Robust, Precise, and Calibration-Free Shape Acquisition with an Off-the-Shelf Camera and Projector

Chunyu Li, Akihiko Torii, and Masatoshi Okutomi
Tokyo Institute of Technology, Meguro, Tokyo, 152-8550, Japan

***Abstract*—Recently, three-dimensional (3D) shape acquisition systems composed of simple commercial devices have received significant attention from both professional and nonprofessional users. In this work, we propose a flexible projector-camera system that can accurately acquire whole shapes of 3D objects without requiring any special calibrations. The proposed system is based on structured-light and Structure-from-Motion (SfM) techniques that use coded patterns as dense discriminative features instead of the local features popularly used in SfM. To this end, recordings of coded patterns are inserted by moving the camera (to capture image data from multiple camera view points) and projection of the patterns by moving the projector. The proposed system can accurately recover even texture-less objects owing to the advantages of both SfM and structured-light. We demonstrate the benefits of this system by several experiments using real data and compare it to current methods.**

***Index Terms*—Structured-Light System, Entire Shape Acquisition, 3D Scanning, Structure from Motion**

I. Introduction

Acquiring 3D shapes of real-world objects has attracted significant attention from both professional and nonprofessional users because it can derive various applications. For example, archaeologists attempt to capture the 3D geometry of cultural relics to make a replica or undertake further analysis of the 3D model [1], [2]. There are potential users who want to create a replica of real-life objects, *e.g.* memorabilia, using a 3D printer. In contrast to such broad attentions, commercial software or services based on 3D shape acquisition are still in the middle of an expansion to general users. High-quality 3D scanning requires expensive special equipment (*e.g.* laser scanners[3]), professional skills, and efforts to use them.

A popular approach developed in the computer vision community is Structure-from-Motion (SfM) [4] and Multi-View Stereo (MVS) [5], [6], [7], [8] that uses multiple images/photos taken from different viewpoints. The SfM-MVS approach can provide low-cost systems because it essentially requires cameras (and computers) only and can achieve flexible scanning of objects and scenes (*e.g.* with drones.) However, the SfM-MVS approach provides 3D reconstruction in poor quality for recovering textureless objects as they need to match distinctive features/patches across images [9] to estimate the depth and shape.

Another approach is a structured-light system [10] that consists of a camera and a projector. Although the structured-light system requires an additional device (the projector), it can measure the depth (3D scan) in high quality even for textureless objects due to active projection. However, a standard structured-light system has some disadvantages in the view of users: the camera and the projector have to be rigidly fixed (mounted on a rig) and calibrated; each 3D scan has to be merged by registering overlapping parts across the scans (typically by the Iterative Closest Point (ICP) [11]) to generate a single global 3D model, which is not a trivial task.

In this paper, we propose a 3D reconstruction method that combines the SfM and the structured-light to quickly and accurately scan the whole shape of 3D objects. The proposed 3D reconstruction can be achieved with commercial off-the-shelf devices, *i.e.* a standard camera (smartphone) and a digital projector, without requiring a rig to fix them or their calibration. Using a camera and a projector, we first take coded patterns from multiple viewpoints by moving the camera and then change the projector position while placing at least one of the camera views to be shared before and after the projector movement. The proposed method computes the global shapes of 3D objects and camera poses (camera parameters) with SfM that uses coded patterns as dense discriminative features instead of local features [9] popularly used. The experiments demonstrate that this 3D scanning system can provide the whole shape of objects in detail with high resolution.

II. Related work

3D reconstruction from imagery is an actively studied topic in computer vision and image processing. One of the most successful methods is SfM [4] followed by (MVS) [5], [6], [7], [8]. SfM computes the camera poses and a sparse 3D reconstruction, and then MVS constructs the dense 3D reconstruction using the camera poses (camera parameters) obtained from SfM. SfM-MVS essentially recovers 3D points by finding discriminative (local) features and by matching them across the images, *e.g.* SIFT [9], DAISY [12], SSD/NCC. The success of reconstruction depends on the amount texture on the scenes and objects to extract a sufficient number of features. It therefore fails to recover such objects with a lack of texture or repetitive patterns because it is hard to find feature matches that are consistent with the true 3D structures.

