Wide-area shape reconstruction by 3D endoscopic system based on CNN decoding, shape registration and fusion

Ryo Furukawa[1], Masaki Mizomori [1], Shinsaku Hiura [1],
Shiro Oka[2], Shinji Tanaka[2] and Hiroshi Kawasaki[3]

[1]Hiroshima City University, Japan,
{ryo-f@, mizomori@ime.,hiura@}hiroshima-cu.ac.jp
[2]Hiroshima University Hospital, Japan, {oka4683, colon}@hiroshima-u.ac.jp
[3]Kyushu University, Japan, kawasaki@ait.kyushu-u.ac.jp

Abstract. For effective *in situ* endoscopic diagnosis and treatment, dense and large areal shape reconstruction is important. For this purpose, we develop 3D endoscopic systems based on active stereo, which projects a grid pattern where grid points are coded by line gaps. One problem of the previous works was that success or failure of 3D reconstruction depends on the stability of feature extraction from the images captured by the endoscope camera. Subsurface scattering or specularities on bio-tissues make this problem difficult. Another problem was that shape reconstruction area was relatively small because of limited field of view of the pattern projector compared to that of the camera. In this paper, to solve the first problem, learning-based approach, *i.e.*, U-Nets, for efficient detection of grid lines and codes at the detected grid points under severe conditions, is proposed. To solve the second problem, an online shape-registration and merging algorithm for sequential frames is proposed. In the experiments, we have shown that we can train U-Nets to extract those features effectively for three specimens of cancers, and also conducted 3D scanning of shapes of a stomach phantom model and a surface inside a human mouth, in which wide-area surfaces are successfully recovered by shape registration and merging.

1 INTRODUCTION

Endoscopic diagnosis and treatment on digestive tracts have become popular and widespread because of effectiveness on finding tumors in early-stage or little suffering on surgery. For this reason, an easy to deploy, accurate tumor size estimation technique is required for endoscopic systems and has been intensively researched. On our continuous works on the development of a 3D endoscope system to automatically measure the shape and size of living tissue based on active stereo, we made non-contact measurement systems by making ultra-small projectors which are possible to be inserted through the instrument channel of

ordinary endoscopes [1–5]. Using those devices, we have successfully measured several *ex vivo* human tumor samples. One significant limitation of the current systems is that it easily fails to recover shapes because of strong subsurface scattering and specular effects which is common in internal tissue. Another issue is that shape reconstruction area was relatively small because of limited field of view of pattern projector compared to that of the camera of the endoscope.

In this paper, to solve the pattern detection problems caused by complicated surface reluctances, such as sub-surface scattering and specularities, we propose a learning-based approach, which is based on CNNs (convolutional neural networks). To apply CNN to oneshot scan, we used two types of U-Nets for line detections (horizontal and vertical) and code detection, since each of the tasks is simplified and easy to learn. Then, at a decoding phase, two outputs of the U-Nets from the single captured image are integrated to make the final output, *i.e.*, detected lines with ID. Using the final output, 3D shapes are reconstructed by light sectioning method using decoded IDs.

Since each region of reconstruction is small, online shape-registration and merging algorithm for sequential frames is required to recover the wide structures of the entire shape. For the purpose, we propose a shape registration and merging algorithm in the paper. In the method, we introduce RBF-based shape densifing algorithm to fill holes between grid lines. Then, ICP based registration is applied followed by incremental fusion of the shape of each frame to the global space, *i.e.*, TSDF in our technique. Final shapes are reconstructed by marching cubes algorithm.

In the experiments, a learning-based technique is evaluated by comparing several real tissues with previous techniques [5], proving the effectiveness of our method. Then, our online shape-registration and merging algorithm is applied to a shape model, *i.e.*, phantom model, of a stomach and a part of a real human body, *i.e.*, inside mouth, to show the successful results of the technique.

