Simultaneous shape registration and active stereo shape reconstruction using modified bundle adjustment

Ryo Furukawa
Hiroshima City University, Japan
ryo-f@hiroshima-cu.ac.jp

Genki Nagamatsu, Hiroshi Kawasaki
Kyushu University, Japan
{nagamatsu.genki.310@s,kawasaki@ait}.kyushu-u.ac.jp

Abstract

Simultaneous registration and shape fusion using 3D scanners have been proposed for conducting wide-area and dense 3D shape reconstruction. However, because the 3D scanners for such a system must be robust and should provide feedback in real time, only a few devices are available, thereby limiting the application of the technique. In this study, we propose a new wide-area scanning algorithm that only requires an off-the-shelf projector and a camera. In our technique, the devices are not necessarily fixed to each other and the relative positions of the devices as well as the scene shapes can be precisely estimated by bundle adjustment (BA) in case of structured light. To efficiently perform shape registration, a robust and dense shape reconstruction is required, which is currently considered to be an open problem for structured light systems. In this study, we suggest a novel network-based feature detection algorithm as well as shape fusion algorithm for the solution.

1. Introduction

Dense 3D shape reconstruction of large area is considered to be important for various applications such as human body capture for medical systems, indoor room modeling for augmented reality systems, and entire shape acquisition for industrial purposes. To achieve this, simultaneous registration and shape fusion of sequentially scanned 3D data have been extensively researched and developed [26, 21, 22]. For these algorithms, 3D sensors are freely moved during the scan to cover a wide-area, therefore, a real-time 3D scanning capability is required. However, it is still not a easy task and only a few commercial sensors fit the requirement [20, 1, 14], thereby, with only a small number of options available for users.

As an alternative, the development of a projector and a camera system using off-the-shelf devices can be a practical option. However, those scanning systems have not been commonly used for wide-area scanning because of several reasons. For some potential applications, one main reason is the practical difficulty that is associated with fixing the projector and camera with each other during the dynamic scanning process. For example, if a video projector is used, it is usually large and heavy and difficult to be fixed tightly under the condition of SLAM-like process, where the system is installed on a vehicle and encounters many bumps and vibrations. Another typical condition can be found at medical systems, *e.g.*, endoscope or laparoscope systems. Those systems are too small and complicated, and thus, it is often impossible to tightly fix the pair of the devices. If the relative position between the projector and the camera is shifted, the pre-calibrated parameters will not be used, thereby resulting in the failure of the entire scan.

In this study, we propose a new multi-frame reconstruction method which does not require the relative position of the devices to be fixed, but estimates all the parameters based on the bundle adjustment (BA). Although the original BA for cameras assumes a certain amount of stable correspondences, it cannot be applied to the structured light systems because the projectors cannot be used to capture images. To solve this problem, we introduce a shape fitting algorithm that is inspired by the iterative closest point (ICP) algorithm and the ICP cost is jointly minimized through the process. After applying BA, since multiple shapes are precisely registered, they are fused into a global shape by truncated signed distance function (TSDF) in our method.

In the paper, we also propose a oneshot shape-reconstruction technique based on grid-based active stereo [15, 31] with a learning-based approach using a convolutional neural network (CNN) and an efficient shape interpolation technique using an anisotropic radial basis function (RBF). In our method, we design a single CNN, which is expected to reduce the computational cost as well as increase the accuracy. To increase the density as well as high frequency, we introduce an anisotropic radial basis function (RBF)-based shape densifying algorithm. The contributions of our study can be given as follows:

1. a new BA technique for structured light systems to re-

construct a consistent wide-area scene by performing sequential scans is proposed;

2. a new grid-based active stereo technique using multiple features, where two-directional parallel lines and the codes are simultaneously detected from the captured patterns using a single CNN, is proposed; and
3. a new shape-interpolation and integration technique using an anisotropic RBF and a weighted TSDF techniques is proposed.

One practical application of our study is a 3D endoscope in which the pattern projector cannot be fixed to the head. We built a micro-sized projector and demonstrated a 3D endoscopic system to successfully recover large areas.

2. Related Work

The structured light technique has been frequently used for practical 3D shape-capturing purposes [29, 32, 23]. Temporal and spatial encoding approaches are the two primary approaches that are used to encode the positional information into patterns. Because temporal encoding requires multiple images, it is not suitable to install the system in moving devices. Conversely, spatial encoding requires only a single image, and it can be implemented in dynamic and moving devices [20, 15, 19, 31]. Therefore, by employing such techniques, a projector-camera system can be freely moved throughout the scanning process to achieve wide-area shape scan. However, it is usually challenging to fix an off-the-shelf projector to a camera. There is an interesting solution for solving the aforementioned difficulty that involves a change in the devices' relative positions rather than fixing them [8, 7]. However, this technique assumes dense and precise correspondences between the camera and projector; therefore, temporal coding is required and cannot be applied to our purpose.

As we will explain later, our technique has resemblance to ICP algorithm, in which multiple shapes are registered by minimizing distances between closest points [2]. Although the original ICP algorithm does not estimate the scale changes, there have been several reports on techniques that consider the scale changes in the ICP framework [6, 18]; further, the scale changes and shape distortions that are caused by the calibration errors are compensated. To efficiently estimate those parameters, Garcia *et al*. [11] used gray-code projection for obtaining dense correspondences between multiple projectors and cameras and used BA. Our technique is inspired by their report; however, we assume dynamic motion of either a projector or a camera, and thus, their methods cannot be applied.