An alternative approach often used in computer vision is a structured-light system which encodes positional information of projector pixels in the projected patterns. The structured-light systems have been investigated and summarized in great detail in [13], [10], [14], [15]. The techniques can be categorized into two types, spatial and temporal encoding techniques.

978-1-5386-3025-9/18/$31.00 ©2018 IEEE

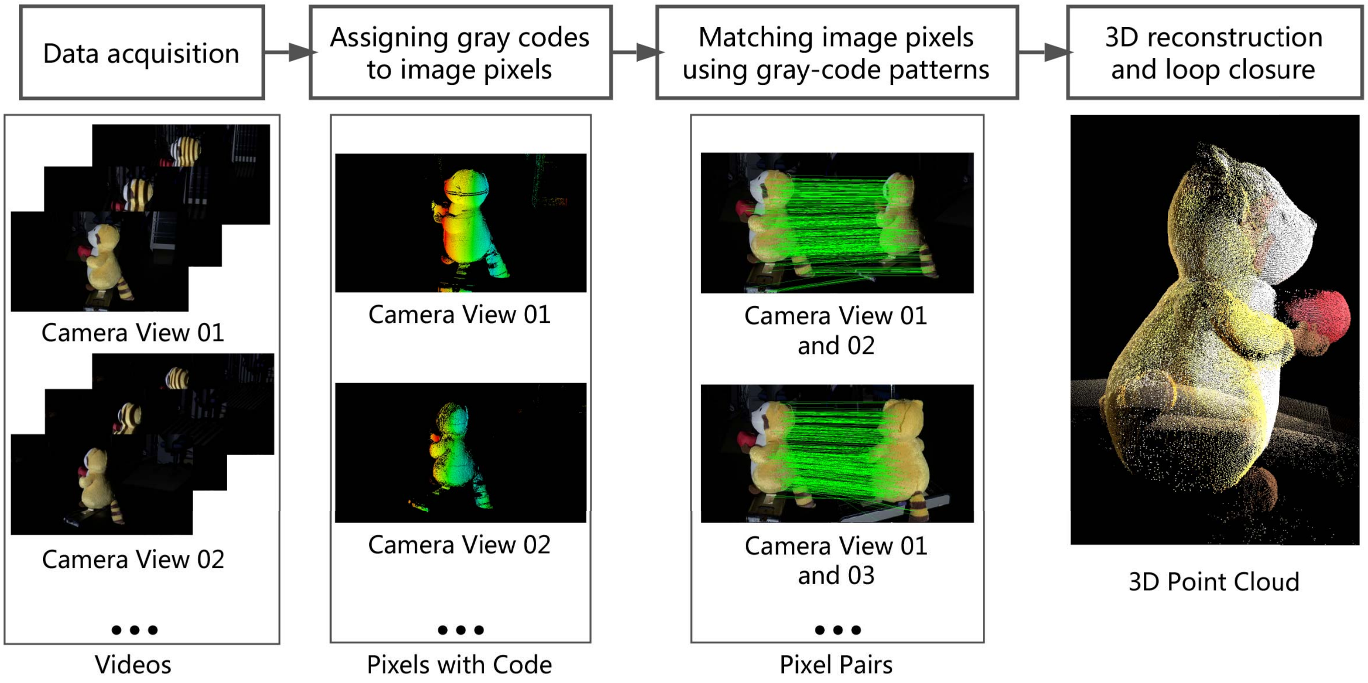


Fig. 1. Overview of the proposed 3D scanning system. First, a sequence of patterns is projected from a digital projector and taken from multiple (at least three) viewpoints by moving the camera. The projector is then moved to a different position as it covers a different side of the target while keeping one of the camera viewpoints at the same position before and after the projector move, *i.e.* two sequences of patterns are taken from the same viewpoint (please also see figure 2). The procedures above are repeated until the whole surface of the target object is scanned.

The spatial encoding enables acquisition of 3D depth with only a single-frame pattern and is therefore commonly used to capture dynamic scenes [16], [17], [18]. The temporal coding technique (time-multiplexing coding) uses the temporal information (the code words are multiplexed in time) to improve the accuracy of patterns [10], [13], [15]. In this work, we exploit a temporal coding technique for achieving high accuracy. In the structured-light system, a common approach to obtaining the whole shape of objects is to align the multiple scans (point clouds) [14] into a single global coordinate using the ICP algorithm [11]. However, aligning with ICP algorithm is not trivial because it does not perform well when an object has repetitive structures or no distinctive 3D structures.

