Structure from Motion

Shanchen Jiang
College of Information Science and Engineering, Ocean University of China, Qingdao, China
Email: jiangshanchen@stu.ouc.edu.cn

***Abstract*—In this paper, we introduce the pipeline of SfM (Structure from Motion) and compare three 3D reconstruction methods used SfM. SfM will analyze each photograph by detecting interest points and using matching algorithm in order to identify specific features. Subsequently, these features will be matched in other photographs. Having analyzed each difference, SfM will perform a bundle adjustment to determine the 3D position of each feature. These parameters will be calculated in other pipeline. Three 3D reconstruction methods used SfM are bundler, colmap and openMVG. We analyze them from running time, number of points, distance of point cloud, distance of surface reconstruction and human eye. On the whole, Bundler has better results.**

I. Introduction

Fast, and accurate graphic analyze are extremely useful for computer vision. In recent years, more and more researchers concentrate on how to improve the quality of 3D reconstruction. Now, there are two major streams in 3D reconstruction techniques[1]. One approach is to acquire data actively. This is an active technique, since it employs sensors which project a sheet of light or bundle of rays to measure distance from the reflected signals using either Time-of-Flight (ToF) or triangulation. And the other is to obtain data passively. The most representative work is SfM. SfM can finish recovering 3D information using photogrammetry techniques by acquiring sequences of 2D photographs.This concept is analogous to biological vision in which human and other living creatures construct the 3D structure of a scene or moving objects from 2D images obtained from a retina.

In this paper, SfM is the focus of our experiments. SfM is done by detecting feature points in images and assigning the corresponding homologous point. Its implementation is usually divided into several steps shown in Fig.1. Without priori knowledge of the camera, SfM can reconstruct a scene by one image set. We select three methods to finish comparison. They are bundler, colmap and openMVG. They will be introduced detailedly in the next section.

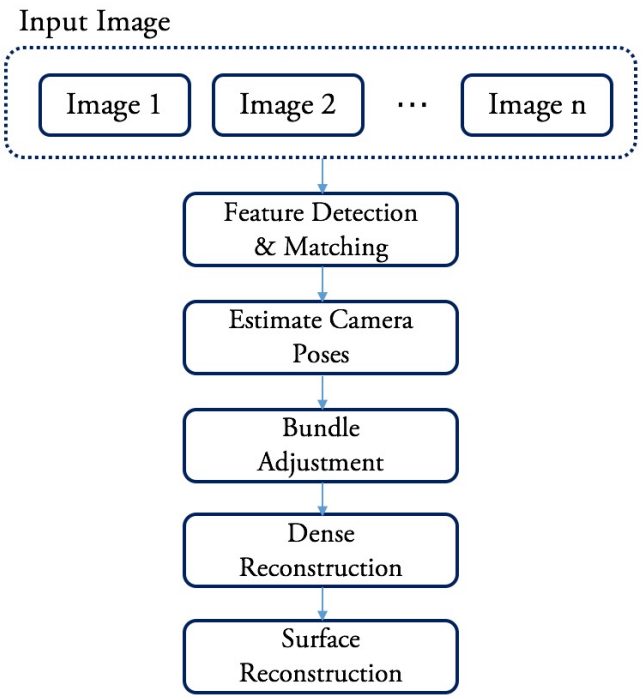


Fig. 1 SfM pipeline

Fig. 2 An earlier version of Bundler was used in the Photo Tourism project

II. Systems of SfM

In this section, three systems used for our comparison will be introduced.

A. Bundler

Bundler is a SfM system for unordered image collections (for instance, images from the Internet) written in C and C++. Bundler takes a set of images, image features, and image matches as input, and produces a 3D reconstruction of camera and scene geometry as output. The home page of Bundler is http://www.cs.cornell.edu/~snavely/bundler/.

B. Colmap

Colmap is an another popular system. It is a general-purpose SfM and Multi-View Stereo (MVS) pipeline with a graphical and command-line interface[2]. It offers a wide range of features for reconstruction of ordered and unordered image collections. The home page of colmap is https://github.com/colmap/colmap.

C. OpenMVG

OpenMVG is a library for computer-vision scientists and especially targeted to the Multiple View Geometry community.

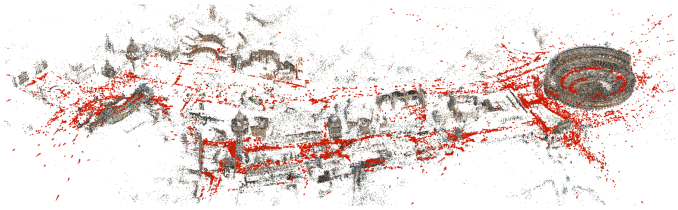


Fig. 3 One project generated by colmap

It is designed to provide an easy access to the classical problem solvers in Multiple View Geometry and solve them accurately. All the features and modules of openMVG are unit tested[3]. Furthermore, it makes it easier for the user to understand and learn the given features.The home page of openMVG is https://github.com/openMVG/openMVG/.

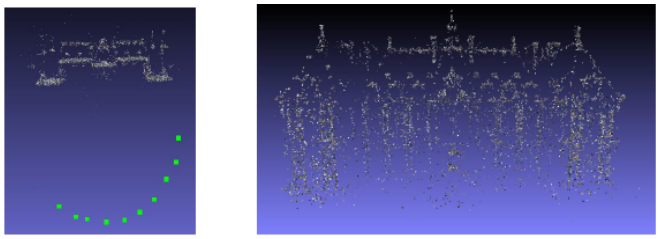


Fig. 4 Estimated camera location and structure generated by openMVG

III. Point Cloud Generation Process

3D point cloud data generation process using the image is shown in Figure 1. In this section, we will introduce them detailedly.

A. Sparse Reconstruction

The generation of sparse point cloud consists of three steps: feature point extraction and matching, camera parameter estimation and bundle adjustment.

