on single image, or image sequences [10].

Image pose estimation is an extensively explored topic in the field of computer vision. A critical review by Xu et al. categorized pose estimation methods into two main categories: structure feature-based localization methods and regression-based pose estimation methods [11]. In structure feature-based methods, a 3D scene of the structure is employed to enable building correspondences between 2D points in a query image (image to be localized) and 3D points from the scene to enable estimating the pose of the image. Regression-based methods employ convolutional neural networks and deep neural networks to regress the pose of RGB images (e.g., PoseNet [12])

Structure feature-based methods are widely used due to their robustness and accuracy [11]. The 2D-3D correspondences in this method are typically established by matching distinctive features from a 2D query image with a database of 3D points, each associated with corresponding feature descriptors. Establishing these 2D-3D correspondences enables the estimation of the camera's position and orientation relative to the scene using a geometry constrain solver (e.g., Perspective-n-Point [13]).

A major drawback of this approach is the assumption that a database of descriptors is available and associated with the 3D points of the scene. However, the outputs of commercial 3D reconstruction and SfM platform does not typically involve image feature descriptors. Hence, a computationally intensive step is needed to establish the descriptors database for all base images. Another drawback related to the database of features descriptors is that the same feature extraction and description method should be used for both the base and query image to ensure consistency and accuracy [14]. Accordingly, if the feature extraction and description algorithm used in building descriptors database exhibits poor performance in matching image pairs with significant change in image environment (e.g., perspective, scale, or illumination), then finding proper number of matches between database and query images might not be possible.

On the other hand, assuming that a database of descriptors is established, if the 3D point cloud (SfM model) of the structure is not dense enough, then establishing high-quality correspondences between the database of descriptors and the query image's descriptors cannot be achieved.

To overcome the aforementioned drawbacks, an efficient structure feature-based image pose estimation approach is proposed to enable registering query images to the 3D scene without the need for database of descriptors and that can perform well in registering query images captured with significant variations in scale, perspective, zoom level, and background visibility compared to the database images.

The proposed algorithm leverages deep-learning-based local feature matchers to build dense 2D-to-2D matches between database images with known calibrated image poses and the query image. Then, the dense matches are used to simultaneously determine the 2D features and the 3D corresponding points, needed to estimate the pose. To estimate the image pose, a perspective-n-point solver within a Random Sample Consensus (RANSAC) framework was implemented.

The proposed approach was implemented on drone-collected cell tower images to test the accuracy of the approach in estimating the position and rotation of query images of the cell tower. The reliability of the proposed method was evaluated by utilizing the projection matrix calculated from the estimated pose in localizing regions of interest (ROIs) in the query images.

2 Methodology

2.1 Overview of the Technique

UAV-based inspection of civil infrastructure involves building SfM model from a large set of sequentially captured images of the structure. The process of 3D reconstruction outputs 3D map of the structure (hereafter, base model) along with calibrated camera poses (hereafter, base images).

In many cases, inspectors collect few images (hereafter, new images) around a specific TRI of high importance and the number of images might not be enough to build an SfM model. Hence, a reliable method is needed to enable registering these images and extract their poses without the need for a 3D reconstruction step. Registering the new images enable building the geometric relationship between the 3D scene and the new images which, in return, enable retrieving information from these images. To allow registering any new image to the base model, a pose estimation approach is proposed.

The approach, shown in Figure 1, starts with a base image that clearly shows the TRI where the inspector needs to perform a focused inspection by collecting new images around that region with varying camera angle, distance from the region, zoom level, or perspective. Prior to registering the new images, a vital step is to enable the inspector to retrieve the closest image relative to the base image the inspector started with. A geometry-based closest image retrieval algorithm is proposed, and it retrieves all the closest images within an applicable pre-defined criteria set by the inspector. From the set of retrieved images, the inspector can select the image with the clearest view and highest details of the all the ROIs that needs to be inspected.

The main enhanced step of this paper is to propose a registration approach that addresses the limitations of existing methods discussed in Section 1. Unlike conventional algorithms, the proposed method enables registering new images of hard-to-access infrastructure with a pre-existing SfM model, even if created by commercial SfM platforms that lack 3D-associated feature descriptors or depth maps, without relying on such data. This broadens the applicability and utility of the registration process.

