Comparative Study of Structure from Motion on Construction Site

Mingyun Kang[1] Sangmin Lee[1] Sebeen Yoon[1] and Taehoon Kim[1]

[1]Architecture Engineering, Seoul National University of Science and Technology, Republic of Korea

kmg4312@seoultech.ac.kr, lsm010714@seoultech.ac.kr, paul6452@seoultech.ac.kr, kimth@seoultech.ac.kr

Abstract

For progress monitoring in the construction industry, understanding the state of construction sites is crucial. However, traditional manual inspection methods are labor-intensive and time-consuming. To address these challenges, various methods for creating 3D models of job sites have been explored. Professional equipment such as LiDAR and laser scanners offer the most accurate means of generating point clouds (PCDs) and constructing 3D models. However, these tools are expensive, cumbersome, and often impractical for frequent use in dynamic construction environments. Recently, with the advancement of deep learning, 3D reconstruction techniques have been extensively studied and applied across various fields. Among these, Structure from Motion (SfM) stands out as a method capable of generating PCDs and estimating camera poses. Based on advancing capabilities of SfM, many research has been conducted to measure progress monitoring in construction field. However, most studies that have utilized SfM for progress monitoring have acquired a large number of images and ensured significant overlap in input data to enhance the robustness of the 3D model. While this approach provides a highly accurate 3D reconstruction, the image acquisition process itself introduces additional labor-intensive tasks. Therefore, this study aims to adhere to the fundamental nature of 3D reconstruction by evaluating the performance of various SfM models using only 26 images captured from a brief video recording at a construction site. The findings aim to evaluate the applicability of various SfM technologies with limited data, in real-world construction scenarios and finally provide insights into their potential and future directions.

Keywords –

Progress Monitoring, Structure from Motion, COLMAP, VGGSfM

1 Introduction

In the construction industry, progress monitoring is essential to ensure project efficiency, quality, and adherence to timelines. Traditional monitoring methods often rely on manual inspections and reports, which can be time-consuming, error-prone, and inconsistent. As a solution, 3D modeling techniques have gained attention for their ability to provide comprehensive information of construction sites [1, 2, 3].

To generate 3D models of construction sites, equipment such as LiDAR and laser scanners are widely used, offering the ability to produce highly accurate and dense point clouds (PCDs). For instance, Hu et al. utilized LiDAR to scan 3D surfaces of opencast quarries, enabling the analysis of slope stability. Similarly, Wang et al. employed laser scanners to rapidly perceive heavy construction equipment such as tower crane, improving safety and productivity in construction environments. Despite their accuracy and utility, these tools are expensive, bulky, require extensive training to operate, and involve significant time for scanning, making them less practical for dynamic and fast-paced construction site conditions [3, 4, 5].

These challenges have driven researchers to explore alternative methods for 3D reconstruction. With the advancement of deep learning and computer vision, techniques using 2D images to create 3D models have garnered significant interest [1]. While such methods may compromise lower accuracy compared to specialized equipment, their accessibility, ease of implementation, and reduced scanning time make them a promising alternative [1, 4]. Among these approaches, Structure from Motion (SfM) stands out as a photogrammetric technique capable of not only constructing 3D models from 2D images taken from varying perspectives, but also estimating camera poses which include their positions and orientations [6].

Using these strengths of SfM, research on progress monitoring based on SfM in the construction field has been actively conducted. However, these studies

focusing on the 3D reconstruction of actual construction sites, particularly interior environments, remain insufficiently explored. Additionally, achieving higher accuracy in SfM models often requires a large number of images and to be overlapped to some extend. Hence, acquiring such extensive image datasets in complex construction sites is challenging and may lead to even more cumbersome issues.

By utilizing image data captured within a very short time frame and comparing multiple SfM frameworks, this study aims to identify the challenges and opportunities of applying SfM in construction environments, thereby contributing to a deeper understanding of its potential in the industry.