1) Feature Point Extraction and Matching: In the process of sparse reconstruction, the extraction and matching of feature points are especially important. It is the basis for calculating the sparse point cloud. The software we use most is SIFT (Scale-invariant feature transform). SIFT is an algorithm for detecting local features. It obtains features by searching for the feature points and their descriptors about scale and orientation, and performs image feature point matching. The processing of extraction and matching by SIFT is shown in Fig.5 and Fig.6.

2) Camera Parameter Estimation: In computer vision applications, it is necessary to establish a geometric model of camera imaging in order to determine the relationship between the three-dimensional geometric position of a point on the surface of a space object and its corresponding point in the image[4]. These geometric model parameters are camera parameters. In most conditions these parameters must be experimented and calculated, the process of solving the parameters called the camera calibration.

Fig. 5 Extraction of feature points by SIFT

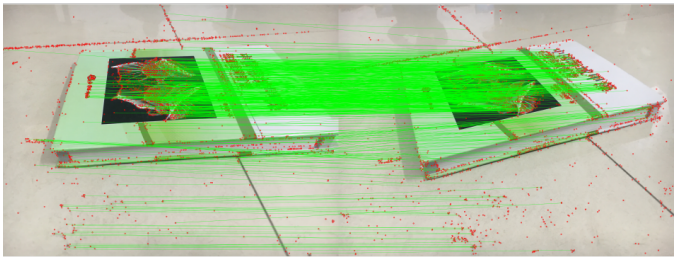


Fig. 6 Matching of feature points by SIFT

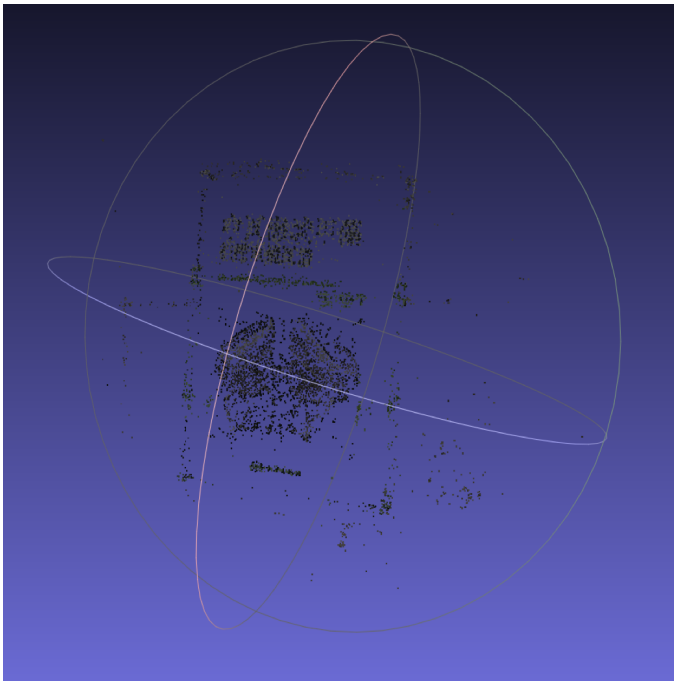


Fig. 7 Result of spares reconstruction

3) Bundle Adjustment: Given a set of measured image feature locations and correspondences, the goal of bundle adjustment is to find 3D point positions and camera param-

eters that minimize the reprojection error. This optimization problem is usually formulated as a non-linear least squares problem. Bundle adjustment is used in almost feature-based 3D scene reconstruction algorithm. From a user provided initial guess the vector of parameters: $[X_j, P_i]_{i,j}$: camera parameters p_{i_i} and the scene structure X_{j_j} are refined in order to minimizes the residual reprojection cost:

$$\underset{[P_i]_i,[X_j]_j}{minimize} \parallel \sum_{j=0}^{m}\sum_{j=0}^{m} x_i^j - P_i X_j \parallel$$

B. Dense Reconstruction

Correspondences between images depicting a static scene cannot only be established for a small set of visually salient regions but can be also extended to the entire image domain. Under the assumption of a static environment, corresponding points in images together with known camera calibration induce a 3D point. Thus, a dense set of correspondences implies a densely sampled surface in 3D. Establishing per-pixel correspondences between only two narrow-baseline images is usually referred as computational stereo. Dense reconstruction addresses the more general problem of obtaining the full 3D geometry from a larger collection of images. Dense reconstruction plays a fundamental role in fully automatic 3D content creation from image data. Accurate dense scene geometry can be augmented with texture images or more general appearance properties enabling photo-realistic rendering of virtual 3D representations.

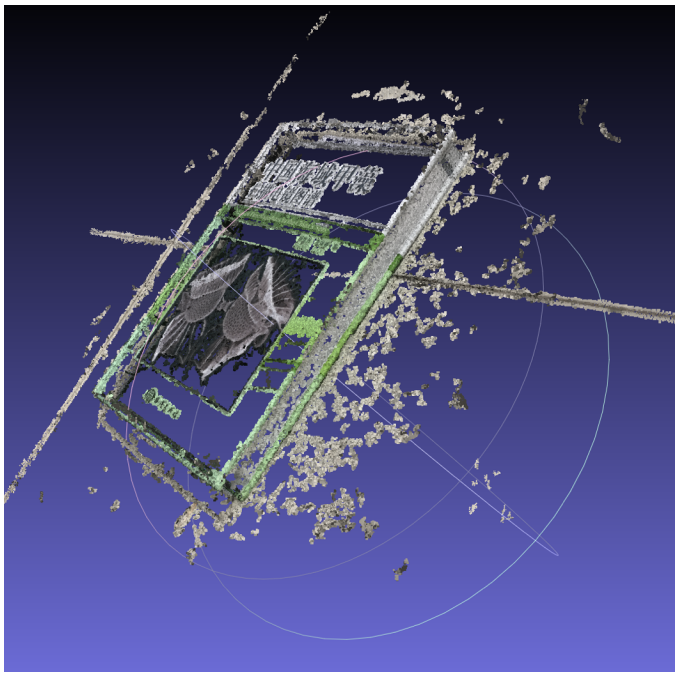


Fig. 8 Result of dense reconstruction

C. Surface Reconstruction

A digital 3D model is a numerical representation of a real object. Two large families of models can be identified: in volumetric models, the local properties all inside the object are represented, while in surface models the surface and the visual appearance of the object are reproduced. Surface reconstruction focus on surface models. Such models can be displayed on a computer or smart-phone display, from any desired view point as a 2D image through perspective projection and rendering. It can also be sent to a processing module for further processing, like for instance in the Computerized Aided Machinery domain, where the model is used to automatically produce from the digital model a physical copy of the real object.

