Supplementary of Batch-based Model Registration for Fast 3D Sherd Reconstruction

Jiepeng Wang[1] Congyi Zhang[1] Peng Wang[1] Xin Li[3] Peter J. Cobb[1]
Christian Theobalt[2] Wenping Wang[3*]

[1]The University of Hong Kong [2]Max Planck Institute for Informatics
[3]Texas A&M University

1. Image Acquisition Scheme

Our image acquisition scheme consists of a customized hardware system (see Fig. 1 (a) in the main paper) and a batch capturing strategy. It is designed to meet the following requirements for high-throughput sherd capturing and reconstruction: (1) efficient acquisition; (2) sufficient coverage for accurate reconstruction; (3) minimal manual labor; (4) portable and easy to deploy in the field.

The hardware of the system has three main parts:

- A turntable consisting of a stepper motor, a stepper motor driver, an Arduino board [1], and a flat board in printed ArUco patterns [7] for camera calibration;
- Three cameras mounted on an aluminum frame. They are at different heights to provide sufficient coverage of the vertical viewing range.
- A controller module running on a PC that controls the motion of the turntable and synchronizes the motion with the cameras shutters.

Acquisition procedure. To capture a *group* of fragments in a batch mode, we place them flat on the turntable, and first take a set of pictures to capture their exposed sides, to be called the *front sides*. Then the fragments are flipped manually on the turntable to photograph their *back sides*. We call all these pictures a *batch*. Three cameras are used to take the pictures and the turntable makes 16 stops to complete a full circle of rotation, i.e. with a rotation angle of 22.5° for each move. Therefore, each batch has $3 \times 16 = 48$ images.

The capturing of each batch of 48 images is controlled and synchronized by a PC controller. Once captured, the batch of images are transmitted to a PC for 3D reconstruction.

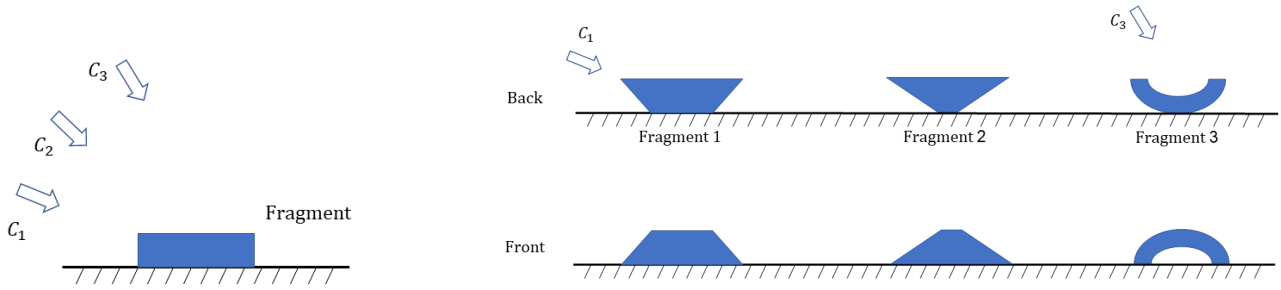
*Corresponding author.

Design justification. In arriving at this setup of devices, we have tested different configurations with different numbers of cameras and images to take in a full circle. Our tests showed that the setup with three cameras provides better coverage in terms of vertical view angles for faithful 3D model reconstruction than those with one or two cameras; and using more cameras would unnecessarily increase the cost and complexity of the capturing device without noticeable improvement of reconstruction accuracy. Meanwhile, we found that taking 16 images by each camera in a full circle provides better coverage of the side view for accurate coverage than using substantially fewer images, while increasing the number of images to more than 16 will unnecessarily increase the time of image processing and without bringing noticeable accuracy improvement. Details of these experiments can be found in Sec. 2.

Scale consistency. Image-based reconstruction often has an issue of scale ambiguity. Without a reference metric, 3D structures reconstructed from images taken in different passes may differ by a global scaling factor. To resolve this ambiguity, we used the ArUco codes on the patterned board as a scale bar, which have known sizes, to normalize the scale of the reconstructed models in world coordinates. Specifically, we first take a batch of photos of fragments. Then, we use the square ArUco tags printed on the board with known width d to recover the camera position C_a in the world coordinates. Then, by comparing the scale between cameras positions from SfM and C_a, we can recover the scale of reconstructed models s.

2. System Parameters

Following the criteria of accuracy, efficiency, and cost in practical sherd acquisition, we tested and optimized two major factors in our design: (1) the number of cameras n and their spatial distribution; (2) the number of divisions k, namely, the number of stops the turntable makes when turning a whole circle (so that a total of $k \cdot n$ images are



(a) Camera angles (b) Fragment shapes

Figure 1: Visualization of cameras and fragments. (a) Relative positions of the 3 cameras and the angles between their view direction and the horizontal plane. (b) a sketch which shows the relationship between 2D fragments with various shapes and cameras' view direction

captured). In the following Sections 2.1 and 2.2, we evaluate the influence of n and k respectively, using a randomly selected group of fragments. The two sides are registered and merged to 9 complete models. The reconstruction results are compared with the 3D scan groundtruth, and the average error of these fragments is evaluated.