2 RELATED WORK

For 3D reconstruction method using endoscopes, techniques using shape from shading (SFS) [6] or binocular stereo [7] have been proposed. However, these techniques often have stringent assumptions on the images that can be processed, or, in the case of binocular stereo, require specialized endoscopes. As an example of active stereo applications in endoscopy, in [8] a single-line laser scanner attached to the head of the scope was used to measure tissue shapes, however, the scope head needed to be directed in parallel to the target, which limited the practical applicability of the technique. Lin *et al.* proposed 3D endoscope system using colored, middle-sized circle dots [9]. Compared to their work, our system uses structured light composed of sharp lines, which can be used for accurate 3D reconstruction using light sectioning triangulation. This is important for obtaining small shape details of the target. Recently, Furukawa *et al.* extended their grid pattern based active stereo system by using DOE (diffractive optical element) with "gap coding" technique solving typical issues

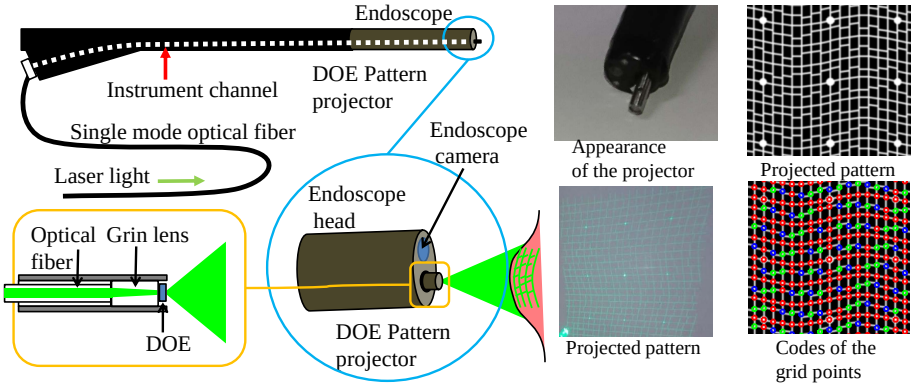

Fig. 1. The system configuration (the left image), the DOE pattern projector (the middle column), and the projected pattern (the right column). The images of the middle column are the appearance of the projector (inserted through the instrument channel) and the pattern illumination projected on a white wall. The images of the right column are the projected pattern and the codewords embedded into the pattern, where S colored in red, L in blue, and R in green. S means edges of the left and the right sides have the same height, L means the left side is higher, and R means the right is higher.

for endoscopic systems [4, 5]. This paper solves practical issues for applying the technique to real bio tissues.

For integrating multiple shapes, registering multiple shapes by ICP algorithm[10] has been a widely-used solution. Similarly, signed distance field (SDF) representation has been widely used for fusing multiple shapes[11]. Recently, KinectFusion[12] integrates those methods so that online shape reconstruction can be realized, where sequentially-captured 3D shapes are incrementally registered and fused into a single model.

3 Overview

3.1 System configuration

A projector-camera system is constructed by inserting a fiber-shaped, micro pattern projector into the instrument channel of a standard endoscope as shown in Fig. 1. For our system, we used a FujiFilm VP-4450HD system coupled with a EG-590WR scope. The DOE-based pattern projector is inserted through the instrument channel of the endoscope the projector slightly protrudes from the endoscope head as shown in Fig. 1 and emits the structured light.

The light source of the projector is a green laser module with a wavelength of 517nm. The laser light is transmitted through a single-mode optical fiber to the head of the DOE projector. In the head, the light is collimated by grin lens, and go through the DOE. The DOE can project a fine, complex pattern at a greater depth range.

In terms of pattern design, we use a grid pattern with gapped lines, whose features are reported to be robust to blurring [4]. The pattern is shown in Fig. 1. The vertical lines of the pattern are all connected and straight, whereas the horizontal line segments are designed so that adjacent line segments have variable vertical gaps at the grid points. With this configuration, a higher-level ternary code emerges from the design with the following three codewords: S (the end-points of both sides have the same height), L (the end-point of the left side is higher), and R (the end-point of the left side is higher). The codes of the pattern of Fig. 1 (right column, top) are shown by color in Fig. 1 (right column, bottom).