2 Literature Review

2.1 Structure from Motion (SfM)

Structure-from-Motion (SfM) is a technique that builds 3D point clouds (structure) from cameras (motion). SfM is typically categorized into incremental and global approaches. In incremental SfM, the process begins by aligning an initial pair of images, followed by the gradual addition of new images. This method iteratively performs feature matching and bundle adjustment (BA), progressively constructing the 3D structure. In contrast, global SfM employs deep learning-based feature matching across multiple images, enabling a faster and more efficient reconstruction process. This approach estimates camera poses and reconstructs the 3D structure in a global, batch-wise manner rather than incrementally.

COLMAP [7] is one of the dominant incremental SfM tool in computer vision field due to its robustness and efficiency. It employs the Scale-Invariant Feature Transform (SIFT) algorithm, which ensures reliable feature matching even under image rotations and lighting variations. Furthermore, COLMAP uses the Ceres Solver-based BA to reduce accumulated errors, thereby enhancing reconstruction accuracy. As mentioned before, with the advancement of deep learning, global SfM methods also have been developed.

VGGSfM [8], one of the newest global SfM model, has been developed as an alternative to traditional incremental SfM methods. Unlike conventional approaches, VGGSfM jointly estimates all camera poses using a Transformer-based architecture, offering a simpler and more differentiable process compared to the combinatorial correspondence chaining step. Additionally, for BA, VGGSfM replaces the nondifferentiable Ceres Solver with the fully differentiable second-order Theseus Solver, further improving computational efficiency and scalability.

2.2 SfM in Construction Site

Although SfM cannot build 3D models as accurately as TLS or LiDAR, it has been widely studied due to its affordability, accessibility, and ability to perform spatial mapping to a certain extent. Ding et al. explored indoor scene understanding using 304 images captured with a widely used consumer-level camera and conducted 3D reconstruction using VisualSfM [5]. The strong 3D reconstruction capability significantly contributes to scene understanding, which has led to extensive research on its applications in progress monitoring.

Studies have also been conducted to evaluate COLMAP's powerful performance within the construction domain, emphasizing its potential applicability in this field. Keyvanfar et al. applied nine different 3D reconstruction models, including COLMAP, to visualize construction sites [9]. In their study, they utilized 138 images captured using an Unmanned Aerial Vehicle (UAV), ensuring a more robust 3D reconstruction by acquiring images with 60–80% overlap.

In addition to studies that solely utilized SfM for progress monitoring, research integrating SfM with other technologies has also been conducted. Mahami et al. and Han et al. employed SfM combined with Multi-View Stereo (MVS) to construct 3D models and align them with as-built Building Information Model (BIM) elements for progress monitoring [10, 11]. Additionally, some studies have leveraged SfM-based camera pose estimation to generate denser 3D models. For instance, Pal et al. developed an SfM-MVS-based 3D reconstruction pipeline to estimate camera parameters from 2D images and incorporated Neural Radiance Field (NeRF) and semantic segmentation to predict progress completion [1].

Although numerous studies have explored the application of SfM in the construction industry, many of these approaches have prioritized achieving more robust and dense 3D reconstructions by utilizing high-resolution cameras or capturing images with significant overlap to enhance feature matching. Consequently, rather than fully exploiting SfM's inherent advantages, these methods introduce constraints related to data acquisition, limiting their practical applicability.

3 Case Study

3.1 Overview

To conduct a 3D reconstruction study in a real construction environment, we visited a construction site located in Seongdong-gu, Seoul, and acquired image data. A specific room within the construction site was selected to compare the performance of different 3D reconstruction methods. Image acquisition was

performed using the Insta ONE X2 fisheye camera, which facilitated rapid data collection by capturing images in all directions with a single shot. The captured footage was converted into perspective images, and a total of 26 images from the selected room were used for the comparative analysis between COLMAP and VGGSfM. All of the converted images have resolution of 4320 × 2880. The workstation used for the case study was equipped with 32 GB of RAM, a Core i7-12700KF CPU, NVIDIA GeForce RTX 3080 Ti GPU and CUDA version 11.3 was utilized for GPU acceleration.

3.2 Evaluation Metrics

To quantitatively compare the results of 3D reconstruction, not only the number of 3D points and processing time but also three key metrics were used: mean track length (MTL), mean observations per image (MOI), and mean reprojection error (MRE). These metrics provide insights into the robustness, contribution, and accuracy of the reconstruction process [12].

