Fully Auto-calibrated Active-stereo-based 3D Endoscopic System using Correspondence Estimation with Graph Convolutional Network

Ryo Furukawa[1], Shiro Oka[2], Takahiro Kotachi[2], Yuki Okamoto[2], Shinji Tanaka[2],
Ryusuke Sagawa[3], and Hiroshi Kawasaki[4]

Abstract— **We have developed a series of 3D endoscopic systems where a micro-sized pattern projector is inserted through the instrument channel of the endoscope and shapes are reconstructed by a structured light technique using captured images of the endoscopic camera. One problem of the previous works is that the accuracy of shape reconstruction is low, because the projector cannot be fixed to the endoscope, and thus, the pose of the pattern projector w.r.t. the camera cannot be pre-calibrated. In this paper, we propose a method to auto-calibrate the pose of the projector without using any special devices nor manual process. Since the technique is one-shot, multiple shapes can be reconstructed from an image sequence and a large 3D scene can be recovered by merging them. Experiments are conducted using the real system.**

Clinical relevance— **For endoscopic diagnosis and treatment, accurate size measurement of tumor is important, *i.e.*, size less than 5mm is recommended to be resected with standard protocol. The technique can be contributed to the purpose.**

I. INTRODUCTION

For endoscopic diagnosis and treatment, accurate size measurement of tumor and organs is important and 3D endoscopic system has been researched [2], [3], [5]. To make the systems practical, they are usually based on triangulation, where micro-sized pattern projector is inserted through the endoscope's instrument channel. One problem is that the pattern projector is not fixed to the endoscope's head, and thus, pre-calibration of extrinsic parameters is not possible. Simple solution is to add markers in the pattern and detect them to conduct auto-calibration [3]. However, robust marker extraction cannot be always achieved and it frequently fails to calibrate in real systems. Another possible solution is to use multiple frames to estimate calibration parameters as well as scene shapes simultaneously [5]. However, since it is essentially difficult to retrieve correspondences between a projector and a camera, it requires iterative steps for convergence, resulting in high computational cost. In this paper, we propose a robust auto-calibration method which requires just a single image without necessity of special markers in the pattern. The proposed method is based on graph convolutional network (GCN) to efficiently find correspondences between projector pattern and captured image. Once correspondences are retrieved, auto-calibration can be conducted used them and subsequently 3D points are reconstructed from the correspondences by triangulation. In the experiments, proposed method is compared with previous methods using the real endoscopic systems with phantoms, confirming the effectiveness of our method. It is also demonstrated that the inside shapes of pig's stomach (ex-vivo) are successfully recovered with our method.

II. RELATED WORKS

The structured light technique has been used for practical applications for 3D scanning purposes [9]. For endoscope systems, since endoscope head always moves, the system should be realtime, and typical solution is oneshot scanning techniques [12], [6], [11]. One severe problem for oneshot scan is that they encode positional information into small regions, patterns tend to be complicated and easily affected and degraded by environmental conditions, such as noise, specularity, blur, etc. Recently learning based techniques ares proposed [3].

U-Net [8] is a standard architecture of FCNN (Fully convolutional neural network), which can receive an image and produces a labeled image. Song *et al.* [10] proposed to decode active stereo pattern using a CNN. They use conventional methods for grid detection, and a CNN is used for classifying specifically designed 256 characters embedded into the grid pattern. We also use U-Net for pattern detection in our method.

Recently, DNN based approach to efficiently find correspondences are proposed and GCN is most recent solution for the purpose [1]. In the paper we also use a full advantage of GCN to find correspondences between projected pattern and captured image.

III. SYSTEM OVERVIEW

For this study, a projector-camera system was constructed by inserting a fiber-shaped, micro pattern projector into the instrument channel of a standard endoscope. We used a Fujifilm EG-590WR endoscope and a pattern projector with a diffractive optical element (DOE) to generate structured-light illumination. The pattern projector can be inserted into the endoscope's instrument channel and patterns are projected from the projector to surfaces in front of the head of the endoscope as shown in Fig. 1. As shown in Fig. 1(b), we used a grid pattern that is robust against subsurface scattering [4]. All vertical edges are connected; horizontal edges have small gaps, representing code symbols S, L and R as shown in Fig. 1(c), where red dots mean that the right and the left edges of the grid point have the same height (code S) blue

*Research supported by 20H00611, 18K19824, 18H04119, 16KK0151 in Japan, and MSRA CORE18.