Poisson reconstruction is one of practical surface reconstruction algorithms now. They show that surface reconstruction from oriented points can be cast as a spatial Poisson problem. This Poisson formulation considers all the points at once, without resorting to heuristic spatial partitioning or blending, and is therefore highly resilient to data noise.

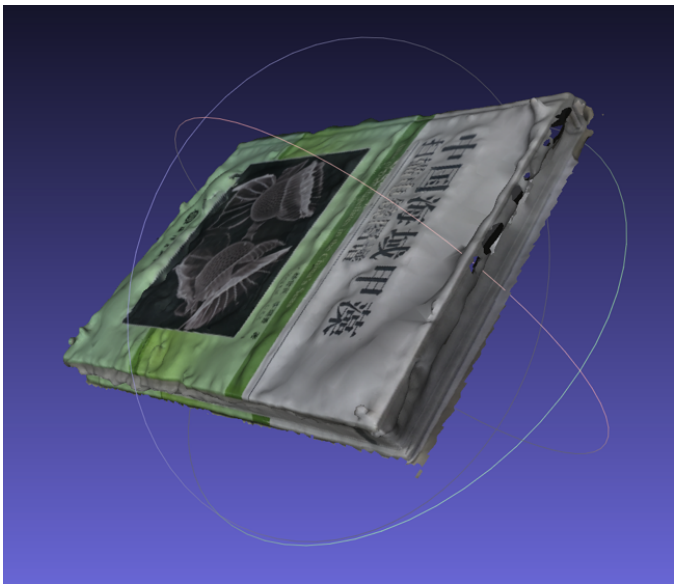


Fig. 9 An example of poisson reconstruction and texture mapping

IV. Dataset

The dataset used in this experiment is MVS Date Set[5]. They are derived from Image Analysis and Computer Graphics at the Technical University of Denmark(DTU). Their experimental setup is shown in Fig.10.

This dataset is aimed at multiple view stereo (MVS) evaluation. The scenes include a wide range of objects in an effort to span the MVS problem. At the same time, the data set also include scenes with very similar objects, e.g. model houses, such that intra class variability can be explored. It consists of 124 different scenes. Each scene has been taken from 49 or 64 position, corresponding to the number of RGB images in each scene or scan. The image resolution is 1600 x 1200. The camera positions and internal camera parameters have been found with high accuracy, via the MATLAB calibration toolbox, which is also the toolbox you need to retrieve these parameters. Lastly, the scenes have been recorded in all 49

or 64 scenes with seven different lighting conditions from directional to diffuse.

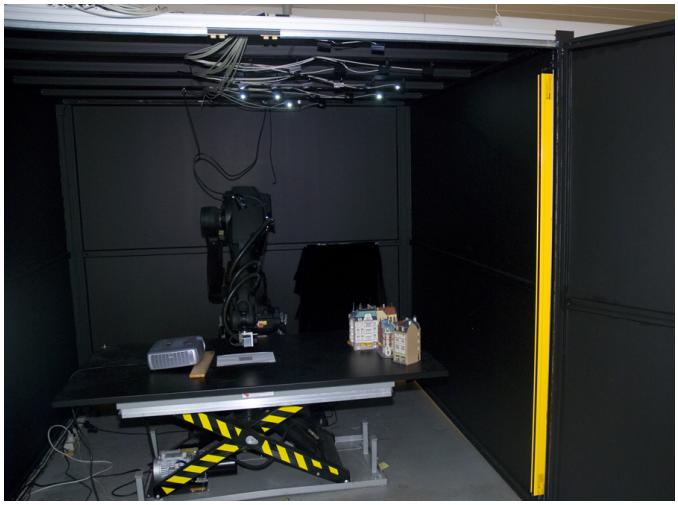


Fig. 10 The experimental setup of MVS Date Set

We select two of them to finish our comparison. T1 and T2 are two packages selected from MVS Date Set. Each of them contains four sections. Each section contains Photo, Point_001 and Surface_001. Photo is used to save image data. From 001 to 004, light intensity increasesy gradually. Point_001 saves the ground truth of spares point clouds. Surface_001 saves the ground truth of surface reconstruction.

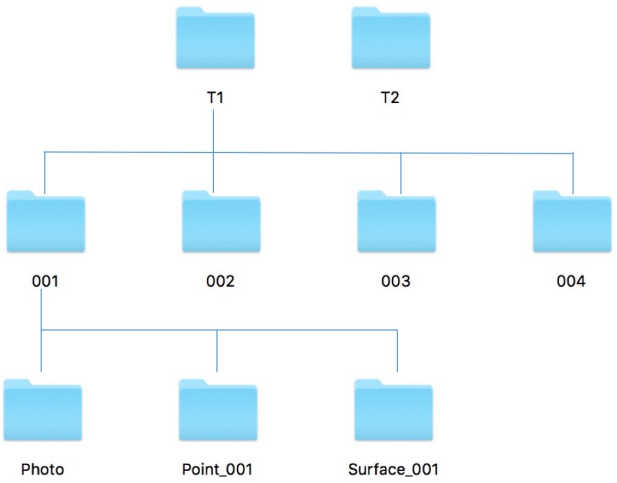


Fig. 11 Our dataset structure

V. Experimental Results

Our experiments focus on the differences between different SfM pipelines. We finish sparse point cloud reconstruction of bundler and colmap, dense point cloud reconstruction of PMVS and poisson surface reconstruction.

Fig.13 shows the ground truth. Our results are shown in Fig.14-16. Fig.14 provides results of sparse point cloud by bundler. Fig.15 provides results of sparse point cloud by colmap. Fig.16 provides results of dense point cloud by PMVS.