For registering the new image and in addition to the base image the inspection started with and the new image that was selected from the pool of retrieved closest images, the concept of shadow image is proposed. Shadow image is another image from the set of base images that is adjacent to the original base image. To automate the process of selecting the shadow image, the proposed algorithm extracts the shadow image by bringing the image next in sequence relative to the original base image (base images are collected sequentially). The poses of the original base image and the shadow image are known from the SfM output. What keeps missing is the pose of the new image.

The foundational step for registering the new image in the proposed approach is extracting dense 2D-to-2D correspondences between the three images: the original base image, the shadow image, and the new image. From the dense correspondences, the common extracted matching features in the three images are retained. Since the poses of the original new image and the shadow image are known, then a direct triangulation is performed on the common matched features between these two images and a 3D point cloud is generated. Moreover, from the common correspondences of the three images, the 2D-to-2D matched features between the original base image and the new image is extracted. The aforementioned steps simultaneously generate the 3D points and their corresponding 2D projections in the new image and these 2D-to-3D correspondences are then fed to a PnP solver to solve the geometric constrain relationship and estimate the 6 DoF pose of the image.

An important outcome of registering the new image to the SfM model is the ability to calculate the new image's projection matrix, which, in turn, facilitates the retrieval of ROIs from the new image. In this study, retrieving ROIs means allowing the inspector to select an ROI from the original base image, and the algorithm autonomously identifies and retrieves the corresponding ROI from the registered new image. The selection of the ROI is achieved in this study by enabling the inspector to simply draw a bounding box around the ROI in the original base image and to retrieve the same ROI form the new image. Hence, interacting with the 3D model to select the ROI points and project the ROI on the new image is not needed. More details provided in the subsequent sections explaining the details of the approach.

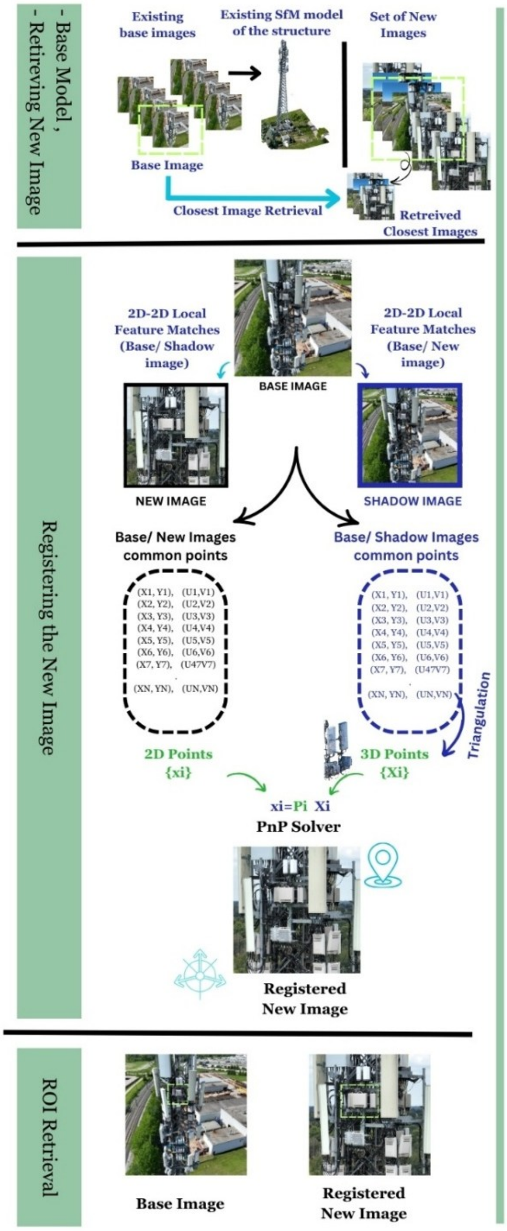

Figure 1. Proposed methodology overview

2.2 Base SfM Model

The SfM technique has been widely adopted for its ability to create accurate 3D scene reconstructions. SfM establishes correspondences from overlapping images to concurrently calibrate unknown camera parameters (intrinsic and extrinsic) while reconstructing the 3D scene. Since SfM can construct the scene and calibrate the extrinsic (position and rotation) of each camera (image), SfM technique was utilized to extract the poses of the base images.

Building high-quality SfM model requires feeding large number of properly overlapping images to the 3D reconstruction algorithm. SfM quality depends on the quality of features matches across the sequenced images. Hence, employing high-resolution camera to collect the images is necessary to ensure reconstruction quality.