Another problem of spatial encoding methods is their instability because that positional information is encoded into small regions and the patterns tend to be complicated and are easily affected and degraded by the environmental conditions. To avoid such limitations, some techniques are observed to be based on geometric constraints rather than decoding [16, 27, 24, 17, 31]. However, because such techniques are still dependent on pattern detection, their results are degraded to some extent. Although there are methods that are available to compensate for such degradation of results [12, 13], they assume that multiple images are captured by projecting multiple patterns; thus, these methods cannot be applied to perform a one-shot scan. Recently, a solution was proposed for the subsurface scattering objects [10, 28]; however, it required a specifically designed pattern, whereas a general technique that used extensively varying patterns was in considerable demand. In this study, we use fully convolutional neural networks (FCNNs) that can be referred to as U-Nets [25] for decoding 2D grid-like patterns that are projected onto the target. A U-Net is an FCNN architecture that receives an image and produces a labeled image. It is known to outperform previous FCNN architectures, including the sliding window convolutional networks [4] in segmentation tasks for medical images. Song *et al*. [30] proposed to decode an active stereo pattern using a CNN.

In the final step, the shape integration of multiple scans is required. Further, the signed distance field (SDF) representation has been extensively used [5] and TSDF has been recently proposed to achieve real-time system [21]. However, both techniques erase high frequency shapes.

3. Overview

3.1. System configuration

Our proposed 3D measurement system comprises a projector and a camera, as depicted in Fig. 1(left). We assume that the devices are not necessarily tightly fixed to each other. The temporally or spatially encoded patterns are projected by a projector and captured by a camera; further, these patterns are decoded to retrieve the correspondences between the image and the pattern by the algorithm. In the experiments, we used grid oneshot scan, which will be explained in the next section. To show the strength of our method, we actually developed an endoscopic system (Fig. 1(right)) in our experiments. In the system, a fiber-shaped micro pattern projector (a diffractive optical element (DOE)-based pattern projector) is inserted through the instrument channel of the endoscope; the projector slightly protrudes from the endoscope head, as depicted in Fig. 2(a), and emits structured light. Because the micro pattern projector is smaller than the channel, the relative position and orientation between the projector and camera varies through scan.

3.2. Grid-based active stereo for endoscope

In the endoscopic system, we use the "grid pattern with gapped codes" described in [10], which is based on grid pat-

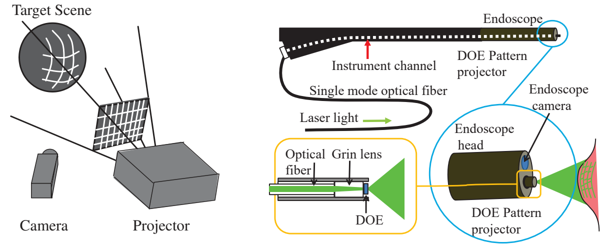


Figure 1. Schematics of the 3D measurement of projector-camera and endoscopic systems.

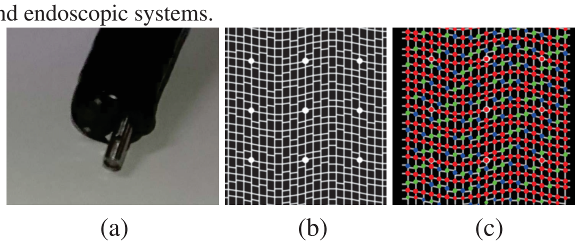


Figure 2. DOE micro projector. (a) The projector inserted through the instrument channel of an endoscope. (b) The projected pattern. (c) The embedded codewords where S is colored in red, L is colored in blue, and R is colored in green. S indicates the edges of the left; further, the right sides exhibit the same height. L indicates that the left side is higher, whereas R indicates that the right side is higher.

tern projection [15, 27, 31]. The basic reconstruction algorithm using a grid pattern projector can be given as follows.

First, a static grid pattern is projected onto the object and is subsequently captured by a camera. Then, the grid structures and marker points are extracted from the captured image. The extracted marker points are used for auto-calibration to estimate the projector pose; note that since the number of the marker is limited, the precision of the estimated poses is usually low and this is our motivation to improve it. Because the pattern comprises a grid structure, only the intersection points are considered to be candidates for retrieving the correspondences between the captured image and the pattern. Normally, there are several grid points that are detected on an epipolar line on the captured image which corresponds to a single grid point on the pattern. In grid-based techniques, the consistency of the graph structure is efficiently used such that a single solution remains. To increase robustness, different pattern features are used, including the colored lines [31], modulated lines [27], or gaps between adjacent edges [10].

The pattern that is used in our experiments is based on [10] and depicted in Fig. 2(b). In the pattern, a discrete feature (gap code) is attached to each of the grid points that are represented by the level gap between the left and right edges of the grid point. The classes of code are either S, L, or R as denoted using different colors in Fig. 2(c). For 3D reconstruction, three types of features must be extracted from the input image.

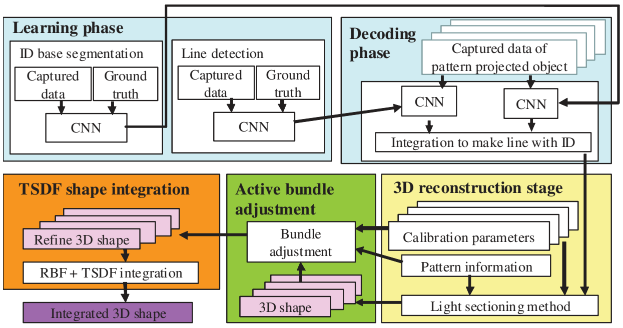


Figure 3. Overview of the wide-area shape reconstruction by simultaneous registration and shape integration from multiple scans.