First, we compare our results from build time. As the Table I shows, with the light intensity increasing from T1-1 to T1-4, the time of reconstruction increases. The same result applies to T2. Because bundler and colmap are used for sparse point cloud, they don't need too much time to build. At the same time, colmap uses less time than bundler. The generation time of dense point cloud is two to three times of the time of sparse point cloud.

TABLE I Build Time

Number	Bundler	PMVS	Colmap
T1-1	234 s	897 s	208 s
T1-2	260 s	884 s	232 s
T1-3	357 s	921 s	277 s
T1-4	349 s	1055 s	292 s
T2-1	327 s	1001 s	258 s
T2-2	356 s	1021 s	277 s
T2-3	355 s	1134 s	292 s
T2-4	398 s	1459 s	290 s

From the table II, we can easily see that the point number of point clouds is huge. Bundler and colmap are both sparse cloud generation software. Bundler detects more points than colmap, almost two times. This is conducive to the next reconstruction work.

TABLE II Point Number

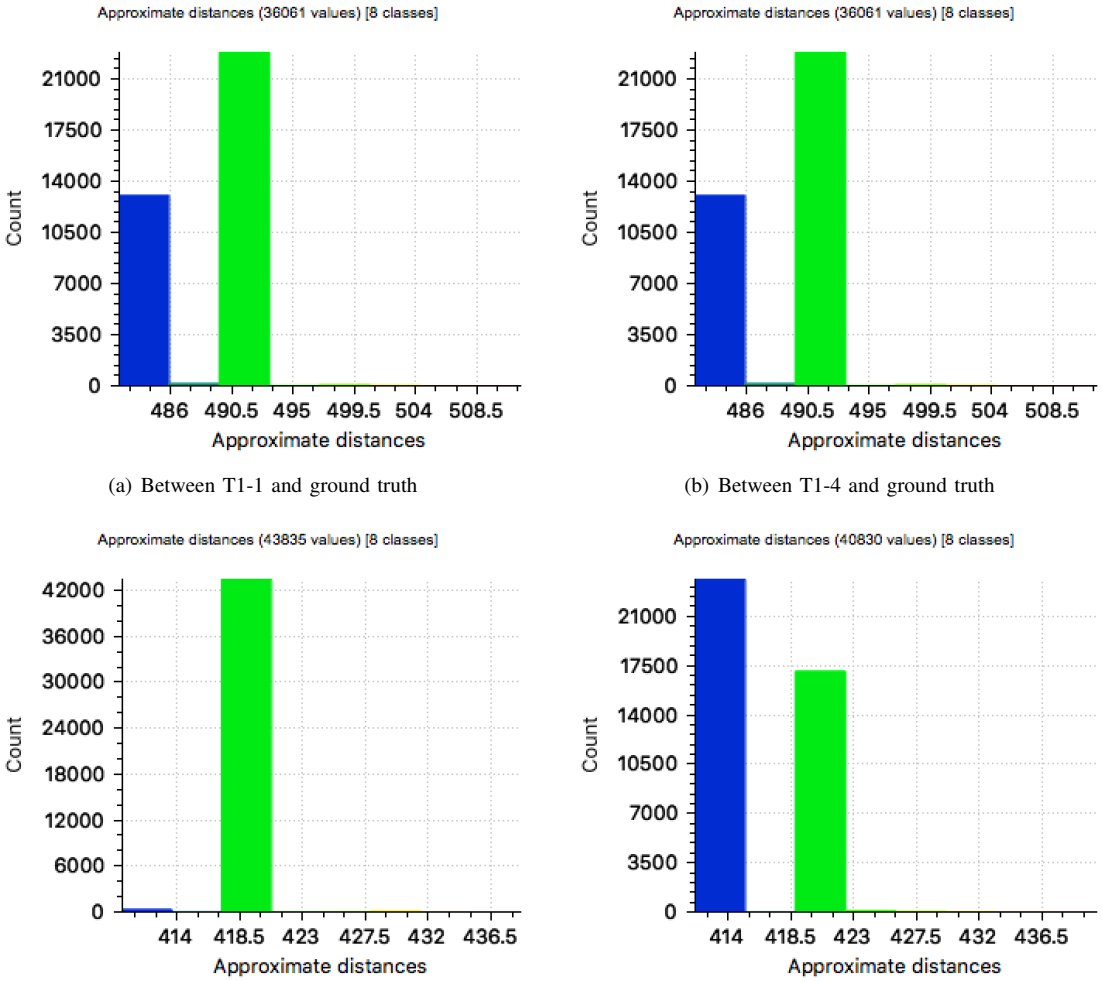
Number	Bundler	PMVS	Colmap
T1-1	36061	141204	16699
T1-2	35645	148242	15354
T1-3	36747	154015	16789
T1-4	41548	158538	18091
T2-1	43835	159780	16317
T2-2	42869	164123	16738
T2-3	38539	164990	16588
T2-4	40830	171673	17870

Table III shows the distance of sparse point clouds between bundler and ground truth. Fig.12 shows the histogram of it. The closer they are, the smaller the gap between him and the truth is. From the table III, T1-4 and T2-4 have higher light intensity, so the average distance is lower, this can also be seen from Fig.12.

TABLE III Distance between different point clouds

Distance between	Average Distance
T1-1 and groundtruth	652.293
T1-4 and groundtruth	488.461
T2-1 and groundtruth	418.08
T2-4 and groundtruth	415.916

Table IV shows the distance of surface between bundler and ground truth. Fig.17 shows the histogram of it. Same to the rule above, the closer they are, the smaller the gap between him and the truth is. Because T1-4 and T2-4 have higher light intensity, so the average distance is lower. Fig.18 shows the results of poisson surface reconstruction. As the number of matching points increases, the more refined the poisson reconstruction is, the smoother the surface is.



