Dense 3D Reconstruction of Endoscopic Polyp

Ankur Deka[1], Yuji Iwahori[2], M. K. Bhuyan[1], Pradipta Sasmal[1] and Kunio Kasugai[3]
[1]*Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India*
[2]*Department of Computer Science, Chubu University, Aichi, Japan*
[3]*Department of Gastroenterology, Aichi Medical University, Aichi, Japan*

Keywords: MIS, SfM, SLAM, Specularity, Malignant, ORB.

Abstract: This paper proposes a model for 3D reconstruction of polyp in endoscopic scene. 3D shape of polyp enables better understanding of the medical condition and can help predict abnormalities like cancer. While there has been significant progress in monocular shape recovery, the same hasn't been the case with endoscopic images due to challenges like specular regions. We take advantage of the advances in shape recovery and suitably apply these with modifications to the scenario of endoscopic images. The model operates on 2 nearby video frames. ORB features are detected and tracked for computing camera motion and initial rough depth estimation. This is followed by a dense pixelwise operation which gives a dense depth map of the scene. Our method shows positive results and strong correspondence with the ground truth.

1 INTRODUCTION

Endoscopy is a Minimally Invasive Surgery (MIS) for examining and operating on the medical condition. It benefits patients through smaller trauma, shorter hospitalization, lesser pain and lower risk of infection than traditional open cavity surgeries. A device called endoscope is inserted into the body through a natural orifice.

In colonoscopy, the colon and the large intestine is examined. One of the major benefits is the detection of malignant (cancerous) polyp in endoscopy through properties such as shape, texture and size of the polyp. Shape cannot be judged directly from 2D images of a monocular endoscope. Specialized endoscopes with a laser light beam head (Nakatani et al., 2007; Hayashibe et al., 2005) or with two cameras mounted on the head for stereo vision (Chang et al., 2014; Stoyanov et al., 2010; Mourgues et al., 2001) are available. However, the sizes of such endoscopes are large. A 3D scanner is developed by Schmalz et al. (2012). Here, we consider a general purpose endoscope, of the sort still most widely used in medical practice. Shape from shading approach using single monocular endoscope has been explored (Iwahori et al., 1990; Wang et al., 2009). Wu et al., (2010) used Multi-view Shape from Shading.

We explore the possibility of using multiple images or video for shape recovery. Shape recovery from multiple images constitute the SfM (Structure from Motion) or SLAM (Simultaneous Localization And Mapping) problem. Both the scene's structure and the camera's ego-motion are unknown, and the challenge is to simultaneous solve them. EKF (Extended Kalman Filter) based Monocular SLAM is used by Grace et al. (2009). Mahmoud et al. (2016) used ORB SLAM. Grace et al. and Mahmoud et al. give real time performance but produce sparse or semi-dense reconstruction which may not be sufficient for the medical practitioner to understand the medical condition.

There has been significant progress in monocular SLAM in terms of both camera tracking and shape recovery. Newcombe et al. (2011) made a dense reconstruction of the scene (non-medical) from an input video. We take some ideas from this paper for our approach.

We propose a simple method for dense 3D shape reconstruction. We use feature based method for tracking the camera and creating an initial sparse depth map. We then use the estimated camera motion to obtain a dense depth map of the scene by minimizing a cost function. Section 3 discusses the proposed method in detail. In many cases, validating a 3D reconstruction algorithm for endoscopy images / video is difficult as ground truth data is not available. We use the Tsukuba dataset for validating

Deka, A., Iwahori, Y., Bhuyan, M., Sasmal, P. and Kasugai, K.
Dense 3D Reconstruction of Endoscopic Polyp.
DOI: 10.5220/0006720701590166
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 2: BIOIMAGING*, pages 159-166
ISBN: 978-989-758-278-3

the algorithm (Peris et al., 2012; Martull et al., 2012). Final test is performed on real endoscope images that we obtained from Aichi Medical University, Japan.

2 NOTATIONS

1) T: Rigid Body Transformation matrix in homogeneous coordinate
T =

$$\begin{bmatrix} \cos\theta_1\cos\theta_2 & cos\theta_1\, sin\theta_2\, sin\theta_3 - sin\theta_1\, cos\theta_3 & cos\theta_1\, sin\theta_2\, cos\theta_3 + sin\theta_1\, sin\theta_3 & c_0 \\ sin\theta_1\, cos\theta_2 & \sin\theta_1\, sin\theta_2\, sin\theta_3 + cos\theta_1\, cos\theta_3 & sin\theta_1\, sin\theta_2\, cos\theta_3 - cos\theta_1\, sin\theta_3 & c_1 \\ -sin\theta_2 & cos\theta_2\, sin\theta_3 & cos\theta_2\, cos\theta_3 & c_2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Where θ_1, θ_2 and θ_3 are the translation angles along the 3 axes and c_0, c_1 and c_2 and the translation values along the 3 axes.