Since the vertical lines of the pattern are straight lines, we can apply light sectioning method for 3D triangulation using these lines. By using light sectioning method, we can get accurate 3D points on these lines, which is important for capturing small details of the target surface.

3.2 Algorithm overview

We record sequence of images captured by the endoscope camera, while projecting the structured light shown in Fig. 1. Then, every image the captured sequence is analyzed to obtain shape information of the frame. The reconstructed shapes are 3D curves corresponding to the vertical lines of the grid pattern. Since the 3D curves are sparse, we convert the shape information to frame-wise depth images, then, process the depth images with the KinectFusion algorithm.

The 3D reconstruction of each frame consists of two stages, such as pattern decoding stage and 3D reconstruction stage as shown in Fig. 2. The pattern decoding stage is processed by CNNs, which are trained to extract grid-like structures, and the gap codes in the captured images. In the 3D reconstruction stage, the extracted grid structures and code information are analyzed, and the IDs of all the detected vertical lines are decided, and 3D curves are reconstructed by light-sectioning method.

For training CNNs (learning phase in Fig. 2), actual patterns are projected onto the strong subsurface scattering objects and captured by a camera. Then, correct lines and code IDs are manually given as the ground truth. It is a tough task even for humans, thus, learning data augmentations such as image translations or rotations are used to decrease the burden. Then, parameters and kernels of U-Net [13] are estimated for lines and IDs independently using deep learning framework so that cost functions are minimized. The cost function is basically a difference between an output of U-Net and the ground truth.

In the decoding phase, the captured image is first applied to CNNs for vertical and horizontal line detections. At the same time, the image is also applied to CNN for region-wise classification of local feature codes embedded into the pattern. Then, both results are combined to produce final output, *i.e.*, detected lines with estimated local codes in the pattern. By using the image with detected lines with pattern ID as the input, 3D shapes are recovered in the 3D reconstruction stage. Since a single local code is not sufficient for unique decision of correspondences, information of connectivity and the epipolar constraints

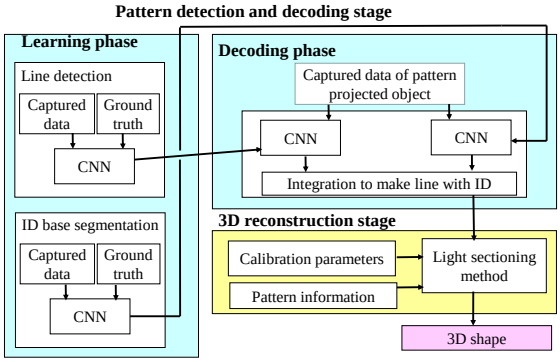

Fig. 2. Overview of CNN-based decoding and 3D reconstruction for oneshot scan. Note that we have two CNNs for vertical and horizontal line detections, and another CNN for decoding IDs of grid points.

are used with a voting scheme to increase robustness, similarly as [14]. Once correspondences of the detected curves are retrieved, 3D shapes are reconstructed by light sectioning method.

Since many of the KinectFusion implementations require depth images, we generate depth images from the sparse 3D curves. Then, the depth images are processed by KinectFusion algorithm. Within the module, the depth images are fused to a volume, where shapes are represented as TSDF (truncated signed distance field). Once all the frames are fused into one volume, the module outputs the fused surface.

4 CNN-based feature detection and decoding for active stereo

A major feature of the projected pattern is a grid-like structure and discrete codes given to each grid point. The grid-like structure is composed of vertical and horizontal line segments. In the pattern, a discrete feature (gap code) is attached to each of the grid point represented by the level gap between the left and right edges of the grid point. The classes of the code are either of S / L / R as shown in Fig. 1 (right column, bottom).

We extract grid-structure and gap-code information using U-Nets [13]. We use U-Nets because this network structure can use global image structures to detect local image features. Because the projected pattern has global structure of grid, we can expect U-Nets use this structure information for detecting local line features to improve performance.