In this work, since the algorithm is implemented on a hard-to-access structure (cell tower) a commercial drone equipped with high-resolution camera was used to collect the images needed for the 3D reconstruction. The images were collected along a pre-defined path designed using a drone mission planning software to ensure the systematic overlapping of the collected images.

2.3 Retrieving Closest New Image(s)

Traditional image retrieval algorithms rely on a database of precomputed image descriptors linked to set of database images. Distance between descriptors in the new and database images are computed and the images with the highest correspondences are retrieved. However, these approaches are computationally demanding and depend on predefined descriptors databases, which may not be available and also, cause potential mismatch.

For UAV-based inspections, drones equipped with sensors like GPS, IMU, and compass typically record and associate metadata or exchangeable image file format (EXIF) data such as camera position, altitude, and heading angle with each collected image. Accordingly, A fast trigonometry-based retrieval algorithm leverages images metadata was developed to efficiently retrieve closest images without relying on complex feature-based methods.

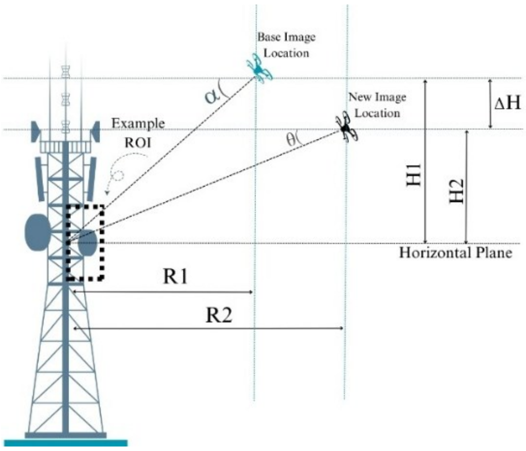

Figure 2. Image Retrieval Illustration

Figure 2 illustrates the proposed technique for retrieving closest new image(s) using a single base image. In the shown illustration, for an arbitrary ROI, a base image is assumed to be captured at a distance (R1) from the ROI with camera pitch angle (α). For the same ROI, a new image is assumed to be captured at distance (R2) from the ROI with camera pitch angle (θ) and looking at the ROI (the main region to be inspected is expected to be near the centre of the image). Subsequently, relative altitudes between base camera and ROI horizontal plane (H1) and between new camera and ROI horizontal plane (H2) and the calculated difference in altitude (ΔH_{calc}) can be calculated as follows:

$$H1 = R1 * \tan(\alpha) \quad (1)$$

$$H2 = R2 * \tan(\theta) \quad (2)$$

$$\Delta H_{calc} = R1 * \tan(\alpha) - R2 * \tan(\theta) \quad (3)$$

The trigonometric relationships in Equations (1-3) form the foundation for EXIF-based image retrieval algorithm. To retrieve the new image(s) closest to the base image, the algorithm starts by reading the heading angle for the base image and all the new images. New images with similar heading to the base image are retained and remaining images are ignored. Then, for the base image and all retained new images, the algorithm will read the longitude, latitude, altitude, and camera pitch angle of each image and calculate (R1), (R2), (H1), (H2), and subsequently, the actual altitude difference (ΔH_{actual}) between the base camera and new camera. Simply, closest images can be filtered by selecting new images with (ΔH_{actual}) close to (ΔH_{calc}). To filter images, either by heading or by ΔH similarity, the inspector can set threshold for the tolerance of the heading or the altitude differences. With higher tolerance, images with larger change in perspective might be retrieved, and vice versa.

This algorithm aims to enhance the speed and quality of UAV image retrieval for images with GPS information. In the case of operating the drone in GPS-denied zones, the conventional content-based image retrieval algorithm can be used to retrieve the closest image.

2.4 New Image Registration

This study introduces improved approach for image registration that leverages state-of-the-art deep learning-based feature matching algorithms to simultaneously construct the 3D scene points and the 2D corresponding points of the new image without the need for pre-defined database of feature descriptors.

As discussed in the “Overview of The Technique” section, three images are utilized to enable registering the new image: base image, its associated shadow image, and a retrieved new image. Robust feature matchers are required to ensure extracting dense matches between the base, shadow, and new image even if the new image was captured under challenging conditions compared to the base image. Deep learning-based feature matching algorithms are employed to ensure accurate and dense feature extraction, even in challenging conditions like varying lighting, perspective, and scale [15].