(a) Between T1-1 and ground truth

(b) Between T1-4 and ground truth

(c) Between T2-1 and ground truth

(d) Between T2-4 and ground truth

Fig. 12 Sparse Point Cloud Distance

TABLE IV Distance between different surfaces

Distance between	Average Distance
T1-1 and groundtruth	491.555
T1-4 and groundtruth	490.190
T2-1 and groundtruth	426.118
T2-4 and groundtruth	424.063

VI. Conclusion

In this paper, we describe the method based on structure from motion to generate 3D point cloud. We have compared two different methods, bundle and colmap. In existing SfM method, we compare them from build time, point number, point distance, surface distance and human eye. We mainly use distance feature for correspondence matching between our result and ground truth.

In the experiment results, we can find out that the fastest method is colmap, its time has dropped by 10%. Correspondingly, the shorter generation time leads to a lower matching result than bundler. Whether from the number of points clouds, or matching degree between result and ground truth, bundler is better.

Fully-automated 3D-modeling is a challenging issue and a fertile research area. The primary benefit of image-based modeling compared to laser scanners is mainly related to its portability and affordability. SfM provides a excellent thought to finish it. Three challenges need to be solved now. First one is the running time. It is unable to meet the requirements of real-time generation. The second one is poor accuracy. The last one is not enough automation. We need to take part in the whole process, especially in the final revision of the model.

References

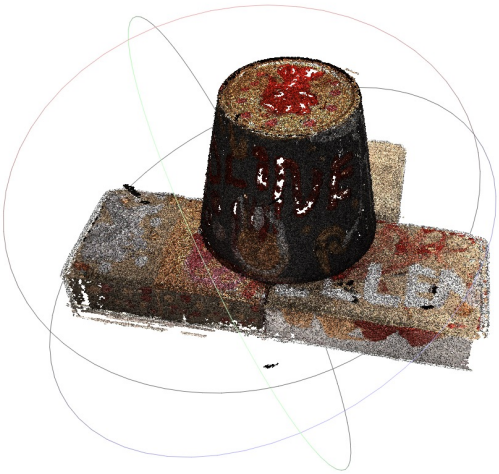
[1] Santoso F, Garratt M A, Pickering M R. 3D Mapping for Visualization of Rigid Structures: A Review and Comparative Study. IEEE Sensors Journal, 2016, 1484-1507.

[2] Schonberger J L, Frahm J M. Structure-from-motion revisited. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, 4104-4113.

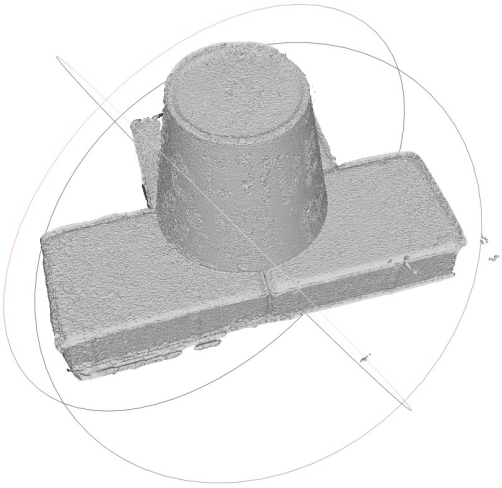
[3] Moulon P, Monasse P, Perrot R. OpenMVG: Open multiple view geometry. International Workshop on Reproducible Research in Pattern Recognition. Springer, Cham, 2016, 60-74.

[4] Pollefeys M, Van Gool L, Vergauwen M. Visual modeling with a hand-held camera. International Journal of Computer Vision, 2004, 207-232.

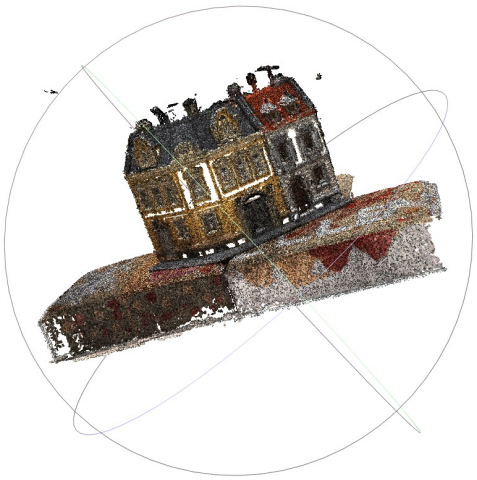
[5] Jensen R, Dahl A, Vogiatzis G. Large scale multi-view stereopsis evaluation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014, 406-413.



