PLAUSIBLE RECONSTRUCTION OF AN APPROXIMATED MESH MODEL FOR NEXT-BEST VIEW PLANNING OF SFM-MVS

R. Moritani [1]*, S. Kanai [2], H. Date [2], Y. Niina [3], R. Honma [3]

[1] Graduate School of Information Science and Technology, Hokkaido University, Japan - r_moritani@sdm.ssi.ist.hokudai.ac.jp
[2] Faculty of Information Science and Technology, Hokkaido University, Japan - (kanai, hdate)@ssi.ist.hokudai.ac.jp
[3] Asia Air Survey Co., Ltd. - (ysh.niina, ryh.honma)@ajiko.co.jp

Commission VI, WG VI/4

KEY WORDS: Surface reconstruction, Quality prediction, Structure from Motion, Multi-View Stereo, View planning, Next-best-view

ABSTRACT:

Structure-from-Motion (SfM) and Multi-View Stereo (MVS) are widely used methods in three dimensional (3D) model reconstruction for an infrastructure maintenance purpose. However, if a set of images is not captured from well-placed positions, the final dense model can contain low-quality regions. Since MVS requires a much longer processing time than SfM as larger amounts of images are provided, it is impossible for surveyors to wait for the SfM–MVS process to complete and evaluate the geometric quality of a final dense model on-site. This challenge results in response inefficiency and the deterioration of dense models in 3D model reconstruction. If the quality of the final dense model can be predicted immediately after SfM, it will be possible to revalidate the images much earlier and to obtain the dense model with better quality than the existing SfM–MVS process. Therefore, we propose a method for reconstructing a more plausible 3D mesh model that accurately approximates the geometry of the final dense model only from sparse point clouds generated from SfM. This approximated mesh model can be generated using Delaunay triangulation for the sparse point clouds and triangle as well as tetrahedron filtering. The approximated model can be used to predict the geometric quality of the final dense model and for an optimization-based view planning. Some experimental results showed that our method is effective in predicting the quality of the final dense model and finding the potentially degraded regions. Moreover, it was confirmed that the average reconstruction errors of the dense model generated by the optimization-based view planning went below tens of millimeters and falls within an acceptable range for an infrastructure maintenance purpose.

1. INTRODUCTION

The increasing cost for inspecting and repairing aging infrastructures has been recently recognized to be a serious social problem, especially in many developed countries (ASCE, 2017). Therefore, to reduce the cost, three dimensional (3D) "as-is" models that can capture the existing status of the infrastructures including their damages and deformations are expected to be effective techniques to efficiently archive and utilize maintenance-related information because the difference between the past and present states of the structures can be easily assessed.

Currently, Structure-from-Motion (SfM) and Multi-View Stereo (MVS) are the most economical and easy-to-use methods for reconstructing the 3D as-is models. These methods can automatically generate a 3D dense model of structures, with rich textures of damaged and degraded regions, from many overlapped images using only a camera. However, if the images are not captured from well-placed positions, the final dense model can contain geometrically low-quality regions (e.g., distortion and holes). Since MVS requires much longer processing time than SfM as larger amounts of images are provided, it is impossible for surveyors to wait for the SfM–MVS process to complete and evaluate the geometric quality of the final dense model on-site. This challenge results in response inefficiency and the deterioration of dense models in 3D model reconstruction. To solve the challenge, it is preferable to estimate the quality of the final dense model and validate the photo-capturing positions (camera poses) at an earlier stage than MVS.

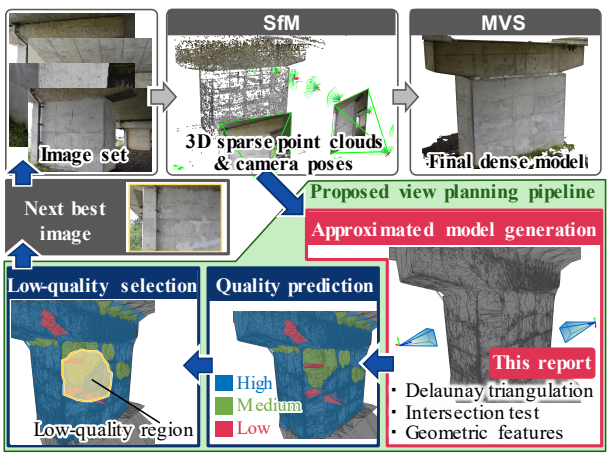


Figure 1. Outline of our view planning system for SfM–MVS (Moritani, 2019).

This is to improve the efficiency and quality of the dense model generation.

So far, we have proposed a view planning system for SfM–MVS (Moritani, 2019) where the geometric quality of the final dense model of MVS can be predicted using only the SfM results, as shown in Fig. 1. The proposed method first generates an approximated triangular mesh model of the final dense model

* Corresponding author

from the SfM results only, i.e., sparse point clouds and camera poses. In our system, the quality of the final dense model is estimated using the quality predictors, and the low-quality regions on the approximated model are automatically selected where more images should be supplemented by making use of mathematical programming using the quality predictor values.