3.3. Basic procedure of the method

A sequence of images is captured by the camera while the structured light is projected. Further, the 3D shapes are recovered for each frame. The 3D reconstruction of each frame comprises the three stages of pattern decoding, auto-calibration, and 3D reconstruction, which are presented in the flowchart in Fig. 3. The pattern decoding stage is processed by CNNs that are trained to extract the grid-like structures and the gap codes in the captured images. The different types of features are simultaneously extracted using a single CNN. In the auto-calibration stage, the camera parameters are self-calibrated by the method proposed in [9] using the special markers detected by the CNN. This is conducted because the projector and camera are not fixed to each other. In the 3D reconstruction stage, the extracted grid structures and code information are analyzed; further, the IDs of all the detected vertical lines are decided by the grid-based reconstruction method that has been previously described, and the 3D curves are reconstructed using a light sectioning method. All the estimated parameters and recovered shapes are used as a initial value to conduct BA to refine the 3D shapes as well as the calibration parameters of the projector-camera system, that can be referred to as active bundle adjustment (active-BA). Finally, the integrated shapes are constructed by the KinectFusion-like algorithm based on TSDF.

4. Bundle Adjustment for Structured Light

4.1. Overview

A structured light system is used to capture a target scene multiple times while the projector and camera system is freely moved. Here, we describe the condition of our shape acquisition, registration and integration process.

1. The target scene is rendered by a combination of data obtained from the pattern projector and camera. Further, the correspondences between the projected pattern and the captured image can be obtained by analyzing the image, *i.e.*, using a grid-based method [15].

The obtained correspondences is sparse because of the feature of grid-based reconstruction.

2. The relative pose between the projector and the camera can be uncalibrated.
3. The same scene is captured multiple times while assuming independent motion of the camera and the projector. We can denote a single capture of the scene as a frame. There are no specific constraints that are imposed on the motions of the camera and projector.
4. The textural information of the scene is not used, *i.e.*, the camera is used only for capturing the projected pattern, but not the textures of the scene, because projected pattern is usually stronger than the textures.
5. In our technique, we do not implement explicit loop detection algorithm, so if a drift becomes so large that correspondences cannot be found, our technique will fail. If correspondences are found, our technique efficiently spread the accumulated errors to entire scene evenly to achieve consistent shape reconstruction.

The objective of this study is to generate a integrated shape when multiple shape measurements are being consistently merged. This should be applied, for example, on the described 3D endoscopic system or any projector-camera system that is used for measuring the 3D shape of a wide area while a projector and a camera cannot be tightly fixed each other.

A naive method to achieve the objective would be to generate 3D reconstruction frame-by-frame and to align the 3D reconstructed shapes using an ICP algorithm. If the projector-camera poses are uncalibrated, they can be auto-calibrated using the projector-camera correspondences. However, such an approach is problematic. The auto-calibration results exhibit genuine ambiguity in terms of scales. Furthermore, the low-precision relative poses between the projector and the camera cause distortion in 3D reconstruction. Because of scale inconsistencies and shape distortions, the shapes from different frames do not fit consistently even after the ICP algorithm is applied.

Instead, we use a BA technique, where the information obtained from the measurements of multiple frames is used as a single solution. Unlike the BA that is conducted when only a camera is used, the texture of the scene cannot be used in BA because the projector cannot be used to capture an image. Therefore, we cannot use the information from “a single point captured from different frames” (note that the projected patterns are not fixed to the scene). Overall, the BA technique cannot be applied without performing any modification.

By considering the defined objective and its associated problems, the only usable constraint between different frames is that their shapes should be fit by a rigid transformation. The ICP algorithm has been a common tool for dealing with this type of constraint. Therefore, we propose the usage of an iterative algorithm when the ICP algorithm and BA techniques are combined. In the proposed process, the shapes of different frames are aligned with the ICP algorithm, and inter-frame correspondences are obtained with the ICP-inspired method. Using the inter-frame correspondences, BA is processed to obtain a better fitting between the shapes of different frames. These processes are repeated until convergence of BA. We can refer to the proposed algorithm as active-BA.

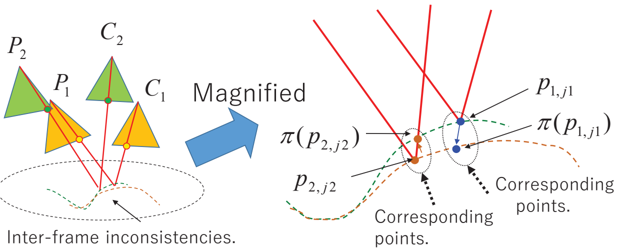


Figure 4. Active bundle adjustment correspondence finding process. This process is equivalent to ICP.

4.2. Active bundle adjustment

The steps of active-BA can be given as follows:

Step 1 The initial parameters of the relative pose between the projector and the camera as well as the positions of the shapes of the frames are given as input.

Step 2 The 3D shape for each frame is reconstructed from the current pose information.

Step 3 The corresponding points between different frames are sampled using a proximity relation between the frame surfaces, which is similar to the process that is used to retrieve the corresponding point pairs in the original ICP algorithm.