Recently, many RGB-D cameras have been developed. For example, Tango [19] (an augmented reality computing platform developed by Google) and Kinect for Xbox One V2 (a motion sensing input device developed by Microsoft) [20] use Time-of-Flight(ToF) technologies to obtain depth information of surroundings. These novel technologies are indeed convenient to acquire 3D information of real-world scenes, however, the accuracy is not sufficient to use them for 3D archiving or 3D printings as demonstrated in section IV.

III. Structured-light SfM system

This section describes the proposed 3D scanning system that can acquire the whole shape of target objects in high quality despite their texture complexities. To achieve this, an off-the-shelf camera and projector are used to capture the data and integrate structured-light and SfM technologies.

The proposed scanning system consists of the following procedures:

- Capture a sequence of projected patterns from multiple viewpoints and record the scene in video format (Section III-A).
- Decode the captured videos to assign gray codes for image pixels (Section III-B). (This process corresponds to the feature detection and description in standard SfM.)
- Obtain the connection of gray codes on the object surface projected from different projector positions by the proposed algorithm to generate pixel mappings of various camera viewpoints among adjacent projector positions (Section III-C). (This process corresponds to the feature matching and tracking in standard SfM.)
- Generate a point cloud by the method of SfM, using pixel mappings (Section III-D), and close the loop of the reconstructed shape (Section III-E).

Figure 1 shows the overview of the proposed system.

A. Data acquisition

This system consists of off-the-shelf devices: a digital camera and a digital projector. While the encoded sequence is being projected onto the measured object, the illuminated scene is recorded by the camera as a video (or a sequence of images). Note that the camera should be kept still while recording at one camera viewpoint.

The video capturing is initiated from one camera viewpoint, the position of the projector is fixed, and then the camera is moved to another viewpoint as illustrated in figure 2. To make the reconstruction with SfM stable, data are obtained from at least three camera viewpoints for each projector position. Next the projector is moved to a different side of the object

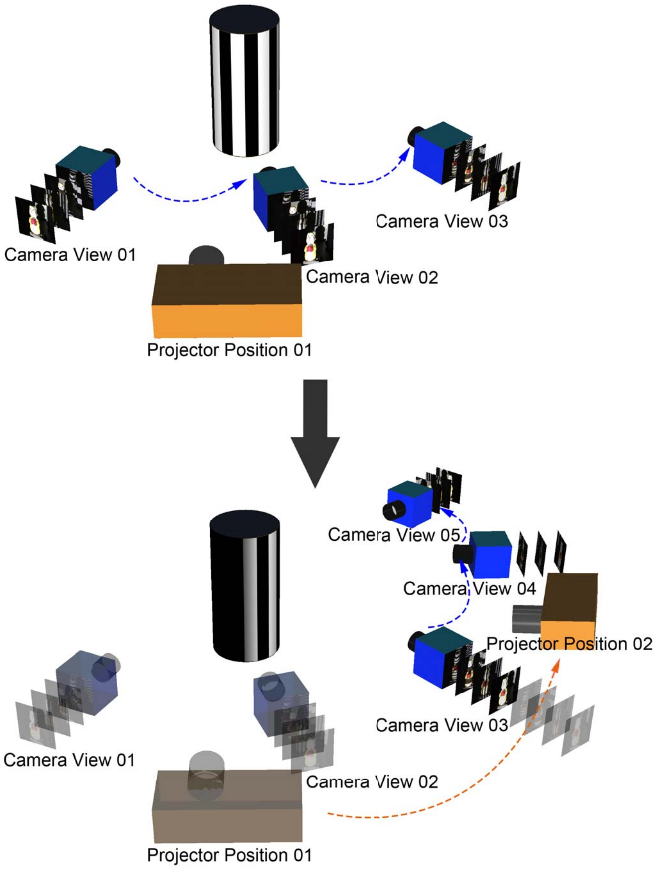


Fig. 2. Data acquisition on the proposed structured-light SfM. A sequence of gray-code patterns is captured from multiple camera viewpoints. The projector is moved to another position, but one of the camera viewpoints remains fixed (as indicated by camera viewpoint 03) to obtain the relationship of the codes on the object surfaces projected from different projector positions.

while fixing the position of the camera, so that two sequences of patterns projected from different projector positions are captured from the same viewpoint as illustrated in the bottom of figure 2. This connects the codes projected by different projector positions. The procedures above are repeated until the whole surface of the target object is scanned.