However, in our previous work (Moritani, 2019), the geometric accuracy of the approximated mesh model was still not enough to plausibly predict the quality of the final dense model. Consequently, appropriate low-quality regions could not always be identified because a considerable number of inappropriate triangles and tetrahedra that do not fit with the dense model surface still remain in the mesh model.

This paper aims to solve this problem. We introduced triangular and tetrahedral mesh filtering to remove the inappropriate triangles and tetrahedra from the approximated mesh model based on the geometric features, so that the approximated mesh model could better fit with the final dense model reconstructed using the MVS. Consequently, we can predict more accurate low-quality regions on the approximated triangular mesh model than those obtained using our previous method (Moritani, 2019).

2. RELATED WORK

Several view planning methods targeted at SfM–MVS process have been studied so far, to derive the best camera poses as well as a navigation strategy that assure completeness and high quality of the reconstructed model with the aid of a computer system.

In these planning methods, some of them have utilized a priori knowledge of the target object to estimate the best camera poses. As a priori knowledge, some "approximated" models of the target object have been used. For example, 2-dimensional map of the building (Jing, 2016), and small-sized 3D models generated from an MVS process (Hepp, 2017), (Schmid, 2012), (Roverts, 2017) performed at a set of images captured from nonoptimized and simple camera poses. However, the approximated models generated only by 2D maps do not accurately represent the 3D surface situation of the target object. Moreover, it is inefficient to perform the time-consuming MVS process only for reconstructing a priori knowledge of the view planning.

Some studies have addressed the approximated model generation only from SfM process for the view planning. The (Qi, 2009) method generates an approximated triangular mesh model from sparse point clouds such that a target object is rotated in front of a fixed camera. However, this approximated model generation method is not applicable to large-scale outdoor scenes with plausible accuracy.

The (Labatut, 2007) method reconstructs a low-resolution triangular mesh model by minimizing the energy function defined on the sparse point clouds generated from the SfM process. Similar to the (Labatut, 2007), approximated triangular mesh models of indoor and outdoor environments are reconstructed from noisy point clouds generated from SfM using planar approximation and graph-cuts (Holzman, 2017). Unfortunately, in both methods, computationally expensive optimization problems must be solved to generate the approximated mesh model; hence, they cannot be adopted as an approximated model generation method for our view planning system.

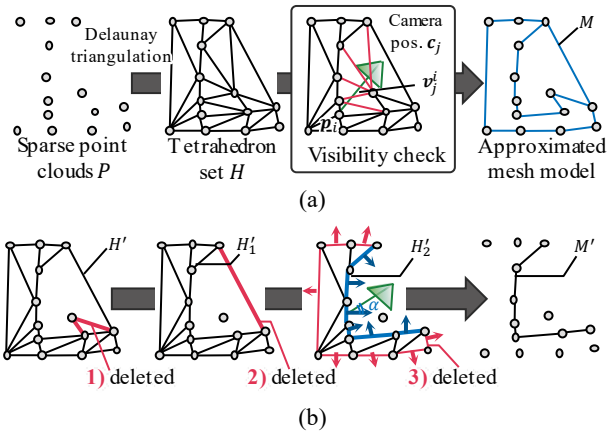


Figure 2. Approximated triangular mesh generation processes. (a) Our previous method (Moritani, 2019). (b) Proposed mesh filtering.

3. QUALITY PREDICTON METHOD

3.1 Improvement of approximated mesh model generation

As shown in Fig.1, the proposed view planning system for SfM–MVS (Moritani, 2019) first generates an approximated triangular mesh model to grasp the spatial occupancy around a target object by checking visibility between the camera position and sparse point clouds generated by SfM. The approximated model is later used to predict the quality measures of the 3D geometry of the final dense model. Our approximated model generation method simplifies the model formerly proposed by (Labatut, 2007) to improve the computational efficiency.

As shown in Fig. 2(a), our approximated model generation method begins with Delaunay triangulation of sparse point clouds P and generates a set of tetrahedral H, followed by the intersection test between every tetrahedron and a set of rays $V_i = \{\boldsymbol{v}_j^i\}$ $(\boldsymbol{v}_j^i = \boldsymbol{p}_i - \boldsymbol{c}_j)$ starting from the j-th camera position $\boldsymbol{c}_j$ to the i-th visible sparse point position $\boldsymbol{p}_i$. If a tetrahedron intersects with the ray, it is deleted, and the remaining set of tetrahedra is defined as H'. Finally, we obtained an approximated mesh model M that is a set of surface boundary meshes of H' and roughly represents the target object surface.