Step 4 Using the obtained inter-frame corresponding points, the corresponding intra-frame information, which is assumed to be known by spatial coding, is propagated to another frame.

Step 5 Using the corresponding intra- and inter-frame information, an algorithm which minimizes both of the reprojection errors of 3D points within each frame and distances between the corresponding points between different frames. Using all the output poses, dense 3D shapes and the relative position between the frames are updated.

Step 6 The above steps are repeated until convergence.

For the relative pose of Step 1, the precision of the initial relative pose may be low; however, it should be sufficiently precise to achieve convergence.

In Step 2, if the correspondences between the projected pattern and the camera image is sparse, the reconstructed shape becomes sparse. It is difficult to align the sparse shapes because the correspondence points between the frames are observed to be small (note that the ICP algorithm does not work between sparse points). Thus, we apply a shape interpolation method to densify the reconstructed shapes. In our method, normal vectors are obtained by 2D regression for each point of the coarse shape data, and RBF-based interpolation is applied using the normal information.

Step 3 is described using Fig. 4. Let the camera and the projector of frame k be C_k and P_k, respectively. By processing the image of frame k (captured with C_k), the 2D correspondences between C_k and P_k are obtained. Let the jth pair of correspondences be a pair of $\mathbf{u^c_{k,j}}$ of C_k and $\mathbf{u^P_{k,j}}$ of P_k. The 3D point obtained by the triangulation of $\mathbf{u^c_{k,j}}$ of C_k and $\mathbf{u^P_{k,j}}$ of P_k can be denoted by $\mathbf{p_{k,j}}$. Further, for all the correspondence pairs, frame k is reconstructed (*i.e.*, repeated for j).

Let frame l be another frame; further, all the correspondence pairs of frame l are also reconstructed. If the reconstructed points $\mathbf{p_{l,j}}$ are sparse, they should be interpolated, and the depth image D_l with the view C_l is obtained. Further, $\mathbf{p_{k,j}}$ is projected onto D_j using the pose and intrinsic parameter of C_l. If the projected pixel is a valid 3D point, we can define this point in the corresponding point of $\mathbf{p_{k,j}}$ in frame l. This corresponding point is $\pi_l(\mathbf{p_{k,j}})$. The 2D projection of $\pi_l(\mathbf{p_{k,j}})$ can be calculated using the camera C_l and projector P_l. Let these 2D points be $\mathbf{v^c_{k,j,l}}$ and $\mathbf{v^P_{k,j,l}}$, respectively. $\mathbf{p_{k,j}}$ and $\pi_l(\mathbf{p^c_{k,j}})$ are corresponding points between different frames. Generally, they are different but are the neighboring points of frames k and j.

In our algorithm, we calculate BA-style reprojection errors of the points $\mathbf{p_{k,j}}$ and $\pi_l(\mathbf{p^c_{k,j}})$, respectively within each frames (k and l), and the distance errors between the corresponding points. Then, the total cost to be minimized is the weighted sum of reprojection errors of all points $\mathbf{p_{k,j}}$ and distance errors of all the pairs of $\mathbf{p_{k,j}}$ and $\pi_l(\mathbf{p^c_{k,j}})$.

The cost function to be minimized is as follows:

$$\begin{aligned} L(I,E,P) = \sum_k \sum_j \{ & \mathrm{reproj}(\mathbf{p_{k,j}}; I_{C_k}, E_{C_k}) \\ & + \mathrm{reproj}(\pi_l(\mathbf{p^c_{k,j}}); I_{P_k}, E_{P_k})\} \\ & + w_c |\mathbf{p_{k,j}} - \pi_l(\mathbf{p^c_{k,j}})|^2 \\ & + w_b \{S(E) - \mathrm{Const}\}^2 \end{aligned} \quad (1)$$

where I_{C_k} and E_{C_k} are intrinsic and extrinsic parameters of camera C_k, I_{P_k} and E_{P_k} are intrinsic and extrinsic parameters of projector P_k, reproj() is BA-style reprojection errors, I is the set of intrinsic parameters I_{C_k} and I_{P_k}, E is the set of extrinsic parameters E_{C_k} and E_{P_k}, and P is the set of $\mathbf{p_{k,j}}$ and $\pi_l(\mathbf{p^c_{k,j}})$. $L(I,E,P)$ is minimized with respect to I,E and P. $S(E)$ is a scale function that determines the scale of the scene, and Const is a constant value. For example, $S(E)$ can be the sum of distances between the positions of the camera or projector for all the frames. We can use the sum of distances for randomly sampled devices and frames to reduce computational costs. w_c and w_b are weights for cost terms. Without the term $w_b(S(E) - C)^2$ that fix the scale of the scene, the whole scene may shrink with similarity transform infinitely, since shrinking the scene reduces the distances $|\mathbf{p_{k,j}} - \pi_l(\mathbf{p^c_{k,j}})|$ without effecting reprojection functions.

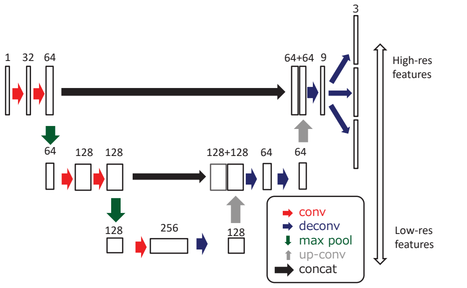


Figure 5. Structure of the U-Net used for feature detection (line/gap code). The numbers represent the dimensions of the feature maps.