4.1 Detection of grid structures

The training process of a U-Net for detecting vertical lines is as follows. First, image samples of the pattern-illuminated scene is collected. Then, the vertical

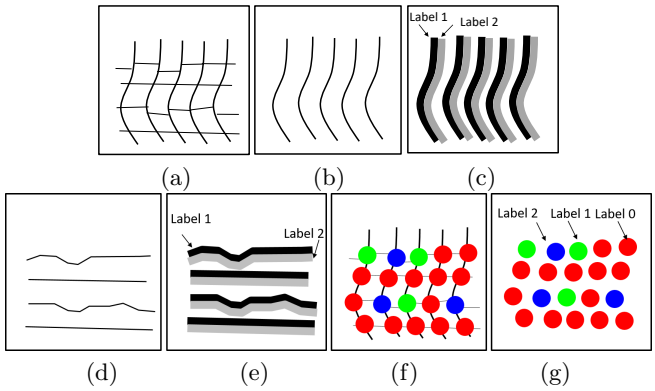

Fig. 3. Training data for U-Nets: (a) An example of captured pattern. (b) Manually annotated vertical line. (c) Manually annotated horizontal line. (d) Training labels for horizontal-line detection. (f) Manually annotated gap codes. (g) Training labels for code detection. In the training data for horizontal-line detection, the discontinuity at the grid points are intentionally connected in the training data. In the training data for code detection, background pixels are treated as "don't care" data for the loss function.

line locations for the image samples are designated manually as curves of 1-dot widths. The 1-dot width curves such as shown in Fig. 3(b) and (d) are too sparse and narrow to be directly used as regions of training data. Thus, regions with 5-dot width of left and right side of the thin curves are extracted, and labeled as 1 and 2, respectively, as shown in Fig. 3 (c) and (e). The rest of the pixels are labeled as 0. These 3 labeled images are used as training data. Then, a U-Net is trained to produce such labeled regions using the loss function of the softmax entropy between the 3-labeled training data and the 3-D feature map produced by the trained U-Net.

By applying the trained U-Net to the image, we can get the 3-labeled image, where left and right side of the vertical curves are labeled as 1 and 2, respectively. Thus, by extracting the 2 horizontally-adjacent pixels where the left is 1 and the right is 2, and connecting those pixels vertically, vertical curve detection is achieved.

The horizontal curve detection is achieved similarly. However, the horizontal edges may be disconnected due to the gaps at the grid points. Even in those cases, training data is provided as continuous curves that go through the center point of the gaps as shown in Fig. 3(e). By optimizing a U-Net using such training data, we can expect results where horizontal curves are detected as continuous at grid points, even if they are actually disconnected by gap codes.

An advantage of using U-Net for line detection of the grid structure is that the U-Net can be implicitly trained to use not only local intensity variation, but also more global information such as repetitive information of grid-like structures. A supporting evidence, that we have experienced is that, if we process an image sample that is scaled so that the training image set does not include the similarly-scaled images, the line-detection performance noticeably worsens.

4.2 Detection of pattern codes

In the proposed method, identification of gap codes is processed by directly applying U-Net to the image signal, not from the line detection results. Thus, the gap code estimation does not depend on line segment detection, which is advantageous for stable detection of gap codes. Note that such a direct method is not easy to implement by conventional image processing.

The training data generation is shown in Fig. 3(bottom row). In the training process, the white background pixels of Fig. 3(bottom row, right column) are treated as "don't care" regions.

The advantage of directly detecting the pattern code is that the stability of the code detection. Since, in the previous work [14], identification of gap codes have been achieved by using results of line detection, failure of line detection or failure of grid-structure analysis consequently leads to code-detection failures. The proposed method is free from such problems of sequential processing.

5 Registration and fusing multiple captured frames

For the KinectFusion implementation, we use Kinfu module of point cloud library[16]. Since this module requires depth images for inputs, we generate depth images from the sparse 3D curves.