B. Assigning gray codes to image pixels

Gray code patterns generated by the projector are encoded to image pixels. The proposed structured-light system uses the same patterns used in [10] which generates $\log_2 x$ bits of gray code to assign the independent value for each pixel where x is the number of projector rows (or columns). In order to improve the accuracy of decoding, the gray code patterns insert the original and its inverse. For example, a sequence of gray code patterns consists of 42 frames (including one white image and one black image) for a projector with a resolution of 1024 × 768.

After capturing the videos at several viewpoints, the keyframes are extracted from each video stream to facilitate the decoding and then the codes assigned to the image pixels are calculated for each view. The decoding algorithm in [10] is followed which outputs a list of $\{CODE, \boldsymbol{u}\}$ of each camera viewpoints. $CODE$ denotes the unique identification in the projected pattern. $\boldsymbol{u}$ denotes the position of extracted gray code in the original image and $\boldsymbol{u} \in \mathbb{R}^2$. If the same gray code is assigned to multiple pixels, the average of pixel coordinates is computed as the final position in the image.

C. Matching image pixels using gray-code patterns

The next step is to match the pixels across images (viewpoints) using the codes assigned to them. Matching image pixels across the camera viewpoints at the common projector position is straightforward because the correspondences can be obtained by connecting the pixels which have the same $CODE$ (figure 3 (a)).

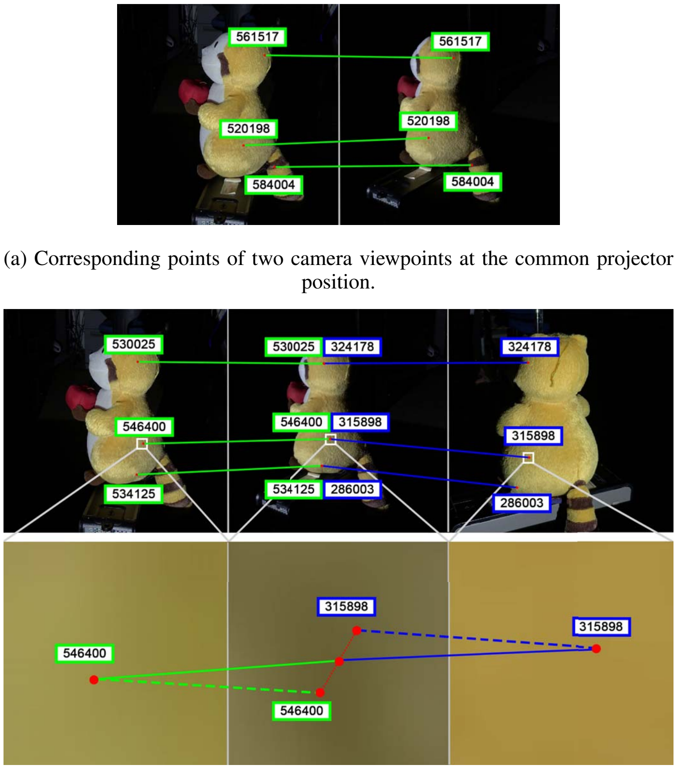
To represent the camera poses and 3D structures in a global coordinate system, it is necessary to find the projector rays intersecting at the same surface. This is achieved by connecting the projector patterns (gray codes) generated from different positions using the camera viewpoint that captures two sequences of projector patterns (as camera viewpoint 03 shown in figure 2). As shown in the middle column of figure 3 (b), the codes located at the same pixel position are linked. More specifically, for each $\{CODE_1, \boldsymbol{u}\}$ projector position N, the nearest code $\{CODE_2, \boldsymbol{v}\}$ of another projector position (N+1) can be found, and the mapping of $CODE_1$ and $CODE_2$ can be obtained. Then the code pairs are merged into one point (midpoint) and matched to the same code extracted from other camera viewpoints, as shown in figure 3 (b). Notice that the corresponding points across different projector positions appear only in the overlapping area of projected patterns on the object surface.