2.1. Number of Images to Capture

The choice of division number k is critical since it affects both reconstruction quality and acquisition efficiency. Previous studies showed that between the neighboring views, an angle from 5° to 60° was appropriate for multi-view depth estimation [14]. In order to figure out the influence of k in our system, we tested the value of k from 8 to 24 with a step size of 4. Table 1 depicts the quantitative evaluation results of reconstructed 3D models under different k values.

Table 1: Quantitative evaluation of different numbers of divisions for each circle. No.: Number of images per circle for a camera; Acc.: average 90% Accuracy(mm) (lower is better); Comp.: average Completeness of reconstructed models(%) at the threshold $0.20mm$ (higher is better); MAE: average Mean Absolute Error(mm); SD: average Standard Deviation(mm). TC: Time cost for Capturing images of two sides (min).

No.	Accu.↓	Comp.↑	MAE↓	SD↓	TC↓
8	0.99	36.69	0.46	0.41	1.4
12	0.13	98.46	0.07	0.07	2.2
16	0.12	99.37	0.07	0.07	3.0
20	0.12	99.41	0.07	0.06	3.6
24	0.13	99.40	0.07	0.08	4.4

As shown in Tab. 1, the reconstruction accuracy increases when more images are used. But when the image number is bigger than 12, no clear improvement can be obtained. This may be due to that although more images can provide more information, many images are redundant and do not help further improve the accuracy. Note that even when $k = 12$ (and totally $k \cdot 3 = 36$ images are taken by the 3 cameras), we still get a good reconstruction with accuracy $0.13mm$. This shows the robustness of our system. Besides, using more images helps cover more regions and increases the completeness. When the number of images is larger than 16, incorporating more images doesn't help further improve completeness much.

Except for accuracy and completeness, acquisition efficiency is another critical consideration. As shown in Table 1, the time cost of image capturing is nearly linear to the number of images. So using fewer images indicates better efficiency.

Through extensive experiments, we set $k = 16$, where the neighboring angle is 22.5°. This configuration provides a good balance between accuracy, completeness, and efficiency.

2.2. Number of Cameras to Use

The vertical and horizontal angles of cameras determine (1) whether the fragments on the turntable can be observed clearly and completely, and (2) how robust and accurate the 3D reconstruction can be. As shown in Fig. 1, the shapes of fragments may be different. For fragment type 2, a camera with a small angle with the horizontal plane is helpful to observe the stripe side of the fragments. For fragment 3, camera 3 is necessary to observe the top side of the fragments. Besides, it should be mentioned that if the orientation of camera 1 is too low, there will be large occlusions between fragments in the captured images while if the orientation of camera 3 is too high (i.e., vertical to the turntable), camera 3 will not contribute too much for the final reconstruction because every image is almost same. Therefore, the arrangement of the cameras is very important.

Given these considerations, we set $n = 3$ and integrate

three identical DLSR cameras on a frame to cover the fragments laid on the board from views of different heights and orientations. The orientations of these 3 cameras are uniformly distributed ($\sim 25°$, $\sim 40°$, and $\sim 55°$, respectively) to have a good coverage of different fragments from various angles.

We tuned the orientations of these cameras, and tested the combinations of these three parameters to see the influence of the number of cameras and their angles. From Table 2, we can see that when only using a single camera, the reconstruction quality is relatively low, because the fragment regions cannot be fully covered. When using two cameras, the reconstruction quality improves. And when 3 cameras are used, the system achieves best model accuracy and completeness.

Considering the importance of model quality in sherd documentation, we use 3 cameras to setup our capturing system.

Table 2: Quantitative evaluation of reconstructed fragments with different cameras. Accu.: average 90% Accuracy of 9 fragments (lower is better); Comp.: average Completeness of reconstructed model at the threshold $0.20mm$ of 9 fragments (higher is better); MAE: average Mean Absolute Error of 9 fragments; SD: average Standard Deviation of 9 fragments.