MTL is the average number of images in which a 3D point is observed. It measures the frequency with which 3D points are visible across the set of input images, providing insight into the robustness of feature matching and triangulation in the reconstruction process.

$$\text{MTL} = \frac{\sum_{i=1}^{N} L_i}{N} \tag{1}$$

Where N denotes total number of 3D points in reconstruction, L_i the number of images in which the i-th 3D point is observed.

MOI is the average number of 3D points observed per image. This metric evaluates how much each image contributes to the 3D reconstruction, indicating the distribution of feature matches across the dataset.

$$\text{MOI} = \frac{\text{Total Observations}}{\text{Number of Images}} \tag{2}$$

MRE quantifies the average distance, in pixels, between the observed 2D points and their corresponding 2D projections from the 3D points. It serves as a key indicator of the accuracy of the reconstruction and the quality of the camera parameters.

$$\text{MRE} = \frac{\sum_{i=1}^{N} |\boldsymbol{p_i} - \widehat{\boldsymbol{p}_i}|}{N} \tag{3}$$

Where N denotes total number of 2D points, p_i the observed 2D coordinates of the i-th point, $\hat{p}_i$ the reprojected 2D coordinates of the i-th point obtained from the estimated 3D point and camera parameters.

4 Result Analysis

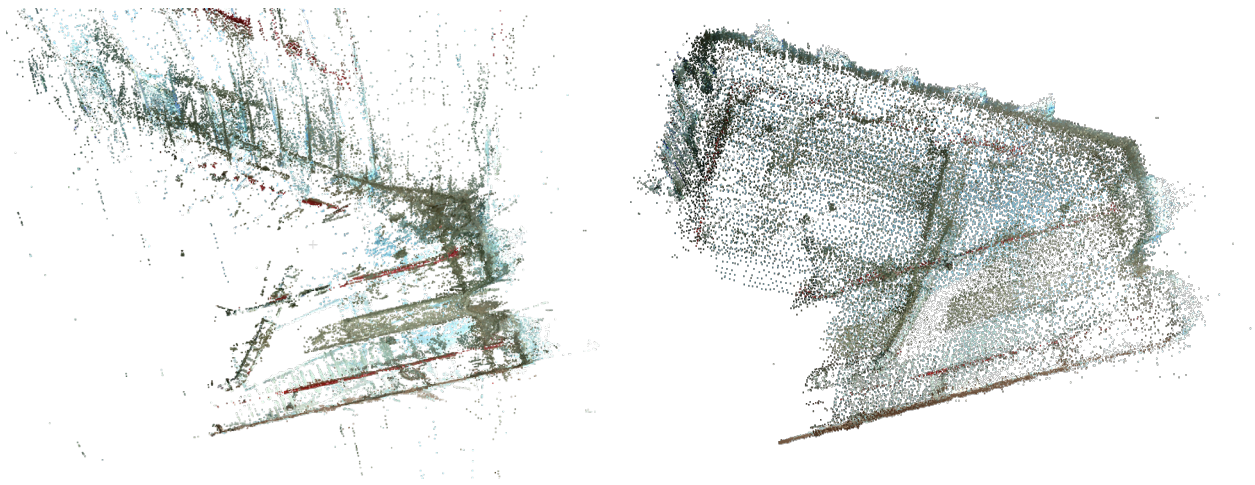
4.1 Quantitative analysis

The quantitative comparison of COLMAP and VGGSfM using the aforementioned metrics is summarized in the following Table 1.