D. 3D reconstruction

Given image pixel matches (feature tracks) across multiple viewpoints, their 3D coordinates and camera poses are estimated using SfM. A standard incremental SfM [4] consisting of camera pose estimation, triangulation of feature tracks, and bundle adjustment is performed. This can be achieved by replacing the SIFT detection and feature matching steps to the structured-light based image pixel matching described above. In this work, a system is built on top of VisualSFM [21] but other SfM pipelines [22], [23] are also compatible.

E. Loop closure

The data acquisition (Section III-A) and image pixel matching (Section III-C) do not explicitly perform the loop closure while scanning around the object. An accumulation of reconstruction errors may result in a considerable surface discrepancy where the 3D model surface should close. The loop closure can be optimally performed by finding the 3D correspondences between overlapping surfaces reconstructed by the first and last viewpoints. In this work, the Iterative Closest Point (ICP) algorithm [11] is used to extract the 3D correspondences. Then the corresponding projection point on each camera plane can be obtained and the feature tracks linked across head and tail camera viewpoints in the loop. Finally, bundle adjustment is employed with the new correspondences to close the loop.



(a) Corresponding points of two camera viewpoints at the common projector position.

(b) Corresponding points of two camera viewpoints at different projector positions.

Fig. 3. Examples of feature matches (a) and tracks (b). The numbers in the rectangles are the code values. The green lines represent the matches between two camera viewpoints at projector position N, and the blue lines represent projector position N+1. The matches are merged at the fixed viewpoint (as the camera viewpoint 03 shown in figure 2) and extended to the track. As shown in the bottom figure, the nearest codes are merged into one point (midpoint) and matched to the same code extracted from other camera viewpoints.

IV. Experimental results

To demonstrate the proposed structured-light SfM system, an object is captured using a smartphone camera (iPhone7) with a video resolution of 1920 × 1080 and a frame rate of 120 fps, and an ASUS P3B data projector with a resolution of 1024 × 768[1]. The object is scanned by projecting the patterns from five projector positions. At each projector position, the projector patterns are captured from three camera viewpoints, *i.e.* the object is taken from 11 viewpoints in total (the left picture of figure 4 (b) shows the estimated camera poses of the Gundam model). It took 7 min to scan the object including the time of moving the cameras and projectors. The computation time for reconstruction is 181 s on an Intel(R) Core(TM) i7-6800K CPU. The experiments were conducted on two different real objects as shown in figure 4 (a) and figure 5 (a). Figure 4 (c) and figure 5 (b) show the reconstructed 3D model using the proposed method.

The reconstruction results are compared to those of a standard SfM-MVS method. The same object is constructed using VisualSFM [21] followed by the dense reconstruction (CMVS [7]). As VisualSFM did not provide any 3D reconstruction from 11 images, 288 images were captured using the smartphone (iPhone7) with a resolution of 1920 × 1080, which is the same camera and resolution used in the proposed method. The results of dense reconstruction are shown in figure 4 (d) and figure 5 (c). The VisualSFM-CMVS reconstruction of the plastic bucket failed because of the lack of texture. The computation time for VisualSFM-CMVS (run on the same computer) is 102 min which is significantly more than the time of the proposed method (even including the data acquisition time). The results clearly show that the VisualSFM-CMVS method only reconstructs the areas with some textures whereas the proposed method provides a more complete shape of the object.

The reconstruction results are also compared with that of Google Tango [19]. A Lenovo Phab 2 Pro Smartphone which includes Tango technology is used. For software, a third-party mobile application called Matterport Scenes [24] is employed, which can scan small spaces or objects and take the 3D models easily using the API of Tango. Google Tango can output the 3D model in real-time and the scanning took 23 s. Figure 4 (e) and figure 5 (d) show the 3D point cloud results reconstructed by Google Tango [19]. Qualitatively, the proposed method can provide a more accurate and detailed shape in comparison with Google Tango.

V. Conclusion

This paper presented a stable and high-accuracy entire shape measurement system based on structured-light and SfM techniques. The proposed system required no special devices and could perform highly accurate 3D reconstruction of real-world objects using standard cameras and projectors. The experimental results demonstrated that the quality of reconstructed models was more complete and accurate in comparison with the standard SfM-MVS and Google Tango models. In future, we will discuss the possibility of estimating detailed shape, illumination, and albedo of the measured 3D model in the proposed system using the Multi-View Inverse Rendering (MVIR) method [25].