In real applications, such as the endoscopic systems in this paper, intrinsics parameters of the camera for all frames k is the same. Thus, we use a common intrinsics for all k.

The active-BA can be regarded as a variation of the ICP algorithm, where ICP is used to estimate only a rigid transformation between the frames; however, in this study, the proposed algorithm estimates the projector-camera relative pose, which deeply affects the shapes of the frames.

5. Implementation

5.1. Simultaneous detection of grid structures and codes

We propose the extraction of grid structure and gap-code information using U-Nets [25]. The structure of the network is depicted in Fig. 5. The numbers in the figure represent the dimensions of the feature maps. Due to this network structure, both the fine and coarse resolution features can be learned. By applying U-Net to an image, feature maps can be generated in the same size as that of the input image. In Fig. 5, the output layers are divided into three groups, each comprising the vertical, horizontal lines, and gap-codes.

The training process of U-Net for vertical-line detection is as follows. First, the image samples of the pattern-illuminated scene are collected. Further, the vertical line locations for the image samples are manually annotated as curves of 1-dot widths. Because the 1-dot width curves are considerably narrow to be directly used as labeled regions of teacher data, regions with 5-dot width to the left and right sides of the thin curves are extracted and are labeled as teacher data (Fig. 6). The line detection and code detection are simultaneously processed using a single U-Net. The training data for code classification are depicted in Fig. 6.

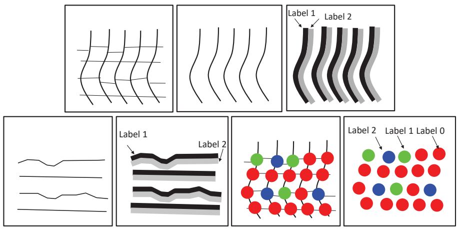


Figure 6. Training data: (top row, left to right) sample image, manually annotated vertical line, labeled image for training vertical line detection; (bottom row, left to right) manually annotated vertical line, labeled image for training horizontal line detection, manually annotated code data, labeled image for training code detection.

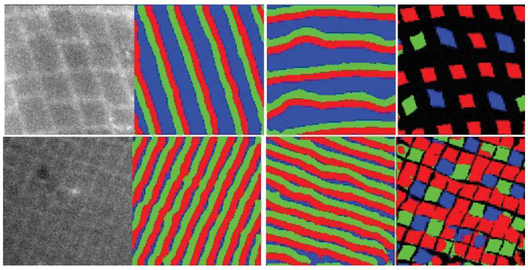


Figure 7. Samples of CNN output after training with data augmentation, (left to right) Sample images, output of vertical line detection, horizontal line detection, and code detection results.

The U-Net is trained using the loss function of the softmax entropy. In this training process, we augment the training data by adding noise and scaling the intensity because the intensity of the illuminated patterns may considerably change. Examples of the U-Net output are depicted in Fig. 7. Further, the dark, bright, noisy images with different grid sizes and up to 40-degree rotation are processed accurately.

By applying the trained U-Net to the image, the 3-way labeled image for vertical line detection is obtained, where the left and right sides of the vertical curves are labeled as 1 and 2, respectively (the green and red regions in Fig. 7). By extracting the borders between these regions, curves can be detected.

The classification of gap codes can be processed by directly applying U-Net to the image signal and not from the line detection results. Thus, the gap code estimation does not depend on line detection, which is an advantage because the line detection errors do not propagate to code decoding.

5.2. RBF-based shape interpolation

To achieve stable calculation of the inter-frame correspondences, the sparse 3D points that are obtained from the sparse projector-camera correspondences should be densified. To achieve this, we use an RBF for interpolation of the 3D curves [3].

The interpolation sizes are defined by RBF functions. In case of using an isotropic RBF, to interpolate the pixels between adjacent lines of the grid, the scale parameter of the Gaussian should be larger or approximately equal to the apparent sizes of the grid in the captured images. However, this setting often oversmooths the measured shape features of the 3D curves, which can be accurately reconstructed by the light sectioning method.

Table 1. Residuals of planes (RMSE [mm]) and plane angles (ground truth is 90°) for our method and KinectFusion (configurations shown in Fig. 10).

	Initial	Proposed	Kinect		
			Fig. 10(b)	(c)	(d)
Plane 1	4.64	1.34	3.12	1.95	2.40
Plane 2	6.63	1.52	2.08	1.74	1.38
Angle	78.3	91.6	90.0	84.8	70.1

To overcome this oversmoothing problem, we propose the use of an anisotropic RBF in which the kernel shape is observed to be long with respect to the direction that is perpendicular to the grid line. In a simple case in which the reconstructed lines are vertical in the image plane, we use a Gaussian kernel that is narrow in the vertical direction (anisotropic RBF). In this case, the vertical size of the kernel becomes 0.3 times of the original size. Further, the shape features along the line direction are expected to be preserved more than isotropic RBF.

6. Experiment

6.1. Evaluation of active bundle adjustment for a structured light system

To evaluate active-BA, we capture a scene multiple times (19 sets in the experiment) using the projector-camera system. A grid pattern is projected onto the object and the images are captured by the camera. The relative positions of the projector and the camera are slightly different for each frame to simulate the loose connection.

Under these conditions, auto-calibration and 3D reconstruction are processed for each frame. Because of frame-wise auto-calibration, the scales of different frames are inconsistent; further, calibration errors are also present. To create meaningful results from the auto-calibration method, we set a constant value as the baseline for all the frames. To emphasize the calibration errors, we also added Gaussian noise to the extrinsic parameters.