Table 1. Summary of quantitative comparison of COLMAP and VGGSfM

Metrics	COLMAP	VGGSfM
Number of 3D points	13,765	**69,612**
Processing time	30m 33s	**10m 35s**
MTL	**1.22082**	0.66221
MOI	1,377.38	**1,993**
MRE	1.08028	**0.01852**

* Bold values indicate better performance



(a) COLMAP (b) VGGSfM

Figure 1 Top view perspective of 3D reconstruction model of COLMAP(a) and VGGSfM(b)

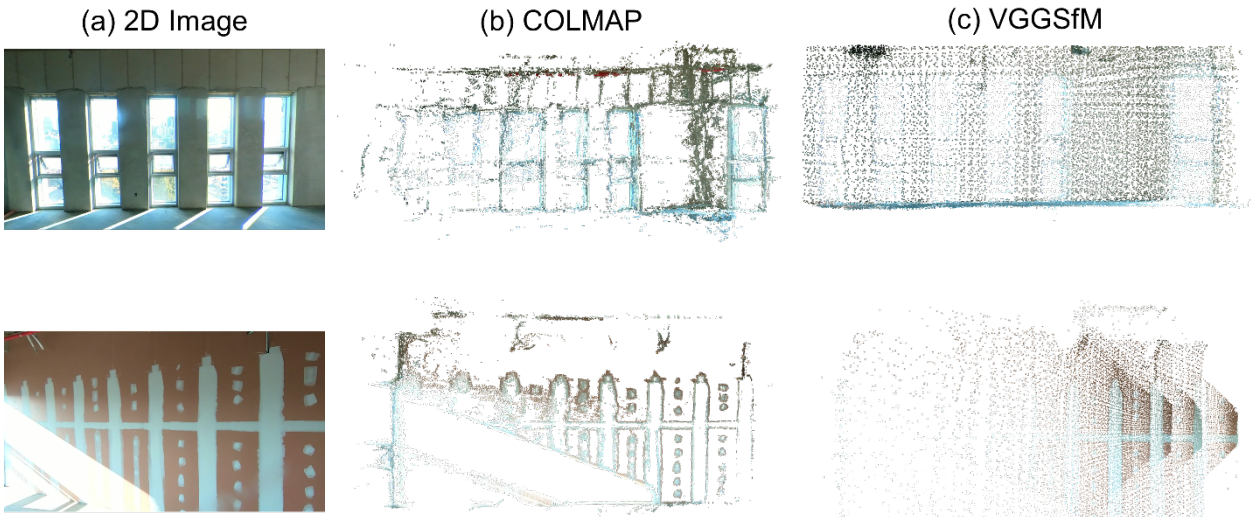


Figure 2 Reconstruction outputs of wall: (a) 2D image of wall (b) COLMAP (c) VGGSfM

The number of generated PCDs was significantly higher for VGGSfM, with 69,612 PCDs compared to 13,765 PCDs for COLMAP, a difference of approximately 55,000 points. In terms of processing time, VGGSfM was approximately 20 minutes faster than COLMAP, completing the task in 10 minutes and 35 seconds compared to COLMAP's 30 minutes and 33 seconds.

While MTL was 0.6 higher for COLMAP, for MOI, VGGSfM showed a higher value by about 600, indicating that images contributed more effectively to PCD generation. Moreover, the MRE for VGGSfM was about 1 lower than COLMAP, suggesting that VGGSfM's PCDs better explain the observed data.

Overall, VGGSfM demonstrated superior performance compared to COLMAP in all metrics except for MTL, which can be attributed to dataset-specific characteristics rather than model performance.

4.2 Qualitive analysis

Figure 1 presents the results of 3D reconstruction projected from a top-view perspective. In the case of COLMAP (Figure 1(a)), the room's shape appears inconsistent and scattered compared to VGGSfM (Figure 1(b)). Additionally, the floor area in the middle of the room is poorly reconstructed in COLMAP, whereas VGGSfM provides a more complete representation.