Acknowledgements. This work was partly supported by JSPS KAKENHI Grant Number 17H00744.

References

[1] R. Li, T. Luo, and H. Zha, “3d digitization and its applications in cultural heritage,” in *Proc. Digital Heritage - Third International Conference*, 2010, pp. 381–388.

[2] L. Niven, T. E. Steele, H. Finke, T. Gernat, and J.-J. Hublin, “Virtual skeletons: using a structured light scanner to create a 3d faunal comparative collection,” *Journal of Archaeological Science*, vol. 36, no. 9, pp. 2018–2023, 2009.

[3] “Faro 3d,” http://www.faro.com/en-us/home.

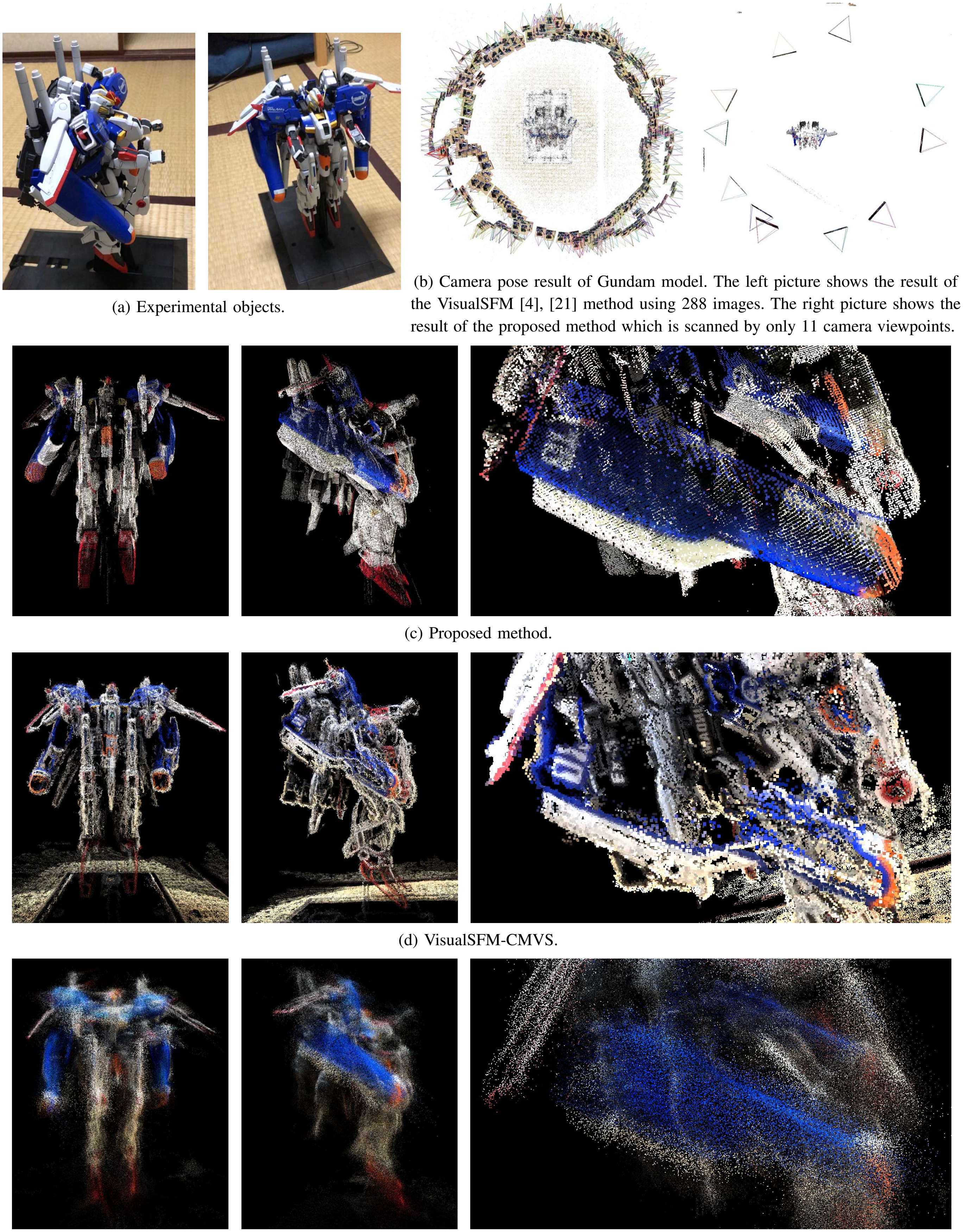
[4] N. Snavely, S. M. Seitz, and R. Szeliski, “Photo tourism: exploring photo collections in 3d,” *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[5] S. Fuhrmann, F. Langguth, and M. Goesele, “MVE - A multi-view reconstruction environment,” in *Eurographics Workshop on Graphics and Cultural Heritage, GCH*, 2014, pp. 11–18.