However, in our previous method (Moritani, 2019), the following inappropriate triangles as shown in Fig. 3 that do not necessarily approximate the final dense model remain on M: 1) locally spiky triangles, 2) widespread triangles, and 3) triangles in the region occluded from all camera poses. These triangles are not completely deleted by the intersection test with rays. Due to these improper triangles, the quality of low-quality regions may often be overestimated and that of high-quality regions may be underestimated. In addition, high-quality regions may, sometimes, be wrongly selected as low-quality regions where additional images should be taken.

To solve the above-mentioned problems, we developed the following three filtering processes against these meshes 1)–3) that appeared on the tetrahedra H' shown in Fig. 2(b).

1) Filtering locally spiky triangles

First, the number of adjacent tetrahedra of a tetrahedron $h(\in H')$ is counted. If it is less than or equal to 1 and the *stretch* (Geuzaine, 2009) of h defined by Equation (1) is less than or equal to 0.1, then h is deleted from H'. The stretch is defined as a ratio of 0 to

1, and the smaller stretch means that the tetrahedron exhibits a shaper shape and higher distortion.

$$Stretch(h) = 6\sqrt{6}W(h)/\left(S(h)\max_{e\in h}(l_e)\right). \quad (1)$$

where $W(h)$ denotes the volume of h, $S(h)$ the surface area of h, and l_e the length of the edge e within h.

We repeated this for all tetrahedra in H' to obtain the new set of tetrahedra H_1'.

2) Filtering widespread triangles

The widespread triangles on H_1' covering the entire outer side of H_1' do not fit into the field of view of a camera; hence, they may be misattributed as low-quality regions based on the quality predictors using the edge length. To delete these triangles, we first examine whether an edge of all the triangles on H_1' can be seen from at least more than two cameras. If it is not possible, the edge is deleted as part of the widespread triangles. The above filtering process is applied to all triangular meshes on the boundary of H_1', and we generate a new set of tetrahedra H_2'.

3) Filtering triangles in the occluded region

The triangles in the region that are not visible from all camera positions must be deleted because they may exist on the backside of the object to be modeled and they do not contribute any quality prediction. Therefore, a triangle on H_2' is first labeled as an *inside* triangle if it is adjoined to two tetrahedra. If it is adjoined to only one tetrahedron, it is labeled as a *surface* triangle. Further, we calculated the normal vector of the *surface* triangles and evaluated the incident angle α between the normal and the vector from the camera position to the centroid of the triangle. If α exceeds the threshold τ_α, the triangle is deleted. The process was applied to all triangles on the boundary of H_2', and the remaining surface triangles finally constitute the approximated triangular mesh model M'.

3.2 Quality predictors

The quality predictor of the final dense model was calculated at every sparse point $i(\in P)$ on the approximated triangular mesh model M'. The predictor $F_x(i)$ quantifies how accurately the final dense model can be reconstructed around a sparse point i. Before estimating the predictors, we normalized the scale of the approximated mesh model to make an average distance $\bar{R}$ equal to 1 among the nearest neighboring points on the mesh model.

The following four quality predictors $F_x(i)$ (Equations (2–5)) are evaluated at and assigned to a sparse point i:

a) *Reliability* $\left(F_r(i)\right)$

The reliability of the 3D reconstruction of the dense model of a local region around i generally decreases as the number of visible cameras $|V_i|$ supporting a sparse point i decreases. Therefore, the reliability $F_r(i)$ of the local region around i is evaluated by Equation (2):

$$F_r(i) = |V_i| \quad (2)$$

b) *Area* $\left(F_a(i)\right)$

The area of a triangle on the approximated mesh model is larger, the reconstruction error of the final dense model generated by MVS tends to be larger. To evaluate the area, the average area of the triangles on M' at a sparse point i is evaluated as $F_a(i)$ for all triangles t_j^i (T^i) including i by Equation (3):

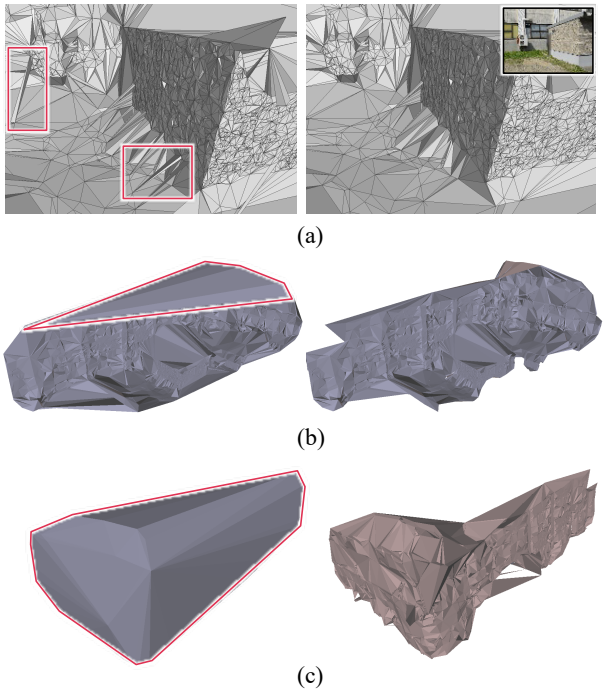


Figure 3. The examples of the proposed filtering effect. The left figures include inappropriate triangles (red frame) before filtering, the right figures after filtering: (a) locally spiky triangles, (b) widespread triangles and (c) triangles in the occluded region