To convert the sparse 3D curves into a dense depth image, we use Radial basis function for interpolation of the 3D curves. Radial basis function (RBF) has been a common tool for 3D shape interpolation from point sets [15]. In the case of the proposed system, we only require 2D depth map for the camera viewpoint of the frame, not a general 3D shape; thus, the problem becomes much simpler.

First, the reconstructed 3D curves are stored in 2D maps in camera view. Then, for each 3D point on the curves, a tangent plane is estimated by fitting the neighbor point set (neighbor points are defined by 2D distances on the 2D view) to a 2D plane by 2D linear regression.

Then, the tangent planes of all the curve points are fused using the weights of the radial basis function. In the proposed system we use 2D Gaussian kernel for the RBF. The resulting height function $h(x, y)$ is

$$h(x,y) = \frac{\sum_i k(x-x_i, y-y_i)\{a_i(x-x_i) + b_i(y-y_i) + z_i\}}{\sum_i k(x-x_i, y-y_i)}, \tag{1}$$

where $k(x, y)$ is an RBF kernel defined by $k(x, y) = \exp(-\frac{x^2+y^2}{2\sigma^2})$, (x_i, y_i) is the 2D position of the i-th point in the camera view, z_i is the depth of the i-th point from the camera view, a_i and b_i are the coefficients of the tangent plane fit by the linear regression, σ is a scale parameter of RBF. In our case, we set this value to about average apparent size of the grid in captured images. We calculate the value of (1) for each pixel of the depth image.

Then, the depth images are processed by KinectFusion algorithm. We used Kinfu module of PCL (point cloud library)[16]. The view pose of the depth image

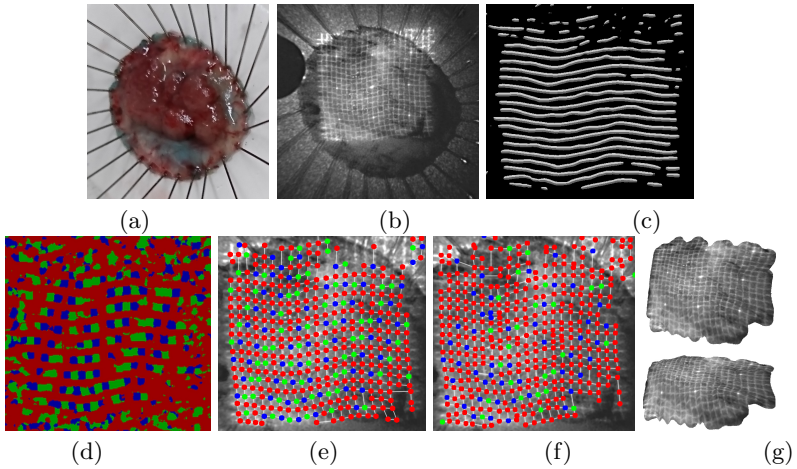

Fig. 4. Grid and code detection results for a specimen of a cancer: (a): The appearance of the sample. (b) The captured image. (c) U-Net output for horizontal-line detection. (d) U-Net output of code detection. (e) Extracted grid-structures and codes. Compare (e) with pattern codes shown in Fig. 1. The counted error rate of (e) was 4.5%. (f) Grid-structures and codes extracted by a previous method[4]. Compare (f) with Fig. 1. The counted error rate of (f) was 18.6%. (g) The reconstructed 3D shape.

is registered with the volume 3D shape represented as TSDF by ICP algorithm using depth error and normal error criterion. Then the depth data is fused into the TSDF. The fused points are extracted after all the frames are processed.