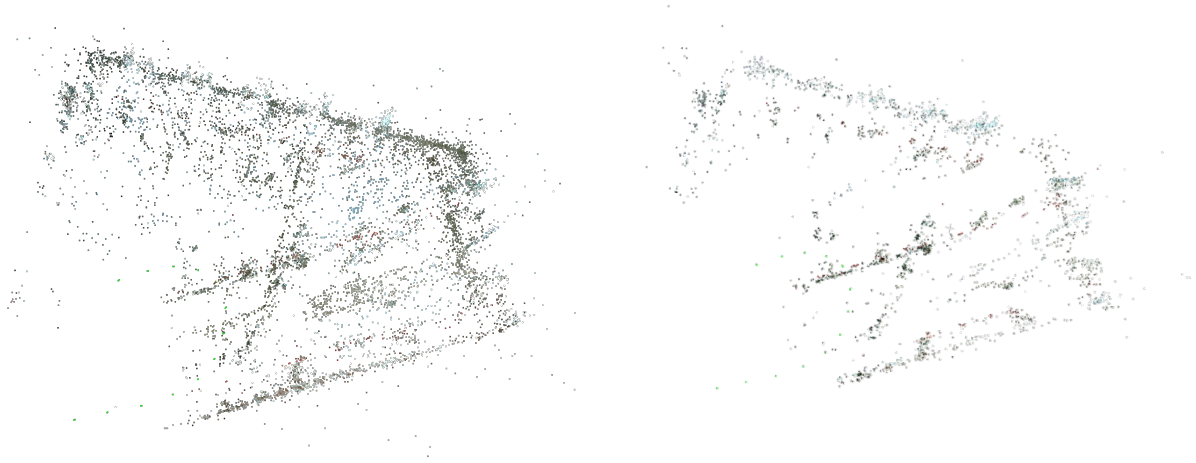
Figure 2 compares the reconstruction outputs of a wall using a 2D image (Figure 2(a)) as input, with results from COLMAP (Figure 2(b)) and VGGSfM (Figure 2(c)). As shown, COLMAP effectively captures distinctive lines and patterns, demonstrating its strength in feature-specific reconstruction. In contrast, VGGSfM produces a more uniformly distributed PCDs, reflecting a balanced and comprehensive reconstruction.

On the other hand, both models share a common limitation compared to laser scanners, as they rely solely on 2D input data for reconstruction. This constraint results in missing information in areas with bright occlusion or insufficient feature data.

5 Discussion

In this section, we compare not only COLMAP and VGGSfM but also various other SfM models. Meshroom [13] is an incremental SfM model that offers ease of use through a graphical user interface (GUI), whereas OpenMVG [14] is a global SfM model based on a command-line interface (CLI). The same set of 26 images used in the case study was used as input data for both models.

The processing time for Meshroom was approximately 21 seconds, while OpenMVG took 19 seconds. The number of generated 3D points was 8,587 for Meshroom and 3,174 for OpenMVG. Although both models demonstrated significantly faster processing times compared to COLMAP and VGGSfM, their reconstructed 3D models were insufficient for scene understanding, as shown in Figure 3. On the other hand, considering the perspective of SfM's camera pose estimation accompanying with fast processing speed, these models can serve as effective preprocessing tools for generating dense 3D models using methods such as MVS or NeRF. Their ability to provide camera positions makes them suitable for applications where precise pose estimation is essential for subsequent reconstruction processes.



(a) Meshroom (b) OpenMVG

Figure 3 Top view perspective of 3D reconstruction model of Meshroom(a) and OpenMVG(b)

Table 2. Summary of Meshroom and OpenMVG

Metrics	Meshroom	OpenMVG
Number of 3D points	**8,587**	3,174
Processing time	21s	**19s**

* Bold values indicate better performance

6 Conclusion

In this study, we explored the applicability of SfM models, in real construction environments. For this purpose, we compared two SfM models: COLMAP, one of the most widely used SfM models, and VGGSfM, a more recently developed framework. Comparison revealed that VGGSfM outperformed COLMAP in terms of the number of 3D points and processing time. Additionally, VGGSfM demonstrated superior performance in all metrics except for MTL.

Qualitative comparisons also highlighted VGGSfM's advantages. VGGSfM more accurately depicted the arrangement of windows and walls, whereas COLMAP exhibited scattered and poorly defined spaces. In contrast, VGGSfM effectively delineated a finite outline of the room, successfully reconstructing the enclosed space.

In discussion section, we also utilized two more SfM model of Meshroom and OpenMVG. These SfM has highly faster processing time of no more than 1 min, however, their number of 3D points were in proportion to the processing time. This result indicates that these models are not adequate for 3D reconstruction of scene understanding, on the other hand, in the perspective of preprocessing for building dense 3D models, they can serve camera poses of images in a short time.