$$F_a(i) = \frac{1}{|T^i|}\Sigma_{t_j^i\in T^i}\, area\left(t_j^i\right) \quad (3)$$

where T^i denotes a set of triangles on M' connected to i.

c) *Edge length* $\left(F_e(i)\right)$

The edge length of the triangles on the approximated triangular mesh model M' tends to be long and the point clouds generated by SfM become sparse, when surface of the target object is poorly textured. Therefore, a longer edge on M' indicates a clue to the low-quality regions on the final dense model. To this end, the average edge length at a sparse point i on the approximated triangular mesh model is evaluated by Equation (4) as $F_e(i)$:

$$F_e(i) = \frac{1}{|D^i|}\Sigma_{j\in D^i} length\left(e_j^i\right) \quad (4)$$

where D^i denotes a set of edges connected to i on the approximated mesh model.

d) *Baseline and height ratio* $\left(F_{bh}(i)\right)$

Higher-quality reconstruction results by MVS are usually obtained from more correct ratio between baseline length and height. As shown in Fig. 4, the base line length is defined as a distance between a pair of the camera positions $\boldsymbol{c}_j$ and $\boldsymbol{c}_k$ visible from a sparse point i, while the base line height is a distance between the position $\boldsymbol{p}_i$ of i and the midpoint $\boldsymbol{c}'_{jk}$ of baseline of visible camera pair(j,k). It is known in photogrammetry that the quality of the final dense model decreases as the ratio is biased (Yan, 2016). Thus, the ratio between baseline length and height $F_{bh}(i)$ defined by Equation (5) was adopted as one of the quality predictors:

$$F_{bh}(i) = \frac{1}{|J_i|}\Sigma_{(j,k)\in J_i}\left(\frac{\|\boldsymbol{c}_j-\boldsymbol{c}_k\|}{\|\boldsymbol{p}_i-\boldsymbol{c}'_{jk}\|}\right) \quad (5)$$

where, J_i denotes a set of all possible camera pair visible from a sparse point i on M'.

To consolidate the four quality predictors to one indicator representing the geometry degradation of the final dense model, first, we converted each of the quality predictors given by Equations (2–5) to normalized energy ∈ [0,1] using the logistic function based on (Mauro, 2014) or quadratic function as Equation (6):

$$E_X(i) = \begin{cases} L(F_X - \mu_X, \sigma_X), & X \in \{a, e\}; \\ 1 - L(F_X - \mu_X, \sigma_X), & X \in \{r\}; \\ 1 - K(F_X, \sigma_X), & X \in \{bh\}, \end{cases} \quad (6)$$

where μ_X denotes the average of F_X, σ_X the standard deviation of F_X, $L(x-\mu,\sigma) = 1/\left(1+\exp\left(-\frac{2(x-\mu)}{\sigma}\right)\right)$ the logistic function for normalization, and $K(x,\sigma) = 1/(1+(x-0.5/\sigma)^2)$ the quadratic function for normalization. In Equation (6), higher energy means that the geometry of the final dense model degrades more.

Next, the energy values E_X are aggregated by taking an average to denote a *geometry degradation indicator* (GDI) at a sparse point i as $E_{GDI}(i)$ using Equation (7) below:

$$E_{GDI}(i) = \left(E_r(i) + E_a(i) + E_e(i) + E_{bh}(i)\right)/4 \quad (7)$$

Therefore, a region with high indicator value $E_{GDI}(i)$ on the approximated triangular mesh model M' indicates that the local region around the sparse point i on the final dense model has much larger possibility of degrading the geometry. It also implies that the effective photos are lacking for the region with high indicator value $E_{GDI}(i)$ and additional photos should be adequately supplemented to improve the reconstruction quality of the region around the sparse point i.

4. QUALITATIVE VALIDATION

We applied the proposed method to 3D reconstruction of a building wall model shown in Fig. 5(c) and qualitatively compared the appropriateness of the approximated mesh models generated from SfM process between the proposed method and our previous method (Moritani, 2019) based on the visual observations. The SfM–MVS process was performed using commercial software (Bentley), and 23 images were initially input to the process.