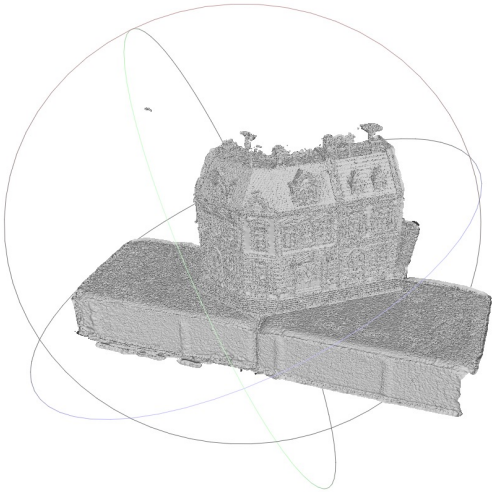
(a) Dense point cloud about T1



(b) Surface Reconstruction about T1

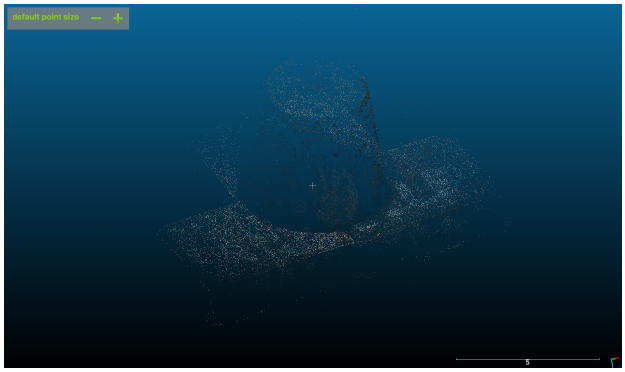


(c) Dense point cloud about T2

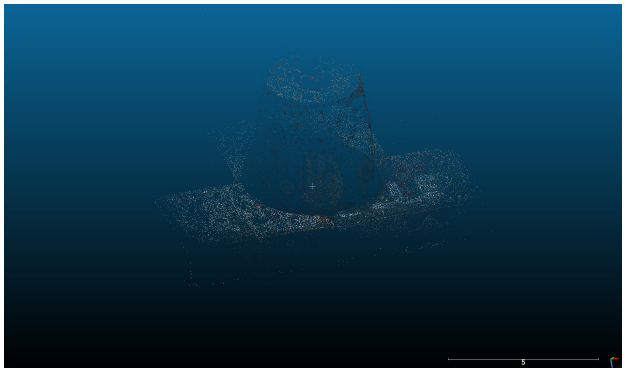


(d) Surface Reconstruction about T2

Fig. 13 Ground truth

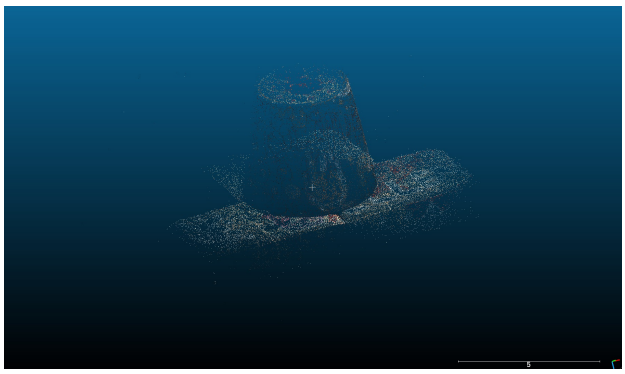


(a) T1-1

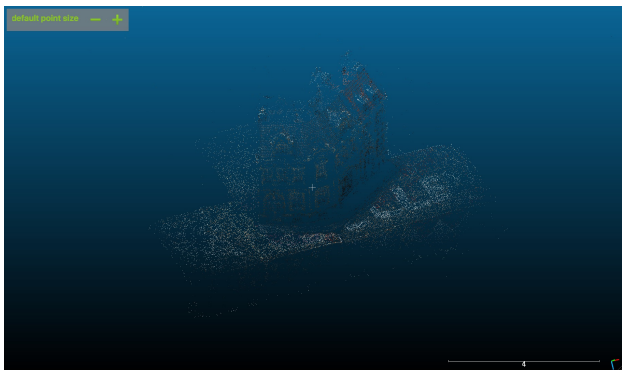


(b) T1-2

(c) T1-3

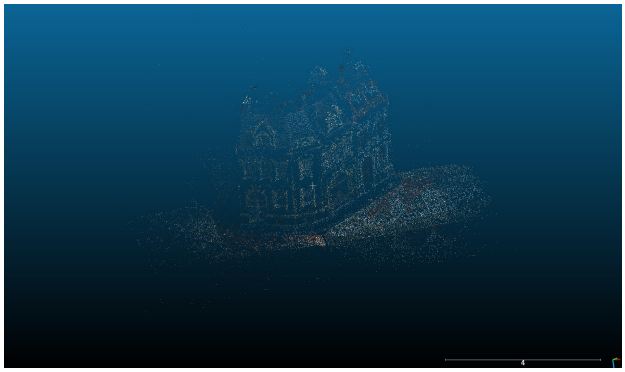


(d) T1-4

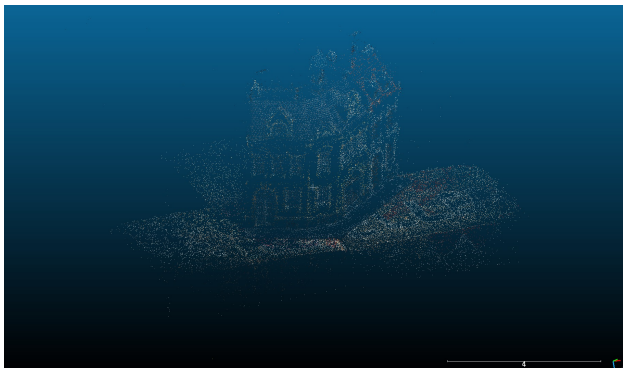


(e) T2-1

(f) T2-2



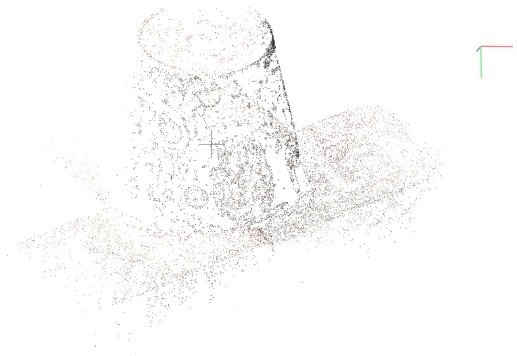