Using these data, the result of ICP alignment for the initial shapes and the result of the active-BA are shown in Fig. 8. While the result of aligning the initial shape exhibits considerable differences between different frames, the differences that are present in the result of active-BA are drastically minimized. We also scan the same object using KinectFusion [21] and results are shown in Fig. 10. Since there is no loop closure mechanism on KinectFusion, inconsistent shapes are reconstructed after entire shape scan.

For performing the evaluation, we apply plane fitting to

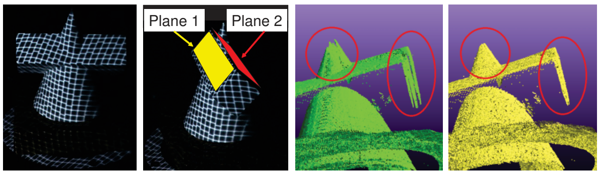


Figure 8. Active bundle adjustment results: (left to right) two captured images, before and after active bundle adjustment.

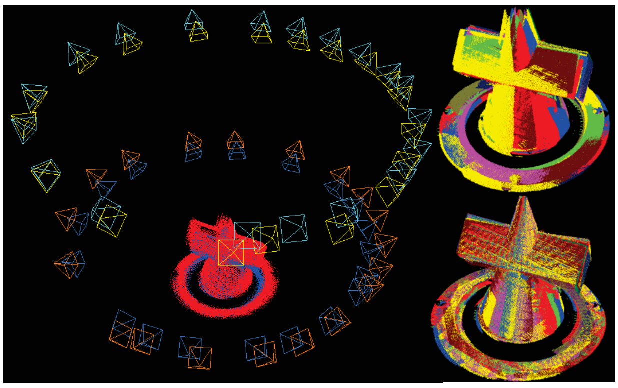
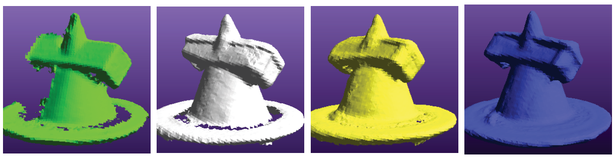


Figure 9. (Left) Camera and projectors' positions and reconstructed points before/after BA. Yellow/cyan for cameras, orange/azure for projectors and red/blue for points. (Right top) 3D points before BA and (right bottom) after BA.



(a) (b) (c) (d)

Figure 10. (a) single captured shape by Kinect v1, (b) KinectFusion result with single rotation around the object, (c) two times rotation and (d) three times rotation. Note that since there is no explicit loop closure mechanism on KinectFusion, a large inconsistency in shape remains even if several rotations are conducted.

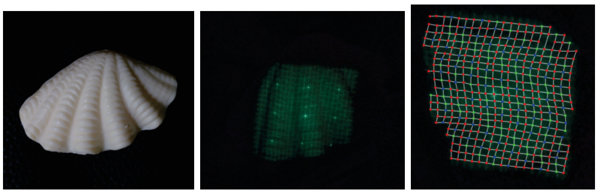
each plane of the object and calculate both the distances from the estimated planes as depicted in Fig. 8, and the RMSEs relative to the ground truth shapes are given in Table 1. Results indicate that our technique can successfully recover consistent shapes because of simultaneous optimization approach considering loop closure.

Table 2. RMSE of each result.

	Fig. 12(b) Anisotropic RBF	Fig. 12(c) Isotropic RBF
RMSE	1.60mm	1.74mm

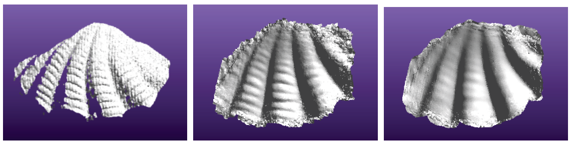
6.2. Evaluation of the effect of anisotropic RBF and TSDF-based shape fusion

To evaluate the anisotropic RBF, we scan a seashell-shaped piece of soap (Fig. 11) including isotropic RBF, and compare the results with the ground truth. The results are presented in Fig. 12 and Table 2. Our proposed method that uses the anisotropic RBF features the finest details, and its RMSE is the optimal as compared to isotropic RBF.



(a) (b) (c)

Figure 11. Grid and code detection results for seashell-shaped soap: (a) the appearance of the sample; (b) the captured image; and (c) extracted grid structures and codes using U-net



(a) (b) (c)

Figure 12. Comparison with the ground truth of a seashell-shaped piece of soap: (a) ground truth captured by gray code; (b) our method with anisotropic RBF; (c) our method with isotropic RBF.

6.3. Active BA and shape integration with TSDF using the endoscopic system

We apply our online shape registration and merging algorithm to the static objects, *i.e.*, phantom of stomach, and results are presented in Fig. 13. In Fig. 13(top), the area of the recovered shape from a frame of the captured sequence is shown by the blue polygon. Fig. 13(bottom left) depicts the captured image of the region highlighted by the yellow rectangle in top figure, where the grid pattern is projected to the surface. In the image, the grid lines are disconnected by the high-frequency shape of the model surface and shapes are distorted, as depicted in Fig. 13(bottom right). The integrated shape generated by active-BA and the online registration and merging algorithm is depicted in Fig. 14(bottom row). A large area is successfully recovered, and the high-frequency shape details are retained. For performing comparisons, we scan the same object with KinectFusion [21] and the result is presented in Fig. 14(top row); the shape is heavily distorted because of sparse reconstruction without consistent shape recovery algorithm. RMSE is 3.63mm for our active-BA, whereas 4.55mm for KinectFusion.