Fig. 5(a) and 5(b) respectively show the approximated triangular mesh model M generated by our previous method (Moritani, 2019) and M' by the proposed method with a threshold of $\tau_\alpha = 90°$. From the GDI distribution shown in Fig. 5(a)(b) on the two approximated mesh models, three lowest-quality regions where more images should be added were selected as indicated "No.1, 2, and 3" in Figs. 3(a) and (b). Fig. 5(c) shows the final dense model with some low-quality regions (holes) generated by the SfM–MVS process from the original 23 input images.

The results show that the geometric quality of M' itself is better than that of M. It is observed that some invalid large triangles appear on M, as shown in Fig. 5(a). Due to these triangles, the previous method incorrectly extracted the regions where the images are not lacking actually. On the other hand, as shown in Fig. 5(b), our proposed method implementing the tetrahedron and triangle filtering deleted invalid triangular meshes that remain on

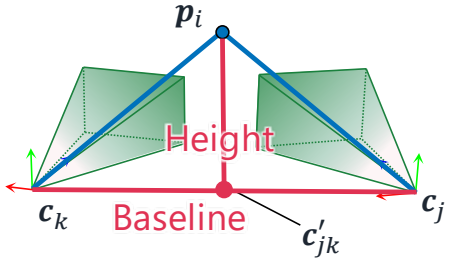


Figure 4. Baseline and height ratio.

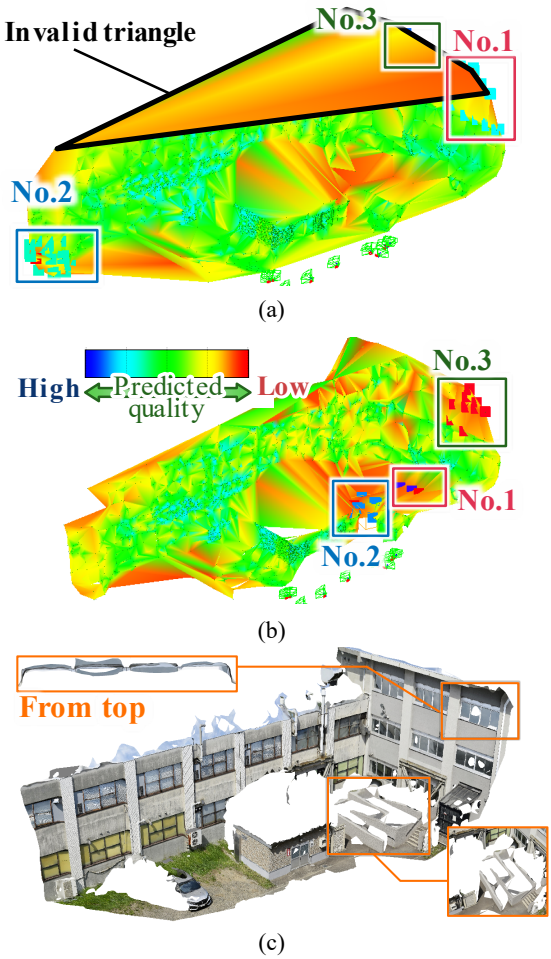


Figure 5. Approximated mesh models and the predicted qualities and final dense model of a building wall. (a)Approximated mesh model M by our previous method (Moritani, 2019), and (b) M' by the proposed method. The color maps show the distributions of the estimated geometry degradation indicator values, where red indicates low quality and blue indicates high quality. "No.1, 2, and 3" show three lowest-quality regions. (c) Final dense model with some low-quality regions (holes) generated by MVS.

M. It is also observed that the GDI distribution on the approximated mesh model of Fig.5(b) better resembles the distribution of the low-quality regions (holes) that appeared on the final dense model Fig. 5(c) than the mesh model of Fig.5(a) does. From this result, it is suggested that the proposed method correctly selected the low-quality regions on M' that correspond to the actual low-quality regions of the final dense model.

Finally, we supplemented 27 images to be targeted to improve the second lowest-quality regions indicated as "No.2" in Fig. 5(b) and generated the final dense model shown in Fig. 6. By comparing Fig. 5(c) and Fig.6, the reconstructed area on the final dense model is sufficiently expanded by the addition of these images. The processing time needed to generate the approximated triangular mesh model included 1.0 sec for the ray intersection test and 0.3 sec for the mesh filtering processed using an Intel i7-5960X @3.00GHz.

5. OPTIMIZATION-BASED VIEW PLANNING AND QUANTITAVIE VALIDATION

Finally, we implemented an optimization-based view planning where the additional camera poses to be supplemented around the target object are derived from a large collection of original images I_0 and their camera poses C_0 by using optimizations. This view planning proceeds as the following steps:

(1) An approximated triangular mesh model M' was generated from the initial image set I_1 and their camera poses C_1 containing a few dozen of images, and the GDI $E_{GDI}(i)$ is evaluated at each sparse point i on M'.