ID	Accu.↓	Comp.↑	MAE↓	SD↓
1	0.31	67.69	0.16	0.11
2	0.45	67.32	0.19	0.17
3	0.28	85.3	0.13	0.11
1,2	0.14	98.13	0.08	0.06
1,3	0.14	96.81	0.08	0.06
2,3	0.13	98.48	0.07	0.06
1,2,3	0.12	99.37	0.07	0.07

3. Implementation Details

Image capturing device In our lab experiments, a computer with Intel(R) i7-7700HQ CPU @ 2.8GHz and 8G RAM is used to control and synchronize 3 cameras and the turntable to capture images. The server that runs our reconstruction algorithms is a machine with two Intel(R) Xeon(R) Silver 4216 Processors, 128G RAM and an GeForce RTX 3090 Graphics card. At the excavation site, we used a mini-computer with a Celeron N5105 CPU @ 2.0GHz, 16G RAM, and 512G SSD to control the camera and turntable. To process the data collected at the excavation site, we instead used a server with a Xeon Gold 5117 CPU @ 2GHz, a Tesla P40 GPU, 192G of RAM, and a 25TB harddrive. A square board (width: 30cm) printed with 400 ArUco tags is used to hold fragments and facilitate the camera pose calibration and scaling. Three Nikon D610 cameras are controlled by the Nikon SDK to take pictures automatically. The cameras are set to have ISO 100, F22, and the shutter speed of 0.5s, or an automatic shutter speed in the field. Note that ISO is a significant parameter for image quality, where a high ISO will cause significant image noise [11], especially in low-light conditions, which will further influence the multi-view reconstruction accuracy. Thus in order to guarantee sufficient light-conditions, we adopt 4 fill lights to provide sufficient lights for using a low ISO setting.

Image segmentation. We used the off-the-shelf U-Net code [12], enhanced with Lovász Loss [2], to segment images of fragments. We setup a green screen behind the turntable (Fig. 1 in the main paper) to facilitate fragment segmentation in captured images. Enhanced with the screen and board, fragment regions in captured images can be effectively segmented by the neural network automatically.

Point cloud clustering The reconstructed model of each side contains multiple disconnected pieces that belong to different fragments. We adopt a *Region Growing* algorithm to separate these fragments. First, all the points are projected to the plane determined by the first and second principle components of PCA [15], which can keep the largest gap between different fragments. Then, on the point cloud, starting from a randomly selected seed point, the points within a distance r to it are considered to be its connected neighbors. This propagates to all the neighbors iteratively until no new neighbor can be found. Each found connected component is a fragment.

Prevention of sliding on the Turntable Since fragments are placed on a turntable, if the turntable rotates with a sudden and jerky motion with big acceleration, the fragments may slide or wiggle on the turntable. This could cause motion blur or even damage the fragments. Therefore, in our design, we adjust the motor's acceleration to ensure a smooth transition to avoid the sliding. At the beginning of each rotation step of the motor, only a small positive angular acceleration is generated to gradually increase the angular speed from zero to the maximum speed. At the end of each rotation step, similarly, the angular speed is decreased gradually with a small negative angular acceleration. Based on our experiments and observations, we find that when the maximal angular speed is controlled under $12\,\mathrm{deg}/s$, and maximal acceleration under about $7\,\mathrm{deg}/s^2$, fragments with different sizes and geometries can stay on the board stably without any sliding. Therefore, we calibrated our motor accordingly to ensure the motion it generates is bounded by these parameters.

4. Capturing Efficiency of Existing Systems

In Table 3 of the main paper, we compared the capturing efficiency of our system and those of existing systems. Since there is no standard benchmark for such a side-by-side comparison, we need to estimate these systems' efficiency. Their efficiency is estimated as follows.

To capture *fresco fragments*, [4] (i.e., [8] in the main paper) developed a system to capture *a single fresco fragment* using structured light. In this scheme, one user performs multiple scans (e.g., 6 by default) on each side of the fragment to capture its geometry acquisition, and a second user takes pictures to get image information. By merging these captures, a throughput of about 10 fragments per hour, or 80 per day, can be achieved. Later, [3] (i.e., [7] in the main paper) used a faster 3D scanner to achieve a better capturing speed of 3 minutes per fragment, namely, a throughput of 20 fragments per hour, or 160 pieces per day. [6] (i.e., [15] in the main paper) designed an automated view planning algorithm to scan *multiple pieces* together. Their experiments showed that their system takes about 44 minutes to scan one side of four fresco fragments, namely, on average 22 minutes to capture both sides of each fragment (i.e, $60/22{*}8h = 22$ fragments per day). All these three acquisition systems are not as efficient as ours. Furthermore, these methods were designed for *flat fresco fragments*, and do not perform very well on curved (e.g., pottery) fragments.

Another category of 3D fragments covers *lithic artifacts*, which are small 3D pieces, and not flat and thin like frescoes. [9] (i.e., [23] in the main paper) tested a photogrammetry-based method to capture two sides of a single artifact by moving around the object and taking about 30 images. This reduces the capturing time to about 10 minutes (6 fragments per hour), or achieves a throughput of $60/10 * 8h = 48$ fragments per day. [9] also tested another option by using a NextEngine Desktop 3D scanner to scan two or more orientations of an object, then merges the individual scans into a complete model. The capturing procedure takes about 90 minutes to finish one piece, reducing to a throughput of only 5 pieces per day, because of lots of manual post-processing operations. [10] (i.e., [27] in the main paper) designed a simple photogrammetry rig with a turntable to improve the efficiency of the phtogrammetry-based method. This system takes about 12 minutes to capture 74 images for an object, reaching a throughput of 5 fragments per hour ($60/12 * 8h = 40$ fragments per day).