Please refer to the supplementary materials for more visual results

6.4. Active BA and shape integration for wide area using projector and camera system

To confirm the strength of our method for common projector and camera based system, we scan the entire room using the pro-cam system as shown in Fig. 15. We rigidly

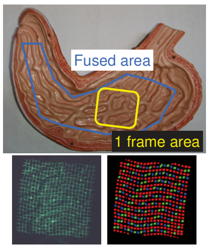


Figure 13. (Top) the phantom model; (bottom left) the captured image of the yellow rectangle in top figure; and (bottom right) the CNN result.

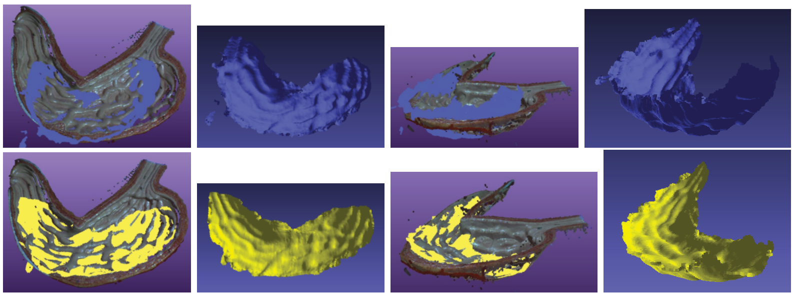


Figure 14. (Upper row) results of KinectFusion, and (bottom row) results of our method. Ground truth 3D shape is also shown in the figures. Note that shapes are heavily distorted by KinectFusion, whereas our technique can recover consistent shape because of global optimization by BA.

attached projector and a camera using professional camera tools, however, since projector is heavy, there is a possiblity that slight motion occurs during moving the system. More importantly, calibration of such in-out scenario is usually not easy, because it is usually difficult to prepare large calibration objects. We apply common calibration technique using natural feature points, which are extracted from the scene for initial shape reconstruction. For comaprison, we conducted a projector and a camera calibration technique using BA [7]. Since the system is tightly fixed, results are mostly comparable, thereby effectiveness of our method is confirmed. Note that if the projector and the camera's relative position is moved, the technique [7] cannot be applied.

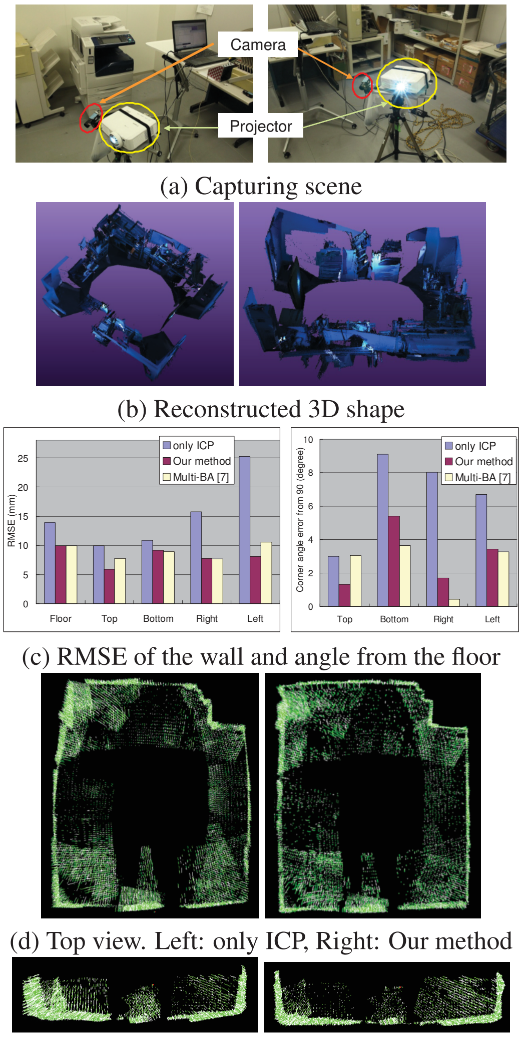
Please refer to the supplementary materials for more visual results

7. Conclusion

We propose and demonstrate a wide-area scanning technique using an off-the-shelf projector and camera. The projector and camera are not required to be fixed with each other and can be freely moved to scan the entire shape of the target object or wide region of the scene. The shapes and relative position between the projector and the camera can be precisely estimated using our technique that can be referred to as active-BA. Because the technique uses ICP to retrieve the correspondences between frames, a robust and dense reconstruction is required. For performing robust shape reconstruction, a single U-Net based network is simultaneously used for both line detection and code-based segmentation. Finally, using our method, wide-area 3D shapes can be successfully reconstructed from both small and large scene. In future, we intend to develop a real-time system for autonomous robot-based navigation.

Acknowlegement

This work is supported by JSPS/KAKENHI 16H02849, 16KK0151, 18H04119, 18K19824, and MSRA CORE14.



(a) Capturing scene

(b) Reconstructed 3D shape

(c) RMSE of the wall and angle from the floor

(d) Top view. Left: only ICP, Right: Our method

(e) Side view. Left: only ICP, Right: Our method

Figure 15. Wide-area with "loop closure." The room, size 3m×5m, was captured by a camera and a projector system, which are tightly fixed each other. Although some parts look disconnected because of visualization reason, shapes are densely captured and closed.