[1] Hiroshima City University, Japan [2] Hiroshima University Hospital, Japan [3] National Institute of Advanced Industrial Science and Technology, Japan [4] Kyushu University, Japan

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

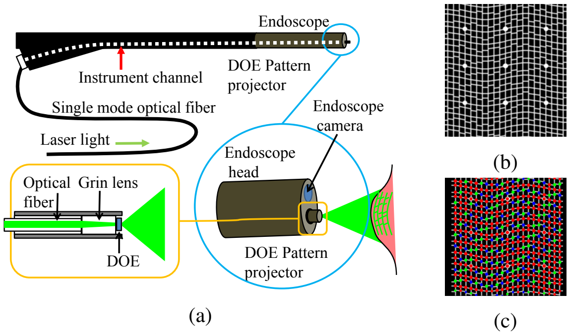

Fig. 1. System configuration and projected pattern: (a) System configuration; (b) Projected pattern with 9 bright markers and gap coding; (c) Codes embedded as gaps at grid points of the projected pattern.

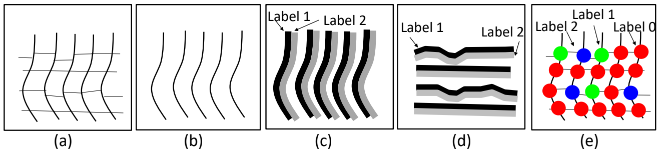

Fig. 2. Training data for U-Nets: (a) Captured image; (b) Manually annotated vertical lines; (c) Labels for training vertical-line detection; (d) Labels for horizontal-line detection; (e) Labels for for code detection.

means the left side is higher (code L), and green means the right is higher (code R). If the system is completely calibrated, by using the connectivity between nodes as well as code at each node, correspondences are efficiently found by 1D search along Epipolar line. Once they are retrieved, 3D points are recovered by triangulation.

IV. FEATURE DETECTION FROM PATTERN-ILLUMINATED IMAGES

As described previously, the projected pattern is a grid structure with code symbols $\{S, L, R\}$ associated with the grid points. We extract the grid-structure and gap-code information using U-Nets [8]. U-Nets uses global image structural context information to detect local features. Since the projected pattern has global grid structures, it can be expected that U-Nets will use such "global" structure to detect "local" line features.

The process of training a U-Net to detect vertical lines is as follows. First, sample images of the pattern-illuminated scene are collected. Then, vertical line locations for the samples are drawn manually as single pixel width curves. Since a single pixel width curve shown in Fig. 2(b) is too narrow to be directly used as training label regions, two new regions with five pixel width are generated on the both sides of the single pixel curves (Fig. 2(c),(d)) By applying the trained U-Net to the endoscope images, curves are detected by extracting the borders between the regions of left and right.

In the paper, to increase the robustness, we also add 9 bright markers into the projected pattern as shown Fig. 1(b) white dots. To utilize these markers in our method, we train CNNs to to classify each of the markers into 5 classes (up to rotational symmetry, 9 markers can be classified into 5 classes). By applying the trained CNNs, every pixel of the captured frames is classified into 6 classes (5 plus one for non-marker).

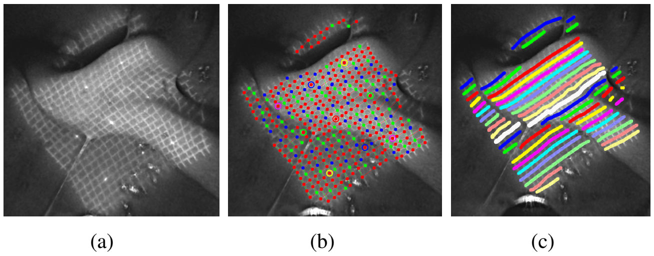

Fig. 3. Training data for GCN: (a)Sample image for GCN training; (b)Grid detection result; (c)Annotation image for correspondence.