Accurately estimating pose of the new image enable calculating the projection matrix. Then, the algorithm can retrieve ROIs selected from the base images and localize the same ROIs in the query images.

3 Experimental Validation

To validate the proposed approach, a lattice telecommunication cell tower in Waterloo (Ontario, Canada) was selected for testing the proposed algorithm. A cell tower was selected for validation as it is type of large and hard-to-access structures that requires systematic and frequent inspections due to its significant and importance as a vital telecommunication infrastructure. The selected tower contains multiple components with highly overlapping wiring and equipment. The cell tower height is approximately 47 meters, with a base width of around 5.5 meters and a width of 4 meters at the location of the tower components.

3.1 Base Cell Tower SfM Model

To construct the base model, we used a commercial DJI Mavic 3 Enterprise drone to capture high-resolution images of the cell tower for the purposes of building high quality SfM model. The drone comprises dual cameras: wide-view camera for regular images, and tele camera with 7x optical zoom. Moreover, the drone is equipped with RTK-GPS. Accordingly, with each collected image, the set of extracted EXIF data contain information needed for executing the various steps of the proposed algorithm. The open-source tool, ExifTool [16], was used to retrieve EXIF data for each image.

To build high quality SfM model, a set containing large number of high-quality and overlapping images of the cell tower was collected. For better results, the collected images should be sequenced and capturing the different components of the cell tower at different perspectives. To ensure systematic collection of images, a commercial software, Drone Harmony [17], for flight path planning and autonomous drone mission was utilized. The planned flight path for image collection was a helical path with a proper pitch distance between helix turns that ensures proper vertical overlap of images. The horizontal overlap of images is guaranteed as the drone is always directed towards the centre of the cell tower. The result of image collection step was a set of 409 images fully covering all the regions of the cell tower. The collected images were then fed to an SfM platform to 3D reconstruct the cell tower.

Different open-source and commercial SfM platforms were tested to create a 3D model of the tower and the platform with the highest quality was selected. In our study, a commercial SfM platform, One3D [18], was utilized to build the base SfM model. The software yielded a 3D reconstruction with relatively high quality compared to other tested platforms. The platform outputs the position and rotation of each camera which allows constructing the projection matrices of all base images.

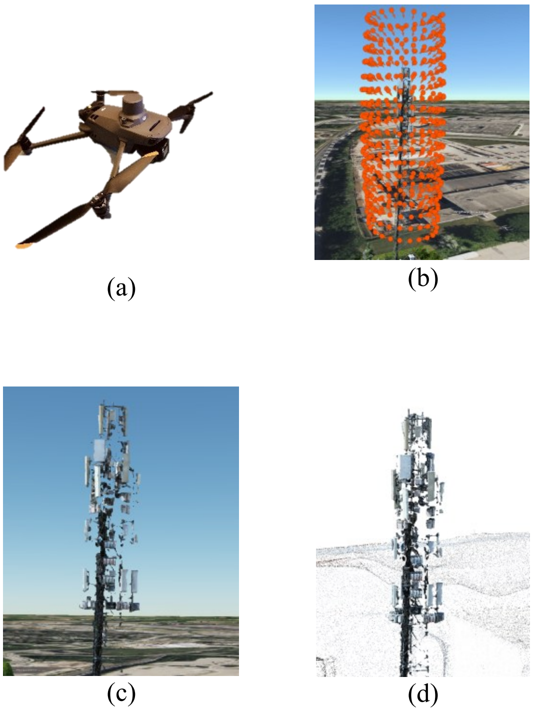

Figure 3: Base model 3D reconstruction (a) DJI Mavic 3D used in this study, (b) collected images and helical flight path (c) generated textured mesh (d) generated point cloud

3.2 Retrieving Closest Cell Tower Image

To validate the proposed image retrieval algorithm, a set of new images were collected at a different day and with different perspectives. New images were collected using wide-view camera and tele camera (7x zoom). In the post processing stage, the images were placed in the same folder directory of the base images to prepare for autonomous image retrieval.

As shown in Figure 4, the algorithm was able to successfully retrieve set of closest images relative to the base image. The algorithm retrieved the images taken with the wide-view camera and the tele camera. As shown in Figure 4, the base image was selected on the basis that the region to inspect located approximately at the center of the image, as shown within a bounding box in (a), the retrieved images clearly show that the main region of the base image is clearly shown in the retrieved images shown in (b). The inspector can retrieve more images by increasing the tolerance of heading and difference in altitude discussed in section “2.3 Retrieving Closest New Image(s)”.