Besides, this study remains certain limitations. Since SfM relies solely on 2D images as input, bright occlusions can lead to a lack of PCDs. Also, the 3D reconstruction was conducted for a single room within the construction site. When attempting to reconstruct multiple rooms, challenges arise due to the inability to extract overlapping features from images capturing transitions through openings, such as doors. Consequently, reconstructing an entire construction site remains a challenge. Future studies will focus on overcoming these limitations by developing methods to integrate reconstructions of multiple rooms, ultimately facilitating a comprehensive understanding of construction sites and contributing to progress monitoring efforts.

7 Acknowledgements

This research was supported by a grant (RS-2022-00143493) from Digital-Based Building Construction and Safety Supervision Technology Research Program funded by Ministry of Land, Infrastructure and Transport of Korean Government.

References

[1] A. Pal, J. J. Lin, S. H. Hsieh, and M. Golparvar-Fard, “Activity-level construction progress monitoring through semantic segmentation of 3D-informed orthographic images,” *Autom Constr*, vol. 157, Jan. 2024, doi: 10.1016/j.autcon.2023.105157.

[2] T. Wang and V. J. L. Gan, “Enhancing 3D reconstruction of textureless indoor scenes with IndoReal multi-view stereo (MVS),” *Autom*

Constr, vol. 166, Oct. 2024, doi: 10.1016/j.autcon.2024.105600.

[3] Y. Jeon, A. S. Kulinan, D. Q. Tran, M. Park, and S. Park, "NeRF-Con : Neural Radiance Fields for Automated Construction Progress Monitoring," Jun. 2024. doi: 10.22260/ISARC2024/0149.

[4] Y. Xu and J. Zhang, "UAV-based bridge geometric shape measurement using automatic bridge component detection and distributed multi-view reconstruction," *Autom Constr*, vol. 140, Aug. 2022, doi: 10.1016/j.autcon.2022.104376.

[5] Y. Ding, X. Zheng, Y. Zhou, H. Xiong, and J. Gong, "Low-cost and efficient indoor 3D reconstruction through annotated hierarchical Structure-from-Motion," *Remote Sens (Basel)*, vol. 11, no. 1, Jan. 2019, doi: 10.3390/rs11010058.

[6] N. Jiang, Z. Cui, and P. Tan, "A Global Linear Method for Camera Pose Registration," in *2013 IEEE International Conference on Computer Vision*, IEEE, Dec. 2013, pp. 481–488. doi: 10.1109/ICCV.2013.66.

[7] J. L. Schonberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 4104–4113. doi: 10.1109/CVPR.2016.445.

[8] J. Wang, N. Karaev, C. Rupprecht, and D. Novotny, "Visual Geometry Grounded Deep Structure From Motion," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.04563

[9] A. Keyvanfar, A. Shafaghat, and M. S. F. Rosley, "Performance comparison analysis of 3D reconstruction modeling software in construction site visualization and mapping," *International Journal of Architectural Computing*, vol. 20, no. 2, pp. 453–475, Jun. 2022, doi: 10.1177/14780771211066876.

[10] H. Mahami, F. Nasirzadeh, A. H. Ahmadabadian, and S. Nahavandi, "Automated progress controlling and monitoring using daily site images and building information modelling," *Buildings*, vol. 9, no. 3, 2019, doi: 10.3390/buildings9030070.

[11] K. Han, J. Degol, M. Golparvar-Fard, and A. M. Asce, "Geometry-and Appearance-Based Reasoning of Construction Progress Monitoring," 2017, doi: 10.1061/(ASCE).

[12] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[13] C. Griwodz *et al.*, "AliceVision Meshroom," Association for Computing Machinery (ACM), Jun. 2021, pp. 241–247. doi: 10.1145/3458305.3478443.

[14] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "OpenMVG: Open Multiple View Geometry," vol. 10214, B. Kerautret, M. Colom, and P. Monasse, Eds., in Lecture Notes in Computer Science, vol. 10214. , Cham: Springer International Publishing, 2017, pp. 60–74. doi: 10.1007/978-3-319-56414-2_5.