6 Experiment

6.1 Evaluation of CNN based line detection

To show effectiveness of the proposed pattern-feature extraction for endoscope images, we measured specimens of cancers that are resected from patients. The appearance, captured image by the 3D endoscope, outputs of the U-Nets for line detection and code labels are shown in Fig. 4(a)-(d) respectively. The grid structures and codes that are extracted from the U-Net results are shown in Fig. 4(e). For comparison, grid-structures and codes detected by a previous method [4] are shown in Fig. 4(f). The 3D reconstruction results of this sample are shown in Fig. 4(g). Although the captured image (Fig. 4(b)) is low-resolution and includes significant noises, the extracted grid structure (Fig. 4(e)) is stable. By comparing Fig. 4(e) with Fig. 1(right column, bottom), we can confirm that the gap codes extracted by the U-Net is reasonably accurate. The manually counted code-detection error rate of Fig. 4(e) was 4.5%, whereas that of the result of baseline method[4] (Fig. 4(e)) was 18.6%. Using the decoded pattern, the 3D shape of the pattern-projected regions are mostly reconstructed as shown in Fig. 4(f).

Grid and code extraction results for other two specimens are shown in Fig. 5, where (a) and (d) are the captured images, (b) and (e) are the extracted grid and

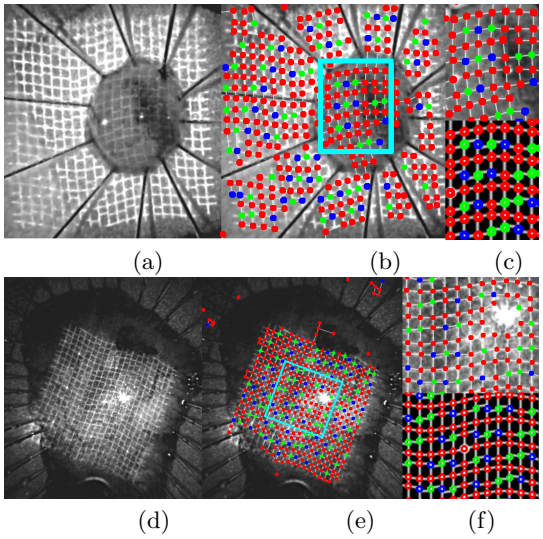

Fig. 5. Grid and code detection results for two specimens of cancers: (a,d) Captured images. (b,e) Extracted grid-structures and codes. (e,f) Magnified regions of (b) and (e), and the corresponding pattern regions.

code structures, and (c) and (f) are the magnified code structures and the corresponding pattern regions. The regions of (c) and (f) are shown as cyan rectangles of in (b) and (e). The specimen of (a-c) was was affected by strong subsurface scattering, however, the extracted codes were reasonably accurate. The image (d) has highly affected by highlights, and the grid structure was missing at the saturated area itself. However, the grids and codes around the saturated area became accurate enough so that the 3D shape can be successfully reconstructed. Those results confirm the stability of our feature-extraction method even if the data condition is low.

6.2 Simultaneous localization and 3D mapping

Then, we apply our online shape registration and merging algorithm to both a phantom model of a stomach and a part of a real human body, *i.e.*, inside a mouth. About calibration, we pre-calibrated the projector-camera system using sphere-based calibration [2].

We first captured shapes of the stomach model for evaluation purpose. Results are shown in Fig. 6. In Fig. 6(a), the area of the recovered shape from a frame of the captured sequence is shown by the red rectangle. Fig. 6(b) is the captured image of the red rectangle where the grid pattern is projected to the surface. In the image, we can observe that grid lines are disconnected by the complicated shape of the surface of the model, however, curves and IDs detected by our method resulting in grids and codes shown in Fig. 6(c). The integrated shape generated by the online registration and merging algorithm is shown in