For *general pottery fragments*, [8] (i.e., [19] in the main paper) used a structured light based scanner to simultaneously scan multiple fragments held by a support frame (through $6 \sim 10$ scans in a circle), which requires manual postprocessing operations to remove the frame. This system has a throughput of $10 \sim 15$ fragments per hour, or $12.5 * 8h = 100$ fragments per day. However, the scanned models are often incomplete and contain missing regions due to occlusion. We also list the data acquisition efficiency of pottery fragments using our Einscan scanner [5] (i.e., E-GT in the main paper), which takes about $20 \sim 30min$ to scan a complete model (Sec. 5.1 in the main paper), and reaches a throughput of $60/25{*}8h = 19$ fragments per day. It can only process one fragment each time, which limits its efficiency.

Compared with these methods, our method can reach a throughput of about 85 fragments per hour with minimized manual efforts, which is significantly faster than other methods.

5. Evaluation Metrics

Given the reconstructed model R and the ground-truth model G, the metrics used for 3D reconstruction quality evaluation are defined as follows:

- The *accuracy* T_a (mm) is defined with respect to a given percentage p_a. Specifically, T_a is a distance threshold such that p_a percentage of the reconstructed points from R have their distances to G smaller than T_a. Following the recommendation of the Middlebury Benchmark [13], we choose the percentage $p_a = 90\%$ to compute the accuracy T_a. A smaller *accuracy* T_a indicates more accurate reconstruction.

- The *completeness* $(\%)$ is a percentage p_c defined with respect to a given distance threshold T_c. Specifically, p_c is percentage of points in G whose distances to R are smaller then T_c. A larger *completeness* p_c indicates a better overlap between the ground truth G and the reconstructed model R. Following the Middlebury Benchmark [13], we choose the distance threshold of $0.20mm$ for computing the *completeness* p_c. For example, $p_c = 95\%$ means that 95% of the points in G have their distances to R less than $0.20mm$.

- The *MAE* and *SD* (mm) are the mean absolute error and standard deviation, respectively, of the point-wise errors for the points of R to G. The point-wise error of a point in R is its closest distance to G.

References

[1] Yusuf Abdullahi Badamasi. The working principle of an arduino. In *2014 11th international conference on electronics, computer and computation (ICECCO)*, pages 1–4. IEEE, 2014.

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the otimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

[3] Benedict Brown, Lara Laken, Philip Dutré, Luc Van Gool, Szymon Rusinkiewicz, and Tim Weyrich. Tools for virtual reassembly of fresco fragments. *International journal of heritage in the digital era*, 1(2):313–329, 2012.

[4] Benedict J Brown, Corey Toler-Franklin, Diego Nehab, Michael Burns, David Dobkin, Andreas Vlachopoulos, Christos Doumas, Szymon Rusinkiewicz, and Tim Weyrich. A system for high-volume acquisition and matching of fresco fragments: Reassembling theran wall paintings. *ACM transactions on graphics (TOG)*, 27(3):1–9, 2008.

[5] Einscan pro 2x. `https://www.einscan.com/handheld-3d-scanner/einscan-pro-2x/`, 2020.

[6] Xinyi Fan, Linguang Zhang, Benedict Brown, and Szymon Rusinkiewicz. Automated view and path planning for scalable multi-object 3d scanning. *ACM Transactions on Graphics (TOG)*, 35(6):1–13, 2016.

[7] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[8] Avshalom Karasik and Uzy Smilansky. 3d scanning technology as a standard archaeological tool for pottery analysis: practice and theory. *Journal of Archaeological Science*, 35(5):1148–1168, 2008.

[9] Matthew Magnani. Three-dimensional alternatives to lithic illustration. *Advances in Archaeological Practice*, 2(4):285–297, 2014.

[10] Samantha T Porter, Morgan Roussel, and Marie Soressi. A simple photogrammetry rig for the reliable creation of 3d artifact models in the field: lithic examples from the early upper paleolithic sequence of les cottés (france). 2015.

[11] Young-Il Pyo, Rae-Hong Park, and SoonKeun Chang. Noise reduction in high-iso images using 3-d collaborative filtering and structure extraction from residual blocks. *IEEE Transactions on Consumer Electronics*, 57(2):687–695, 2011.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[13] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006.

[14] Shuhan Shen. Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes. *IEEE transactions on image processing*, 22(5):1901–1914, 2013.

[15] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.