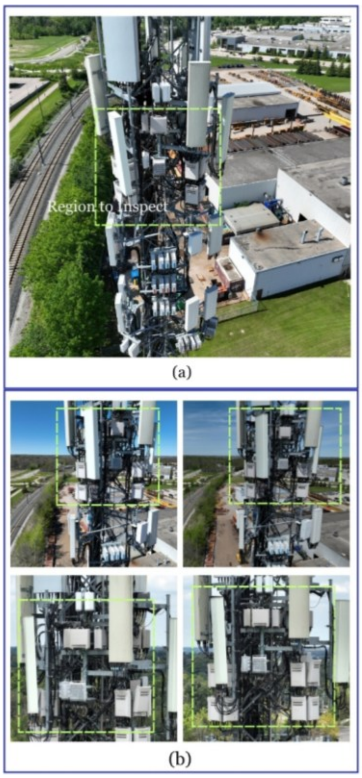

Figure 4: Image Retrieval Algorithm Outputs. (a) base image, (b) example retrieved closest images.

3.3 Image Registration Output

The first step in this approach is to select local feature matching algorithms that can build dense 2D-to-2D common correspondences between base and shadow image and between base and new image. By common correspondences we mean features that are shown in the base, shadow image, and new image, simultaneously.

Three state-of-the-art deep learning-based detector-free local feature matching algorithms were employed in this study to build the dense correspondences: Efficient LoFTR [19], RoMa [20], and AspanFormer [21]. The RoMa model creates feature pyramids by integrating coarse-level features with fine-tuned specialized features, using classification-based loss for global matches and robust regression loss for fine-tuning to improve correspondence accuracy. Efficient LoFTR is an extension to LoFTR (Local Feature Matching with Transformers) algorithm. It enhances LoFTR by refining attention mechanisms and incorporating a two-stage correlation layer to enhance efficiency and accuracy in feature matching. AspanFormer is another transformer-based algorithm that employs a hierarchical attention structure and adaptive attention spans to dynamically focus on relevant regions. This adaptive design ensures accurate feature matching across varying scales and challenging scenarios. Each algorithm approaches the feature matching problem from a different perspective. Hence, in this study, the three algorithms were cascaded together to build dense correspondences that are spread across the TRI. Figure 5 shows the results of common features between the base and shadow and between base and new image. It worth noting that all features captured in the non-overlapping regions in the three images were discarded to reduce outliers and that feature that are not common in the three images were excluded.

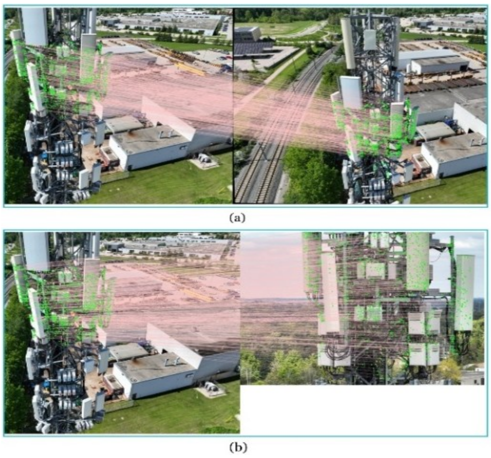

Figure 5: Feature Matching Output. (a) base image with shadow image, and (b) base image with the new image.

To build the 2D-to-3D correspondences required to localize the new image, the common features between the base and shadow images, show in Figure 5 (a), were triangulated to generate the 3D scene points and the common features between the base and new image, shown in Figure 5 (b) were extracted as the 2D projections.

The PnP algorithm was employed to minimize the reprojection error of a geometric constrain relationship. For PnP solver, and in addition to the 2D-3D correspondences, the intrinsic matrix **K** for the camera is needed and was acquired by calibrating the drone camera in the lab using camera calibration board. Additionally, a PnP solver type should be specified. For this study, the Efficient PnP solver (EPnP [22]) was implemented within a Random Sample Consensus (RANSAC) framework, in order to handle outliers.