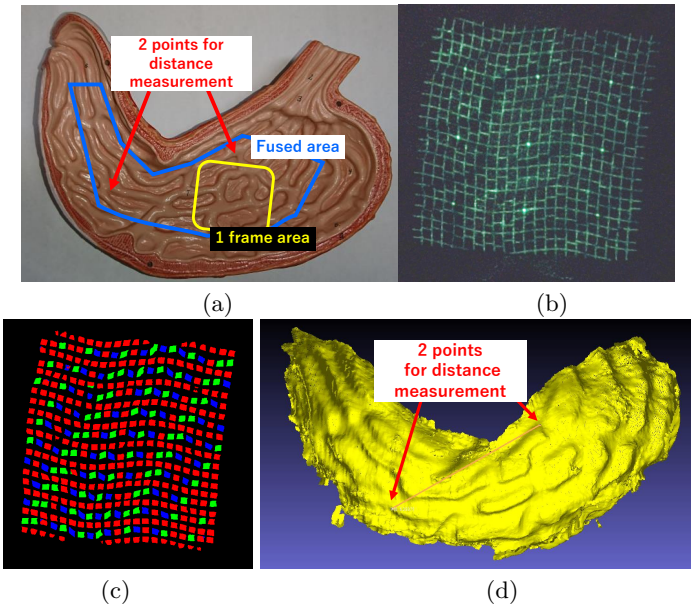

Fig. 6. An example of capturing a phantom model of a stomach: (a)The appearance of the phantom model. (b)A captured image of red rectangle in (a). (c) The CNN result of grid and code detection of (b). Compare (c) with Fig. 1. (d) Fused shape (the region of blue polygon in (a)).

Fig. 6(d). We can confirm that a large area is successfully recovered as well as keeping high-frequency shape details. For quantitative evaluation, we compared distances between corresponding points shown in Fig. 6(a) and (d), as shown in Table 1.

Table 1. Estimated and true distances between points shown in Fig. 6.

	Real size	Estimated size
Distance between two points in Fig. 6	67mm	63mm

Finally, we captured shapes inside a mouth of a human. A captured image, the pattern detection result, the single-frame shape from the shown image, and the final integrated shape are shown in Fig. 7. With this experiments, we can confirm that the grid-structure and codes are robustly detected even with live tissues captured by an ordinary endoscopic system. In addition, a large area is successfully recovered without losing high-frequency shape details, which are clearly observed in Fig. 7(h) where small shape details in the top (a subimage of (a)) is also shown in the right 3D CG shading results (a subimage of (d)).

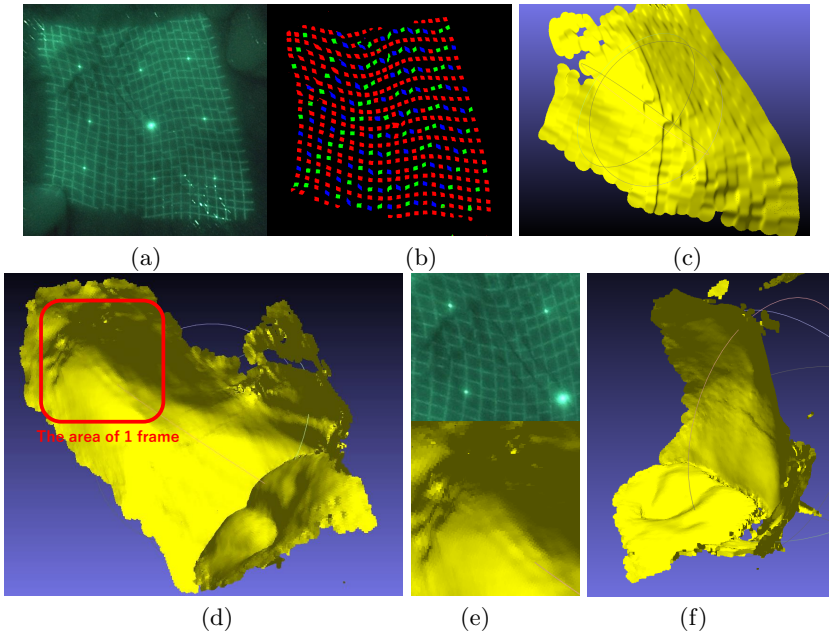

Fig. 7. An example of capturing surfaces inside a mouth: (a) A captured image. (b) The extracted grid-structures and codes of (a). Compare (b) with Fig. 1. (c) The reconstructed shape from (a). (d) The merged shape. (e) Small shape details restored by shape fusion(top: subimage of(a), bottom: subimage of (d)). (f) The merged shape from another viewpoint.