[6] Y. Furukawa and J. Ponce, “Accurate, dense, and robust multiview stereopsis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, 2010.

[7] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, “Towards internet-scale multi-view stereo,” in *Proc. CVPR*, 2010, pp. 1434–1441.

[1] As in other structured-light systems, the projector’s pixels must be clearly captured by the camera, thus it is suggested to use a camera which has higher resolution than the projector.



(a) Experimental objects.

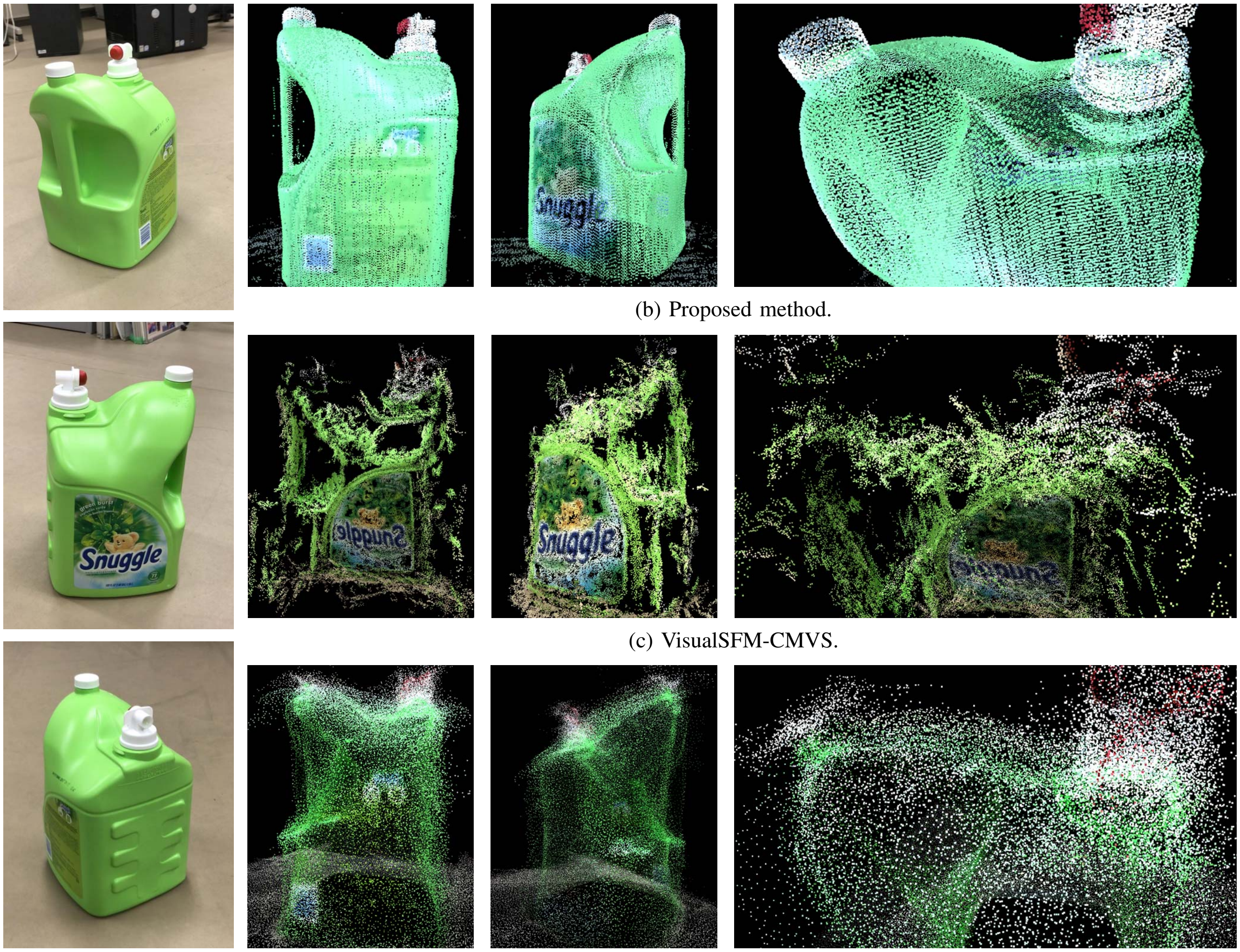
(b) Camera pose result of Gundam model. The left picture shows the result of the VisualSFM [4], [21] method using 288 images. The right picture shows the result of the proposed method which is scanned by only 11 camera viewpoints.