(2) The N_{lq} regions of low quality are automatically selected from M' based on the distribution of E_{GDI} on M', and N_{lq} target points $\{k\} \in K, (|K| = N_{lq})$ each of which is placed around at the center of a low-quality region are selected among the set of sparse points of M'. The target point selection is performed by solving a combinatorial optimization using the greedy method proposed in (Moritani, 2019).

(3) In each low-quality region around a target point $k(\in K)$, the best camera pose $\tilde{\boldsymbol{c}}(k)$ that could most effectively improve the geometric quality of the region around k is selected among the remaining set of images $I_0 - I_1$ by solving the following minimization problem in Equation (8).

$$\tilde{\boldsymbol{c}}(k) = \min_{i \in C_{visible(k)}} \{NBV_{i,k}^{posture} \times (w_f NBV_{i,k}^{positionFront} + w_{bh} NBV_{i,k}^{positionBH})\} \quad (8)$$

where $C_{visible}(k)$ $(\subset C_0 - C_1)$ denotes a set of the camera poses from which the target point k is visible excludingC_1. $NBV_{i,k}^{posture}$, $NBV_{i,k}^{positionFront}$,and $NBV_{i,k}^{positionBH}$ indicates how the selected camera pose i accurately fits with the theoretical best pose to capture the target point k, and are defined by Equations (9–11);

$$NBV_{i,k}^{posture} = 1 - (\boldsymbol{v}_i^{opticalaxis} \cdot (\boldsymbol{p}_k - \boldsymbol{c}_i)/\|\boldsymbol{p}_k - \boldsymbol{c}_i\|), \quad (9)$$

$$NBV_{i,k}^{positionFront} = 1 - (\boldsymbol{n}_k \cdot (\boldsymbol{c}_i - \boldsymbol{p}_k)/\|\boldsymbol{c}_i - \boldsymbol{p}_k\|), \quad (10)$$

$$NBV_{i,k}^{positionBH} = |\|\boldsymbol{c}_r - \boldsymbol{c}_i\|/\|\boldsymbol{p}_k - \boldsymbol{c}'_{ri}\| - 0.536|. \quad (11)$$

Here, $NBV_{i,k}^{posture}$ indicates how precisely the optical axis $\boldsymbol{v}_i^{opticalaxis}$ of the selected camera i is oriented toward the target point k. $NBV_{i,k}^{positionFront}$ indicates how precisely the position of the selected camera i aligns the surface normal direction $\boldsymbol{n}_k$ of M'at the target point k, in other words, how precisely the camera i covers the low-quality regions centered at k. $NBV_{i,k}^{positionBH}$ indicates how precisely the selected camera i is relatively positioned with the other existing cameras in terms of the baseline and height ratio, and how effectively it could improve the quality of the dense model. In these equations, $\boldsymbol{c}_i$ denotes the $i-$th camera position, $\boldsymbol{v}_i^{opticalaxis}$ the optical axis of a camera

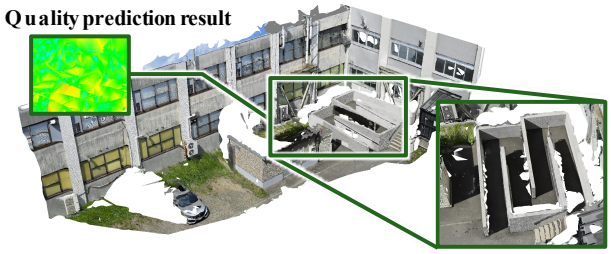


Figure 6. Final dense model and the geometry degradation indicator distribution (a color map on the top-left) on the approximated mesh model after 27 images are added.

Figure 7. The bridge column used for the experiment

i, $\boldsymbol{p}_k$ the target point position and $\boldsymbol{n}_k$ the normal vector of M'at the target point k. $\boldsymbol{c}_r$ denotes the reference camera position that can capture the largest area in the $k-$th low-quality region among $C_{visible}(k)$, $\boldsymbol{c}'_{ri}$ the midpoint between $\boldsymbol{c}_r$ and $\boldsymbol{c}_i$, and 0.536 is the constant of the baseline and height ratio when the angle between $\overline{\boldsymbol{p}_k\boldsymbol{c}_i}$ and $\overline{\boldsymbol{p}_k\boldsymbol{c}'_{ri}}$ takes 30°.

(4) If N_{lq} best camera poses $\{\tilde{\boldsymbol{c}}(k) | \, k \in [1, N_{lq}]\}$ corresponding to the N_{lq} regions of low qualities are selected from $C_0 - C_1$ as the solutions of the minimization problem of Equation (8), these best poses C_a and the images I_a are added to the initial first camera poses C_1 and image set I_1 to obtain the optimized camera poses $C_2 = C_1 \cup C_a$ and $I_2 = I_1 \cup I_a$.

(5) After updating $C_1 \leftarrow C_2$ and $I_1 \leftarrow I_2$, new C_1 and I_1 are inputted again to the SfM process to get an improved version of the approximated triangular model of the final dense model, and we repeated the steps from (1) to (4) until the iterations reaches the specified number or the number of the optimized images in I_2 exceeds the predefined threshold.