To test the proposed algorithm, the estimated camera position of the new image should be compared to ground truth values. However, since new images have unknown ground truth values and base images have known calibrated poses, 10 base images were selected from the pool of base images and were treated as new images to evaluate the localization approach. Table 1 shows the absolute distance between actual and estimated camera positions.

Table 1. Evaluation of Image Localization Approach

Images ID	Error in Camera Position	Error in Camera Rotation (degrees)
1	0.066	0.739
2	0.022	0.180
3	0.090	0.078
4	0.075	0.544
5	0.078	0.094
6	0.061	0.685
7	0.017	0.314
8	0.041	0.089
9	0.094	0.371
10	0.064	0.620

It can be shown from Table 1 that the average error in camera position is around 0.06 m and the average error in rotation vector is 0.37°. This implies that the proposed dense feature matching-based camera pose estimation approach can accurately and reliably register new images.

In addition to the 10 base images, the whole dataset was tested to investigate the ability of the methodology to register images with different view angles and perspectives. The average mean error in camera position was found to be 0.069 m and the mean error in rotation vector was found to be 0.41°.

For the case of cell tower, and due to the fine and complex details on the structure, the image resolution was kept high during the feature matching process. This is needed to ensure building robust correspondences that guarantees high registration accuracy. High resolution images increase the inference time for deep learning-based matchers. In our experiment, an NVIDIA GEFORCE RTX 3080 GPU with 12 GB VRAM was used. The inference time for each pair takes ~2.5 seconds using AspanFormer, ~7 seconds using EfficientLOFTR, and ~9 seconds using RoMA. This inference time makes this method not compatible with real-time applications. However, the method is designed to enhance offline visual inspection and ROI retrieval from new images.

3.4 Localizing ROIs in New Cell Tower Images

Once the new image is localized, the inspector can map ROIs from base (old) images to new images to track and assess the condition changes of the ROI. In this paper, and to avoid the need to interact with the 3D scene of the structure to identify the ROI, the algorithm is designed so that the inspector selects an ROI in the base image and the features within that ROI and their corresponding features in the shadow image will be selected and triangulated using both images' projection matrices. The 3D points of the ROI are then projected back in the new image using the calculated projection matrix estimated using during image registration. Figure 6 presents different examples of ROIs selected in the base image (with the letter B in the corner) and the corresponding ROI retrieved from the new image (with the letter N). The last column in Figure 6 (with letter Z) is for zoomed retrieved ROI.

It worth mentioning that for this example the new image was selected from the set captured by the tele camera with 7x zoom. This reflects the robustness of the algorithm in localizing images and retrieving ROIs even in challenging scenarios and severe change in perspective, illumination, and scale. Figure 6 illustrates two main advantages of the proposed approach: 1) select ROIs within the region of focused inspection and retrieve the same ROI with clear details to assess condition inspection and temporal changes (as shown in first three rows of Figure 6), and 2) Inspect changes in asset dismantling or installation on the cell tower as shown in the last row where the retrieved ROI shows the change in asset installation on the selected region.

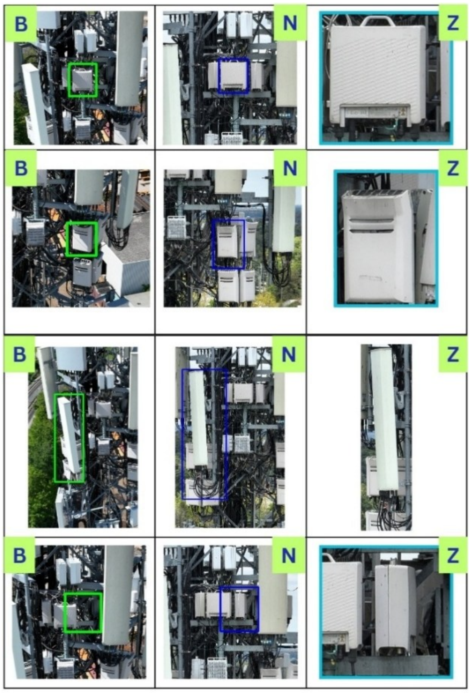

Figure 6: ROI retrieval results. First column for selected ROI in base image (B). Second column for retrieved ROI from the new image (N). Third column for zoomed in view of the ROI (Z).