7 Conclusion

This paper proposed a CNN-based grid pattern detection algorithm for active stereo to solve pattern degradation problem caused by subsurface scattering and specularities. Two independent networks, *i.e.*U-Nets, are constructed and trained for both line detection and code based segmentation purposes, respectively. They are integrated to retrieve robust and accurate line detection results with pattern IDs. With our experiments using several target objects with strong subsurface scattering and specular effects, the proposed method shows stable detection of the grid structure and codes that are embedded into the grid points. In addition, 3D shapes of strong subsurface scattering objects are successfully reconstructed, which is only scarcely reconstructed even with the previous technique which is designed to robust to blurring effects. In the future, *in-vivo* experiments for test and real diagnosis purposes are important for real system.

Acknowledgment

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

References

1. Aoki, H., Furukawa, R., Aoyama, M., Hiura, S., Asada, N., Sagawa, R., , Kawasaki, H., Tanaka, S., Yoshida, S., , Sanomura, Y.: Proposal on 3D endoscope by using grid-based active stereo. In: The 35th EMBC. (2013)
2. Furukawa, R., Aoyama, M., Hiura, S., Aoki, H., Kominami, Y., Sanomura, Y., Yoshida, S., Tanaka, S., Sagawa, R., Kawasaki, H.: Calibration of a 3D endoscopic system based on active stereo method for shape measurement of biological tissues and specimen. In: The 36th EMBC. (2014) 4991–4994
3. Furukawa, R., Masutani, R., Miyazaki, D., Baba, M., Hiura, S., Visentini-Scarzanella, M., Morinaga, H., Kawasaki, H., Sagawa, R.: 2-DOF auto-calibration for a 3D endoscope system based on active stereo. In: The 37th EMBC. (Aug 2015) 7937–7941
4. Furukawa, R., Sanomura, Y., Tanaka, S., Yoshida, S., Sagawa, R., Visentini-Scarzanella, M., , Kawasaki, H.: 3D endoscope system using DOE projector. In: The 38th EMBC. (2016) 2091–2094
5. Furukawa, R., Naito, M., Miyazaki, D., Baba, M., Hiura, S., Kawasaki, H.: HDR image synthesis technique for active stereo 3D endoscope system. In: The 39th EMBC. (2017) 1–4
6. Visentini-Scarzanella, M., Stoyanov, D., Yang, G.: Metric depth recovery from monocular images using shape-from-shading and specularities. In: ICIP, Orlando, USA (2012) 25 –28
7. Stoyanov, D., Visentini-Scarzanella, M., Pratt, P., Yang, G.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: MICCAI. (2010) 275–282
8. Grasa, O., Bernal, E., Casado, S., Gil, I., Montiel, J.: Visual slam for handheld monocular endoscope. Medical Imaging, IEEE Transactions on **33**(1) (Jan 2014) 135–146
9. Lin, J., Clancy, N.T., Stoyanov, D., Elson, D.S.: Tissue surface reconstruction aided by local normal information using a self-calibrated endoscopic structured light system. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 405–412
10. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Robotics-DL tentative, International Society for Optics and Photonics (1992) 586–606
11. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, ACM (1996) 303–312
12. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on, IEEE (2011) 127–136
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, Springer (2015) 234–241
14. Furukawa, R., Morinaga, H., Sanomura, Y., Tanaka, S., Yoshida, S., Kawasaki, H.: Shape acquisition and registration for 3D endoscope based on grid pattern projection. In: The 14th ECCV. Volume Part VI. (2016) 399–415
15. Carr, J.C., Fright, W.R., Beatson, R.K.: Surface interpolation with radial basis functions for medical imaging. IEEE transactions on medical imaging **16**(1) (1997) 96–107
16. Rusu, R.B., Cousins, S.: 3d is here: Point cloud library (pcl). In: Robotics and automation (ICRA), 2011 IEEE International Conference on, IEEE (2011) 1–4