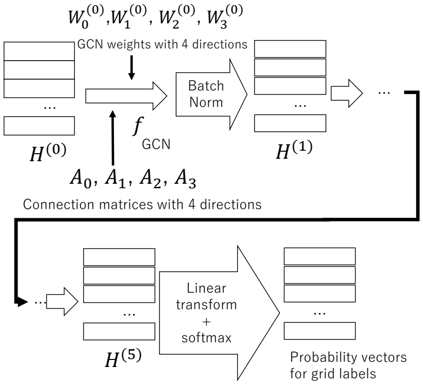
$W_0^{(0)}, W_1^{(0)}, W_2^{(0)}, W_3^{(0)}$
GCN weights with 4 directions
Batch Norm
f
GCN
$H^{(0)}$
$H^{(1)}$
A_0, A_1, A_2, A_3
Connection matrices with 4 directions
Linear transform + softmax
$H^{(5)}$
Probability vectors for grid labels

Fig. 4. Network architecture of GCN for correspondence estimation. f is the GCN and an activation function.

In the output image from U-Net, the lines of the grid can be extracted as the boundaries between the two regions. By performing 8-neighbor labeling process on extracted boundary curves, different labels are assigned for each curve. Then, intersections between vertical and horizontal curves are extracted as grid points. By sorting the set of grid points on one vertical curve by y coordinates, the adjacency relationships between these grid points along the vertical curve is determined. By applying this process for all both vertical and horizontal curves, the grid structure of the extracted curves can be represented as a graph.

For each grid point, a feature vector is assigned using outputs of U-Net, such as 2D coordinates on the image, estimated code and estimated marker class. Fig. 3(a),(b) show an example of an image and a graph extracted from the image. This attribute is embedded in feature vectors of 2 + 4 + 6 dimensions (2D coordinates, 4D code classes (3 types of codes + unknown code), 6D marker classes (5 types of markers and non-markers)).

V. CORRESPONDENCE ESTIMATION USING GCN

For 3D measurement, the detected grid points should be assigned to the grid points of the projected pattern shown in Fig. 1(b) as known as correspondence problem for stereo. In this study, we propose a method to estimate this assignment using Graph Convolutional Network (GCN) [1].

A graph extracted as Sec. IV typically has the following characteristics. (1) It is a grid graph with several extra edges and nodes by some errors. (2) For each grid point, the positions, the code information, and the marker information as well as its connected grid point's IDs are given as attributes. (3) The extra edges and nodes are generated by erroneously detected grid points or an incorrectly connected edges. Similarly many graph structures might be missing due to occlusions or serious noises. Also, the code and marker information of the grid points include errors. (4) Although the code is a significant clue, it is difficult to determine the correspondence only from codes, because the code pattern has repetitions as shown in Fig. 1(c).

In [4], the correspondences were estimated using both code information and epipolar constraints. However, if the extrinsic parameters of the projector are unknown or have large errors, the epipolar constraint cannot be used. In this research, we propose to solve this problem by applying GCN on the grid graph without using the epipolar constraint. Note that there are alternative methods to GCN, such as MRF(Markov random field) approaches as belief propagation, or converting the grid graph to 2D image-like data applying 2D CNNs. However, we adopted GCN over these approaches due to the following reasons. (1) In MRF approach, it is necessary to design a cost function to be optimized, which is not easy in this problem. (2) Due to defects of grid structures of the graph, and it is often difficult to convert it to 2D image-like data. Whereas, A GCN can be applied to the observed graph itself without conversion to 2D data. (3) A GCN can use information of both the grid-point properties of each grid point as well as the adjacent grid points simultaneously without manually designing a cost function.

The layer operation of a GCN is applied to a data matrix $H^{(l)}$, where l-th layer feature vectors of all the grid points are stacked to one matrix, and produces $H^{(l+1)}$, which is $(l+1)$-th layer data matrix. It can be represented as

$$H^{(l+1)} = f(H^{(l)}, A, W^{(l)}) = \sigma(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(l)}W^{(l)}),$$

where $\hat{A} = A + I$ is the adjacency matrix of graph G with added self-connections, I is the identity matrix, $\hat{D}$ is the degree matrix of $\hat{A}$, $W^{(l)}$ is the Weight matrix of this layer, and σ is an activation function (LeLU).