(6) Finally, the final dense mesh model was generated from the latest optimized image set I_1 via the MVS process.

To confirm the effectiveness of the optimization-based view planning, we applied it to reconstruct the 3D dense mesh model of a bridge column. We first prepared a collection of the 110 original images I_0 densely taken around a bridge column shown in Fig. 7. Among the 110 original images, we manually selected 21 initial images as I_1 and generated 3D sparse point clouds $P$and the camera poses of the images C_1 using the SfM process implemented in the commercial software (Bentley). Next, the approximated triangular mesh model $M'$was generated from these 21 initial images I_1 using our proposed method. Furthermore, the proposed optimization-based view planning was conducted under the setting of $N_{lq} = 6$ to derive 6 low-quality regions on M' and 6 best camera poses corresponding to them. By repeating these steps 7 times, finally, 63 optimized

camera poses C_7 for the bridge column were selected from the original image poses C_0.

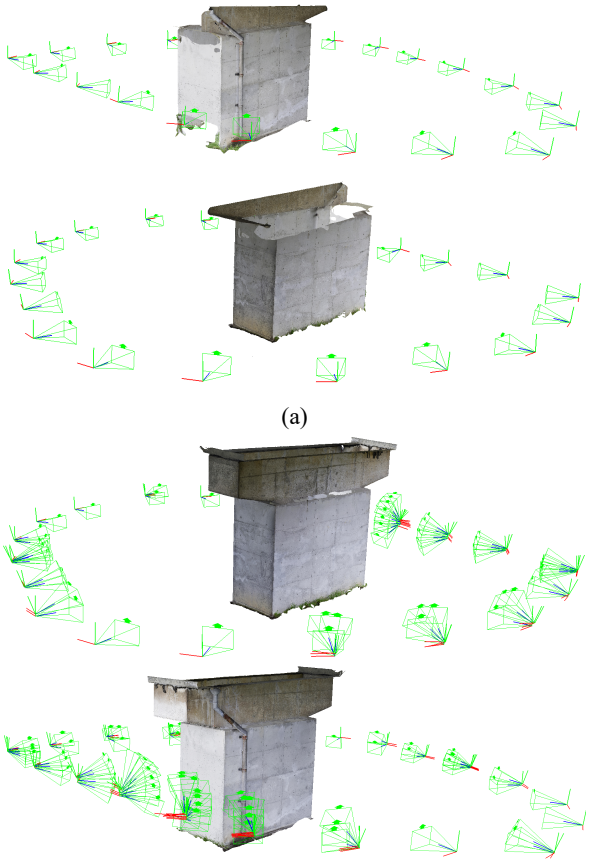
The final dense model generated from the initial camera poses $C_1, (|C_1| = 21)$ and that from the optimized camera poses $C_7, (|C_7| = 63)$ are compared in Fig. 8. It is clearly observed that major parts of the upper and lower portions of the column are missing in the final dense model reconstructed from C_1. On the contrary, a whole geometry of the column was fully reconstructed from the optimized camera poses C_7. Based on the observation, our proposed method of approximated triangular mesh model generation and the optimization-based view planning works as expected.

Finally, we evaluated a reconstruction error of the final dense model generated from the optimized camera poses by comparing it with the reference point clouds captured using a terrestrial laser scanner. Fig. 9 shows the error distributions. It can be observed that in most of the regions on the column surface, the error went below fifty millimeters. The absolute average error of absolute distance between point clouds to mesh model was 3.30 mm and the standard deviation was 0.07 mm. Therefore, the dense model reconstructed using our proposed method has the sufficient degree of accuracy when using it as "as-is" models for infrastructure maintenance purpose.

6. CONCLUSIONS

We proposed an improved method to reconstruct a plausible mesh model that approximates the final dense model geometry only from sparse point clouds obtained from the SfM results based on the triangle and tetrahedron filtering. We also proposed the quality predictors that estimate the degree of the geometric qualities of the final dense model defined only on the approximated triangular mesh model and camera poses. Moreover, an optimization-based view planning method that can rationally select a set of best camera poses based on the quality predictors. The proposed method was applied to two case studies, and we confirmed that the approximated mesh model generated by our method can effectively predict the quality of the final dense model and can identify the regions where more images should be added to improve the geometric quality of the regions. The reconstruction error of a dense model reconstructed using our proposed method was also evaluated by comparing it with the laser-scanned point clouds, and it was confirmed that their average error amount stayed tens of millimeters and falls within an acceptable range when used for infrastructure maintenance purpose.

In future works, we will attempt to complete our view planning system where the next photo-capturing positions are automatically derived from the quality predictors defined on the approximated mesh model.