(c) Proposed method.

(d) VisualSFM-CMVS.

(e) Google Tango.

Fig. 4. Reconstruction of the Gundam model. The proposed method recovers fine details (c) when compared with the VisualSFM-CMVS method [4], [7] (d) and Google Tango [19] (e).



(b) Proposed method.

(c) VisualSFM-CMVS.

(a) Experimental objects.

(d) Google Tango.

Fig. 5. Reconstruction of the plastic bucket. (a) shows the experimental objects. The proposed method recovers fine details (b) when compared with the VisualSFM-CMVS method [4], [7] (c) and Google Tango [19] (d). The VisualSFM-CMVS reconstruction of the plastic bucket is incomplete because of the lack of texture.

[8] M. Jancosek and T. Pajdla, "Multi-view reconstruction preserving weakly-supported surfaces," in *Proc. CVPR*, 2011, pp. 3121–3128.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] K. Herakleous and C. Poullis, "3dunderworld-sls: An open-source structured-light scanning system for rapid geometry acquisition," *CoRR*, vol. abs/1406.6595, 2014.

[11] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[12] E. Tola, V. Lepetit, and P. Fua, "DAISY: an efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, 2010.

[13] D. Caspi, N. Kiryati, and J. Shamir, "Range imaging with adaptive color structured light," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 5, pp. 470–480, 1998.

[14] H. Nguyen, D. Nguyen, Z. Wang, H. Kieu, and M. Le, "Real-time, high-accuracy 3d imaging and shape measurement," *Applied optics*, vol. 54, no. 1, pp. A9–A17, 2015.

[15] J. Salvi, J. Pagès, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognition*, vol. 37, no. 4, pp. 827–849, 2004.

[16] H. Kawasaki, R. Furukawa, R. Sagawa, and Y. Yagi, "Dynamic scene shape reconstruction using a single structured light pattern," in *Proc. CVPR*, 2008.

[17] R. Furukawa, R. Sagawa, H. Kawasaki, K. Sakashita, Y. Yagi, and N. Asada, "One-shot entire shape acquisition method using multiple projectors and cameras," in *Proc. Fourth Pacific-Rim Symposium on Image and Video Technology (PSIVT)*. IEEE, 2010, pp. 107–114.

[18] H. Kawasaki, T. Hirukawa, and R. Furukawa, "Registration and entire shape acquisition for grid based active one-shot scanning techniques," in *Proc. ICPR*, 2016, pp. 1743–1749.

[19] "Tango," https://get.google.com/tango/.

[20] "Kinect for xbox one," http://www.xbox.com/en-US/xbox-one/accessories/kinect.

[21] C. Wu, "Visualsfm : A visual structure from motion system," http://ccwu.me/vsfm/.

[22] J. L. Schoenberger, "Colmap," https://colmap.github.io/.

[23] P. Moulon, "openmvg: "open multiple view geometry"," http://imagine.enpc.fr/~moulonp/openMVG/index.html.

[24] "Matterport scenes," https://matterport.com/matterport-scenes/.

[25] K. Kim, A. Torii, and M. Okutomi, "Multi-view inverse rendering under arbitrary illumination and albedo," in *Proc. ECCV*, 2016, pp. 750–767.