References

[1] Artec. United States Patent Application 2009005924, 2007j. 1

[2] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992. 2

[3] J. C. Carr, W. R. Fright, and R. K. Beatson. Surface interpolation with radial basis functions for medical imaging. *IEEE transactions on medical imaging*, 16(1):96–107, 1997. 6

[4] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012. 2

[5] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH 96. ACM*, pages 303–312, 1996. 2

[6] S. Du, N. Zheng, L. Xiong, S. Ying, and J. Xue. Scaling iterative closest point algorithm for registration of m?d point sets. *Journal of Visual Communication and Image Representation*, 21(5):442 – 452, 2010. Special issue on Multi-camera Imaging, Coding and Innovative Display. 2

[7] R. Furukawa, K. Inose, and H. Kawasaki. Multi-view reconstruction for projector camera systems based on bundle adjustment (oral). In *IEEE International Workshop on Projector-Camera Systems 2009*, pages 69–76, 2009. 2, 8

[8] R. Furukawa and H. Kawasaki. Uncalibrated multiple image stereo system with arbitrarily movable camera and projector for wide range scanning. In *IEEE Conf. 3DIM*, pages 302–309, 2005. 2

[9] R. Furukawa, R. Masutani, D. Miyazaki, M. Baba, S. Hiura, M. Visentini-Scarzanella, H. Morinaga, H. Kawasaki, and R. Sagawa. 2-DOF auto-calibration for a 3D endoscope system based on active stereo. In *The 37th EMBC*, pages 7937–7941, Aug 2015. 3

[10] R. Furukawa, H. Morinaga, Y. Sanomura, S. Tanaka, S. Yoshida, and H. Kawasaki. Shape acquisition and registration for 3D endoscope based on grid pattern projection. In *The 14th ECCV*, volume Part VI, pages 399–415, 2016. 2, 3

[11] R. R. Garcia and A. Zakhor. Geometric calibration for a multi-camera-projector system. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 467–474, Jan 2013. 2

[12] M. Gupta and S. K. Nayar. Micro Phase Shifting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Jun 2012. 2

[13] M. Gupta, Q. Yin, and S. K. Nayar. Structured light in sunlight. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013. 2

[14] Intel. Intel realsense SR300. 1

[15] H. Kawasaki, R. Furukawa, , R. Sagawa, and Y. Yagi. Dynamic scene shape reconstruction using a single structured light pattern. In *CVPR*, pages 1–8, June 23-28 2008. 1, 2, 3

[16] H. Kawasaki, S. Ono, Y. Horita, Y. Shiba, R. Furukawa, and S. Hiura. Active one-shot scan for wide depth range using a light field projector based on coded aperture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3568–3576, 2015. 2

[17] T. P. Koninckx and L. Van Gool. Real-time range acquisition by adaptive structured light. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):432–445, 2006. 2

[18] C. Li, J. Xue, N. Zheng, S. Du, J. Zhu, and Z. Tian. Fast and robust isotropic scaling iterative closest point algorithm. In *2011 18th IEEE International Conference on Image Processing*, pages 1485–1488, Sept 2011. 2

[19] Mesa Imaging AG. SwissRanger SR-4000, 2011. http://www.swissranger .ch/index.php. 2

[20] Microsoft. Xbox 360 Kinect, 2010. http://www.xbox.com/en-US/kinect. 1, 2

[21] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *IEEEISMAR*, pages 127–136, 2011. 1, 2, 6, 7

[22] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, June 2015. 1

[23] M. O'Toole, S. Achar, S. G. Narasimhan, and K. N. Kutulakos. Homogeneous codes for energy-efficient illumination and imaging. *ACM Trans. Graph.*, 34(4):35:1–35:13, July 2015. 2

[24] M. Proesmans and L. Van Gool. One-shot 3d-shape and texture acquisition of facial data. In *Audio-and Video-based Biometric Person Authentication*, pages 411–418. Springer, 1997. 2

[25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2, 5

[26] R. B. Rusu and S. Cousins. 3d is here: Point cloud library (pcl). In *ICRA*, pages 1–4, 2011. 1

[27] R. Sagawa, Y. Ota, Y. Yagi, R. Furukawa, N. Asada, and H. Kawasaki. Dense 3D reconstruction method using a single pattern for fast moving object. In *ICCV*, pages 1779–1786, 2009. 2, 3

[28] R. Sagawa and Y. Satoh. Illuminant-camera communication to observe moving objects under strong external light by spread spectrum modulation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2317–2325, July 2017. 2

[29] J. Salvi, J. Pages, and J. Batlle. Pattern codification strategies in structured light systems. *Pattern Recognition*, 37(4):827–849, 4 2004. 2

[30] L. Song, S. Tang, and Z. Song. A robust structured light pattern decoding method for single-shot 3d reconstruction. In *2017 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 668–672, July 2017. 2

[31] A. Ulusoy, F. Calakli, and G. Taubin. One-shot scanning using de bruijn spaced grids. In *Proc. The 2009 IEEE International Workshop on 3-D Digital Imaging and Modeling*, 2009. 1, 2, 3

[32] J. Wang, A. C. Sankaranarayanan, M. Gupta, and S. G. Narasimhan. Dual structured light 3d using a 1d sensor. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 383–398, 2016. 2