In [1], the network is an undirected graph, but in the case of this research, the information of a grid point can be obtained from adjacent nodes of 4 directions of top, bottom, right and left, and information from different directions has different meanings. Thus, for treating the 4 directions differently, we set A_0, A_1, A_2, A_3 as the adjacency matrix of the directed graph that includes only the connections of top, bottom, right, and left directions, respectively. Calculation of $H^{(l+1)}$ is performed by the following formula.

$$H^{(l+1)} = \sum_{d=0}^{3} f(H^{(l)}, A_d, W_d^{(l)}), \quad (1)$$

where $W_d^{(l)}$ is the weight matrix of layer l specific to direction $d \in \{0, 1, 2, 3\}$.

After repeating Eq. (1) and batch normalization by 5 times, two 22-class probability vectors are calculated from the latent variable vectors $H^{(5)}$ by the fully connected linear transformation, and the softmax function(Fig. 4). The classification into 22 classes is because the original pattern has 21 types of vertical and horizontal lines, respectively, and the ”unknown” class has been added for each of the vertical and horizontal lines. By applying the GCN to the feature vectors and the adjacency matrix of the graph, classes of vertical and horizontal lines are estimated for the grid points.

Training of the GCN is done by supervised manner. First, a graph structure is extracted from an actual endoscopic image using U-Net. Separately, for the same endoscopic image, the class IDs of both vertical and horizontal lines are manually annotated as shown in Fig. 3 (c), which are used as a teacher data.

VI. AUTO-CALIBRATION OF PROJECTOR

In this research, it is assumed that the intrinsic parameters of the endoscope camera are known, since there are many tools for camera calibration. On the other hand, projector calibration is not common and still challenging especially for a static pattern. In our method, three intrinsic parameters, such as focal length, two principal points for xy, and the six extrinsic parameters of the projector will be auto-calibrated. In terms of intrinsic parameters, once they are calibrated, they never change, and thus, it is necessary just once before operation. In terms of extrinsic parameters, since the pattern projector is inserted through the instrument channel, the projector pose may change all the time and should be calibrated at each frame. In addition, since the motion can be mostly explained by two degrees of freedom, such as one-dimensional translation along the channel and rotation about the center line of the channel [2], the rest 4-DOF changes other than the above two DOF from a 6DOF of rigid transformation. Such restriction also allows the scale to be determined, which is normally impossible for auto-calibration, and leads to stabilization of the auto-calibration.

In the common calibration method, the correspondence problem is usually solved by using the epipolar constraints. In the method, even if the projector is uncalibrated, since correspondences are acquired by the aforementioned method, auto-calibration of the projector can be performed. Since the corresponding points determined by GCN include outliers, the Random Sample Consensus (RANSAC) algorithm is used to efficiently exclude the outliers. In this research, we estimate the projector parameters using 10 corresponding points, and if the number of inliers that match the estimation result is more than a threshold, we re-optimize the parameters using the all inliners and estimate the error. This is repeated, and the parameter set with the minimum error becomes the result. Once the parameters of projector is estimated as above, outliers can be determined by checking the epipolar error.

After the first auto-calibration is performed, if the parameters of the projector pose due to the operation of the endoscope, and the error of the epipolar constraint exceeds a threshold, the pose parameters are re-calibrated by minimizing the epipolar errors. Since the outliers of the correspondences can be determined by epipolar constraint, it can be updated without using RANSAC. For each pair of corresponding points, an error of epipolar constraint is

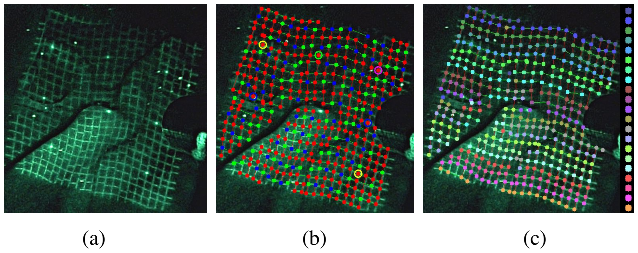

Fig. 5. Correspondence labeling result; (a) Source image; (b) Extracted grid graph with codes and marker estimations; (c) Labeling result of vertical positions (colors can be compared with the scale at the right-side border).

calculated with respect to the above nine parameters, where the cost is the squared distance between the epipolar line of the projector point drawn to the camera image and the corresponding camera point. Since the GCN does not use epipolar constraint, correspondences from the GCN that is epipolar constraint inlier are considered to be correct with strong certainty. By triangulation of these corresponding points, three-dimensional point information can be obtained.