4 Conclusion

This study introduces and validates an innovative method for enhancing image localization and pose estimation by leveraging state-of-the-art feature matching algorithms. Key contributions in this study include:

1. An efficient image registration method that overcomes the limitations of existing methods that require database of feature descriptors associated with 3D scene points.
2. Simplified metadata-based image retrieval algorithm for drone-collected images.
3. Accurate ROI retrieval, allowing inspectors to select ROIs on a base image, and automatically retrieve and localize the same ROI in the new image, without the need to interact with the 3D reconstruction for ROI selection.

The image localization algorithm achieved an absolute error in camera position of around 6 cm. Tests on new images from a 7x optical zoom camera shows the performance of the algorithm even in challenging scenarios of image environment. This approach significantly improves UAV-based inspections of large and hard-to-access infrastructure, offering broad applicability for diverse structures.

References

[1] B. F. Spencer, V. Hoskere, and Y. Narazaki. Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering*, 5(2):199–222, 2019.

[2] W. W. Greenwood, J. P. Lynch, and D. Zekkos. Applications of UAVs in civil infrastructure. *J. Infrastruct. Syst.*, 25(2):04019002, 2019.

[3] S. Chen, D. F. Laefer, E. Mangina, S. M. I. Zolanvari, and J. Byrne. UAV bridge inspection through evaluated 3D reconstructions. *J. Bridge Eng.*, 24(4):05019001, 2019.

[4] L. Zhou et al. UAV vision-based crack quantification and visualization of bridges: system design and engineering application. *Structural Health Monitoring*, 2024.

[5] M. Mandirola, C. Casarotti, S. Peloso, I. Lanese, E. Brunesi, and I. Senaldi. Use of UAS for damage inspection and assessment of bridge infrastructures. *International Journal of Disaster Risk Reduction*, 72:102824, 2022.

[6] J.-L. Xiao, J.-S. Fan, Y.-F. Liu, B.-L. Li, and J.-G. Nie. Region of interest (ROI) extraction and crack detection for UAV-based bridge inspection using point cloud segmentation and 3D-to-2D projection. *Automation in Construction*, 158:105226, 2024.

[7] C. M. Yeum, J. Choi, and S. J. Dyke. Automated region-of-interest localization and classification for vision-based visual assessment of civil infrastructure. *Structural Health Monitoring*, 18(3):675–689, 2019.

[8] N. M. Shajahan, T. Kuruvila, A. S. Kumar, and D. Davis. Automated inspection of monopole tower using drones and computer vision. In *Proceedings of the 2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS)*, pages 187–192, Singapore, Singapore, 2019.

[9] H. Manninen, C. J. Ramlal, A. Singh, J. Kilter, and M. Landsberg. Multi-stage deep learning networks for automated assessment of electricity transmission infrastructure using fly-by images. *Electric Power Systems Research*, 209:107948, 2022.

[10] Z. A. Al-Sabbag, C. M. Yeum, and S. Narasimhan. Enabling human–machine collaboration in infrastructure inspections through mixed reality. *Advanced Engineering Informatics*, 53:101709, 2022.

[11] M. Xu et al. A critical analysis of image-based camera pose estimation techniques. *Neurocomputing*, 570:127125, Feb. 2024.

[12] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. *arXiv preprint arXiv:1505.07427*, 2015.

[13] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[14] E. Deretey, M. T. Ahmed, J. A. Marshall, and M. Greenspan. Visual indoor positioning with a single camera using PnP. In *Proceedings of the 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–9, Banff, AB, Canada, 2015.

[15] S. Xu, S. Chen, R. Xu, C. Wang, P. Lu, and L. Guo. Local feature matching using deep learning: A survey. *Information Fusion*, 107:102344, 2024.

[16] *ExifTool by Phil Harvey*. On-line: http://exiftool.org, Accessed: Dec. 26, 2024.

[17] *Drone Harmony*. On-line: https://app.droneharmony.com, Accessed: Dec. 22, 2024.

[18] *One3D - guide*. On-line: https://app.one3d.ai/dashboard/guide, Accessed: Dec. 22, 2024.

[19] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. *arXiv preprint arXiv:2305.09016*, 2024.

[20] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg. RoMa: Robust dense feature matching. *arXiv preprint arXiv:2305.15404*, 2023.

[21] H. Chen et al. ASpanFormer: Detector-free image matching with adaptive span transformer. *arXiv preprint arXiv:2208.14201*, 2022.

[22] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An accurate O(n) solution to the PnP problem. *Int. J. Comput. Vis.*, 81(2):155–166, 2009.