(a)

(b)

Figure 8. The reconstructed final dense models and the camera poses: (a) initial camera poses $C_1, (|C_1| = 21)$, (b) optimized camera poses $C_7, (|C_7| = 63)$.

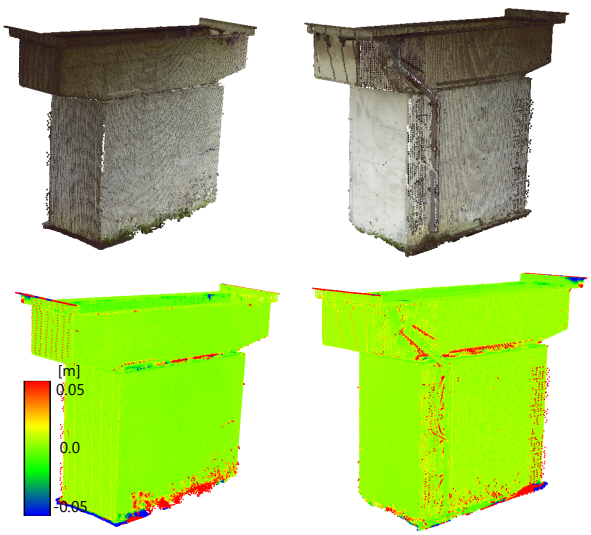


Figure 9. Top row shows the point clouds of the bridge column captured by a terrestrial laser scanner and bottom row depicts the error distributions of the final dense mesh model on the laser-scanned point clouds.

REFERENCES

ASCE (American Society of Civil Engineers), 2017: infrastructure report card, https://www.infrastructurereportcard.org/wp-content/uploads/2019/02/Full-2017-Report-Card-FINAL.pdf (Retrieved 3 May, 2020)

Bentley, ContextCapture. https://www.bentley.com/en/products/brands/contextcapture, (Retrieved 31 January, 2020).

Geuzaine, C., Remacle, J., 2009: Gmsh: A 3-D Finite Element Mesh Generator With Built-in Pre- And Post-processing Facilities. *Int. J. Numer. Meth. Engng.*, *79*(11), 1309–1331. https://doi.org/10.1002/nme.2579

Hepp, B., Nießner, M.S., Hilliges, O., 2018: Plan3D: Viewpoint and Trajectory Optimization for Aerial Multi-View Stereo Reconstruction. *ACM Trans. Graph.*, *38*(1), 1–17. https://doi.org/10.1145/3233794

Holzmann, T., Oswald, M.R., Pollefeys, M., Fraundorfer, F., Bischof, H., 2017: Plane-based Surface Regularization for Urban 3D Reconstruction, in *Br. Mach. Vis. Conf. (BMVC)*.

Jing, W., Polden, J., Tao, P. Y., Lin, W., Shimada, K., 2016: View planning for 3D shape reconstruction of buildings with unmanned aerial vehicles, in *2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 1–6.

Labatut, P., Pons, J.P., Keriven, R., 2007: Efficient Multi-View Reconstruction of Large-Scale Scenes using Interest Points, Delaunay Triangulation and Graph Cuts. *2007 IEEE 11th Int. Conf. Comput. Vis.*, 1–8.

Mauro, M., Riemenschneider, H., Signoroni, A., Leonardi, R., Van Gool, L.J., 2014: A Unified Framework for Content-Aware View Selection and Planning Through View Importance, in *Br. Mach. Vis. Conf. (BMVC)*. 1–11.

Moritani, R., Kanai, S., Date, H., Niina, Y., Honma, R., 2019: Quality Prediction of Dense Points Generated by Structure from Motion for High-quality and Efficient As-is Model Reconstruction. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, *XLII-2/W13*, 95–101.

Qi, P., Reitmayr, G., Drummond, T., 2009: ProFORMA: Probabilistic Feature-based On-line Rapid Model Acquisition. *Br. Mach. Vis. Conf. (BMVC)*, 1–11.

Roberts, M., Shah, S., Dey, D., Truong, A., Sinha, S., Kapoor, A., Hanrahan, P., Joshi, N., 2017: Submodular Trajectory Optimization for Aerial 3D Scanning, in *Proceedings of the IEEE International Conference on Computer Vision*, 5334–5343. doi: 10.1109/ICCV.2017.569.

Schmid, K., Hirschmüller, H., Dömel, A., Grixa, I., Suppa, M., Hirzinger, G., 2012: View Planning for Multi-View Stereo 3D Reconstruction Using an Autonomous Multicopter. *J.l Intell. Robot. Syst.*, *65*(1), 309–323. doi: 10.1007/s10846-011-9576-2.

Yan, L., Fei, L., Chen, C., Ye, Z., Zhu, R., 2016: A Multi-View Dense Image Matching Method for High-Resolution Aerial Imagery Based on a Graph Network. *Remote Sens*. 8, 799. https://doi.org/10.3390/rs8100799

Revised January 2020