(g) T2-3



(h) T2-4

Fig. 14 Sparse point clouds by Bundler

(a) T1-1



(b) T1-2

(c) T1-3

(d) T1-4

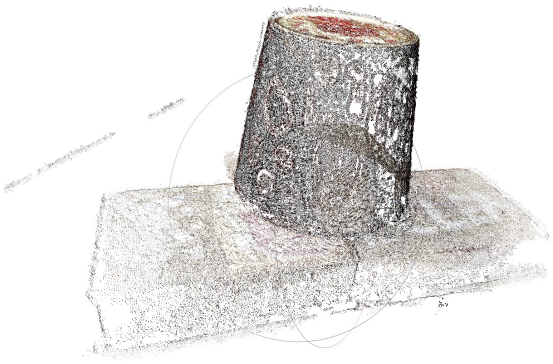
(e) T2-1

(f) T2-2

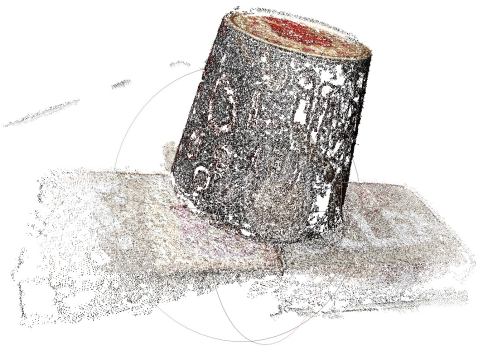
(g) T2-3

(h) T2-4

Fig. 15 Sparse point clouds by Colmap



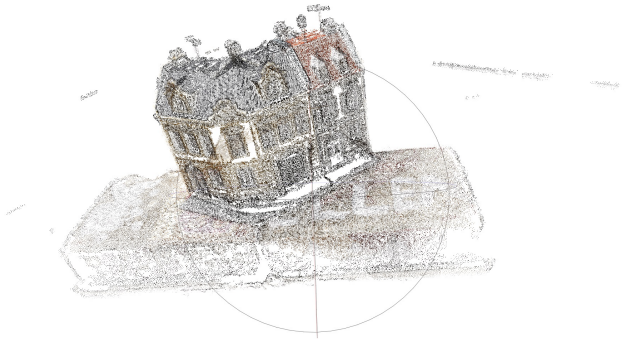
(a) T1-1



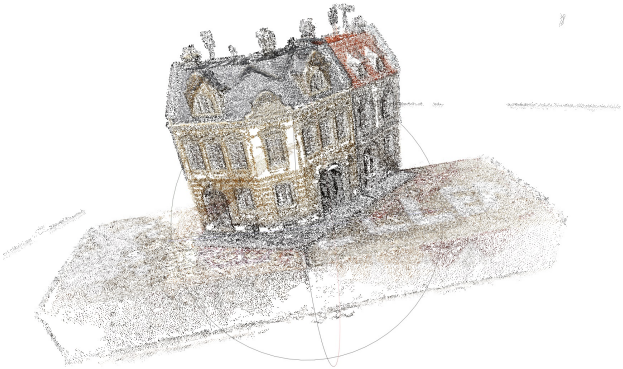
(b) T1-2

(c) T1-3

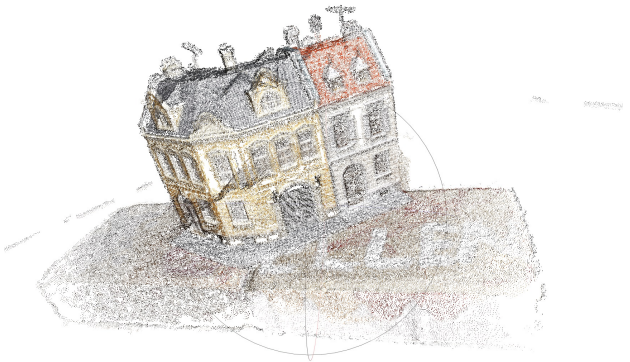
(d) T1-4



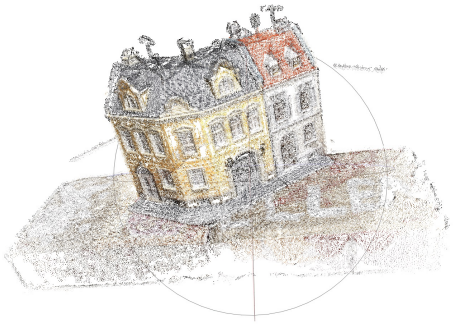
(e) T2-1



(f) T2-2

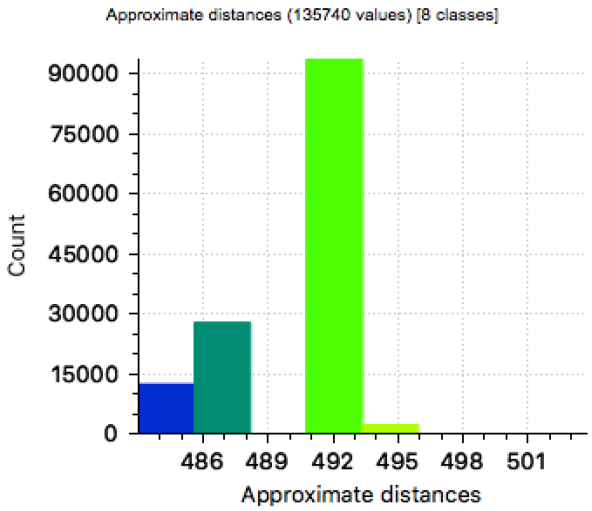