VII. Results

Fig. 5 shows an example of the correspondence estimation results of the proposed method. The source images are surfaces inside a pig's stomach (ex-vivo). We confirmed that the obtained correspondences were mostly correct, in spite of occlusions shape discontinuities, and large specularities.

Next, we measured a surface of a stomach phantom model (Fig. 6 (a)). The shape of the phantom is also measured by gray-code projection as a ground-truth shape. We conducted a pre-calibration of the projector-camera parameters using a sphere using method of [2]. Next, we bended the endoscope so that the fiber-shaped projector moves inside the instrument channel. Then, we conducted the auto-calibration method and measured a surface on the phantom. To compare the result with the ground-truth shape, we captured multiple frames of shape by moving the endoscopic head, scanning a small region on the phantom shown in Fig. 6 (a). We fused the multiple shapes using KinectFusion[7] and compared the fused shape with the ground truth shape. Fig. 6 (b) is an orthographic view with the ground truth shape, and we confirmed that the shape and the size of the fused shape was almost the same as the ground truth shape. On the contrary, without either initial auto-calibration using RANSAC nor calibration updating, did not output a proper shape (Fig. 6 (c)). It is because, while scanning around the region, the projector was rotating inside the instrument channel.

We also conducted a simple validation of using all the correspondence points utilizing GCN-based estimation, instead of just using markers. In the case of Fig. 6, auto-calibration was also possible by only using marker positions. However, by adding random outliers by 30% onto the correspondence data, the RANSAC calibration failed, since the number of sample pairs was too small (60 pairs for 10 frames).

VIII. Conclusion

In the paper, we propose an efficient correspondence search algorithm based on GCN, which allows auto-calibration without any manual process nor calibration tools. By using the real endoscopic system where the projector is not fixed to the head of it, we successfully reconstructed 3D surfaces inside a pig's stomach from a sequence of images of over 200 frames. We also demonstrated to acquire an unified and wide shape inside the stomach by merging the reconstructed shapes.

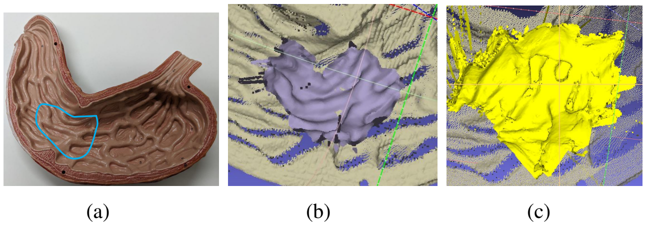

Fig. 6. 3D measurement of a phantom: (a)The phantom shape and the scanned region; (b) An orthographic view of the fused shape on the ground-truth shape as a background; (c) Output shape of KinectFusion where calibration parameters were not updated.

References

[1] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

[2] 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 *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7937–7941. IEEE, 2015.

[3] R. Furukawa, M. Mizomori, S. Hiura, S. Oka, S. Tanaka, and H. Kawasaki. Wide-area shape reconstruction by 3d endoscopic system based on cnn decoding, shape registration and fusion. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 139–150. Springer, 2018.

[4] 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 *European Conference on Computer Vision*, pages 399–415. Springer, 2016.

[5] R. Furukawa, G. Nagamatsu, S. Oka, T. Kotachi, Y. Okamoto, S. Tanaka, and H. Kawasaki. Simultaneous shape and camera-projector parameter estimation for 3d endoscopic system using cnn-based grid-oneshot scan. *Healthcare Technology Letters*, 6(6):249–254, 2019.

[6] H. Kawasaki, R. Furukawa, R. Sagawa, and Y. Yagi. Dynamic scene shape reconstruction using a single structured light pattern. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

[7] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011.

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

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

[10] 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. IEEE, 2017.

[11] A. O. Ulusoy, F. Calakli, and G. Taubin. One-shot scanning using de bruijn spaced grids. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 1786–1792. IEEE, 2009.

[12] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 19:4–12, April 2012.