(g) T2-3

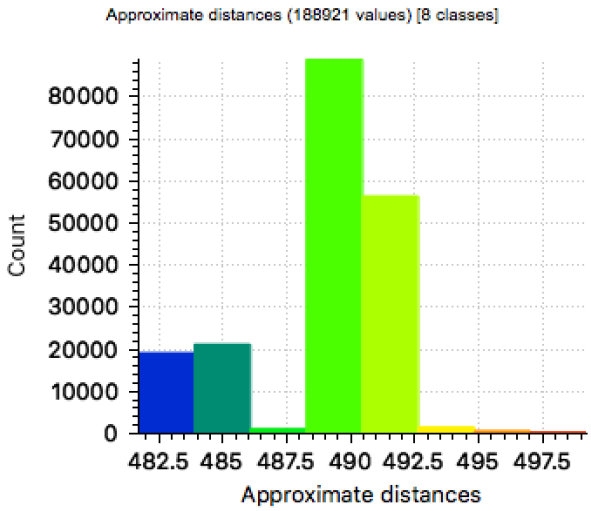


(h) T2-4

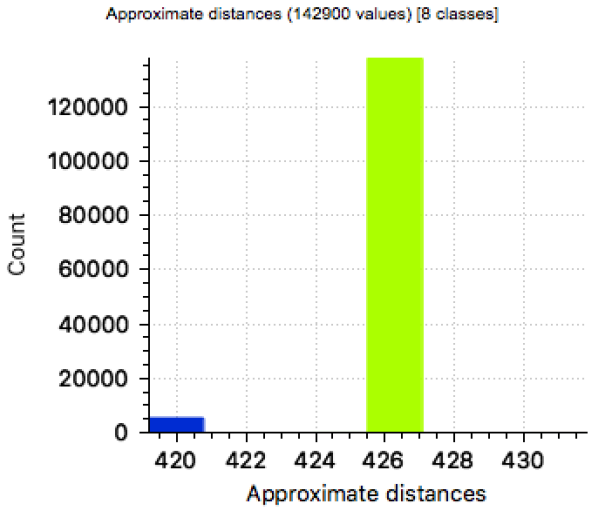
Fig. 16 Dense point clouds by PMVS



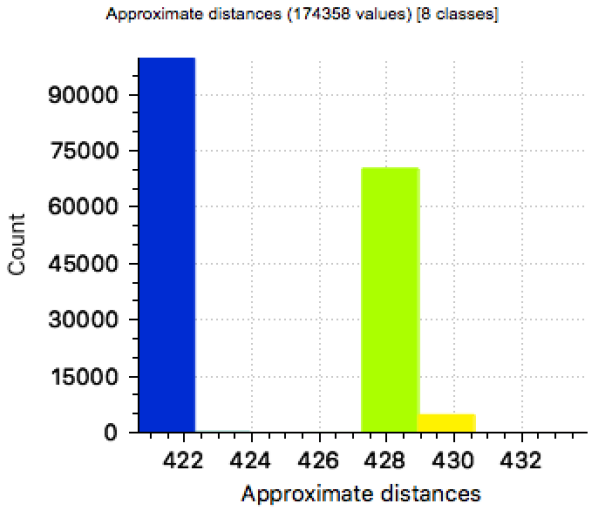
(a) Between T1-1 and ground truth



(b) Between T1-4 and ground truth

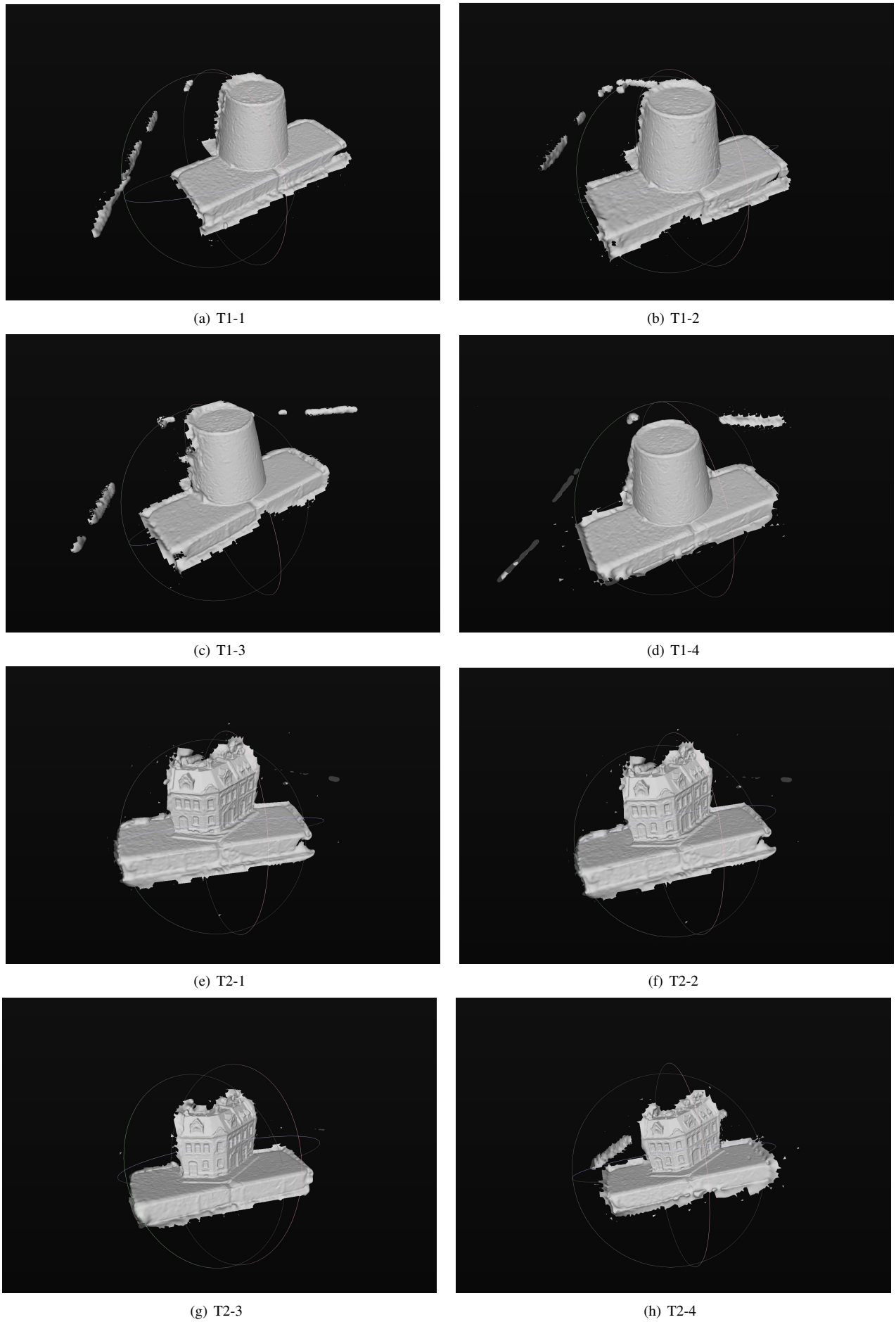


(c) Between T2-1 and ground truth



(d) Between T2-4 and ground truth

Fig. 17 Surface distance



(a) T1-1 (b) T1-2

(c) T1-3 (d) T1-4

(e) T2-1 (f) T2-2

(g) T2-3 (h) T2-4

Fig. 18 Poisson surface reconstruction by result of PMVS