Multiplication by T converts from converts from coordinate system of one camera frame to another.

2) $\Omega \subset R^2$: image domain
Any image point $\mathbf{u}=(u,v)^T \in \Omega$

3) $\mathbf{I}: \Omega \rightarrow R^3$
RGB value at pixel point

4) $\xi(\mathbf{u}): \Omega \rightarrow R$
Inverse Depth Map: The range of values the reciprocal of depth can take.
d= $\xi(\mathbf{u})$ gives 1/depth

5) K: Camera intrinsic matrix

$$K = \begin{bmatrix} f_x & s & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} K_{11} & K_{21} & K_{31} \\ K_{21} & K_{22} & K_{32} \\ K_{31} & K_{23} & K_{33} \end{bmatrix}$$

f_x, f_y are focal lengths of the camera in the x and y axes respectively. We use K_{ij}'s to refer to the terms in the camera intrinsic matrix. All elements in these matrices are in units of pixels.

6) $\Pi^{-1}(\mathbf{u}, \mathrm{d}) = \left(\frac{1}{\mathrm{d}}\right) * \mathrm{K}^{-1} * \mathring{u}$
where $\mathring{u}=(u,v,1)^T$

7) $\Pi(\mathbf{x})$: Dehomogenization function
$\Pi(\mathbf{x}) = (x/z, y/z)$

8) $K^* = \begin{bmatrix} & & 0 \\ & K & 0 \\ & & 0 \end{bmatrix} = \begin{bmatrix} f_x & s & x_0 & 0 \\ 0 & f_y & y_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

9) $K^{-1} = \begin{bmatrix} K_{11}^{-1} & K_{21}^{-1} & K_{31}^{-1} \\ K_{21}^{-1} & K_{22}^{-1} & K_{32}^{-1} \\ K_{31}^{-1} & K_{23}^{-1} & K_{33}^{-1} \end{bmatrix}$

1) $K^{-1*} = \begin{bmatrix} & & \\ & K^{-1} & \\ & & \\ 0 & 0 & d \end{bmatrix}$

2) ¥ = $[c_0 \quad c_1 \quad c_2 \quad d_1 \quad d_2 \quad . \quad . \quad . \quad d_n]^T$

Where c_0, c_1 and c_2 denote the translations in the 3 axes respectively and d1 to dn denote the inverse depth values at each of the n feature points in the view of image 1.

A point $\mathbf{u_1}$ in image 1 (Im_1) corresponds to a point $\mathbf{u_2}$ in image 2 (Im_2), which can be found as:
$\mathbf{u_2} = \Pi\ (KT_{21}\Pi^{-1}(\mathbf{u_1}, \mathrm{d}))$

We can represent this operation using only matrix multiplications in homogeneous coordinates:

$$\begin{bmatrix} u_1 * w \\ v_1 * w \\ w \end{bmatrix} = (K^*) \times (T_{21}) \times (1/d) \times (K^{-1}*) \times \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix}$$

$$= (1/d) \times (K^*) \times (T_{21}) \times (K^{-1}*) \times \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix}$$

$$= (1/d) \times \Upsilon$$

Where $\Upsilon = (K^*) \times (T_{21}) \times (K^{-1}*) \times \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix}$

$$= \begin{bmatrix} \Upsilon_0 \\ \Upsilon_1 \\ \Upsilon_2 \end{bmatrix}$$

3 PROPOSED METHOD

The outline of the algorithm is shown in Fig.1. Two nearby frames from a video are taken as input. These two images are of the same scene with a slight movement of the camera (3DOF translation). Specularity is removed using method of [17]. The endoscope camera is calibrated using a 3rd party software.

The algorithm can be divided into 2 parts. These two parts deal with the feature points and remaining points respectively. The idea is that a few number of number of good and distinct feature points are generally available in endoscope scene. So we use those to track the camera and obtain depth at those points. For the remaining points we use the obtained camera tracking information to obtain depth.

3.1 Feature Points

Feature points are detected and matched in the 2 images. We randomly initialize the camera transformation matrix. Thereafter, we keep changing the transformation matrix till the pairs of corresponding feature points are correctly mapped

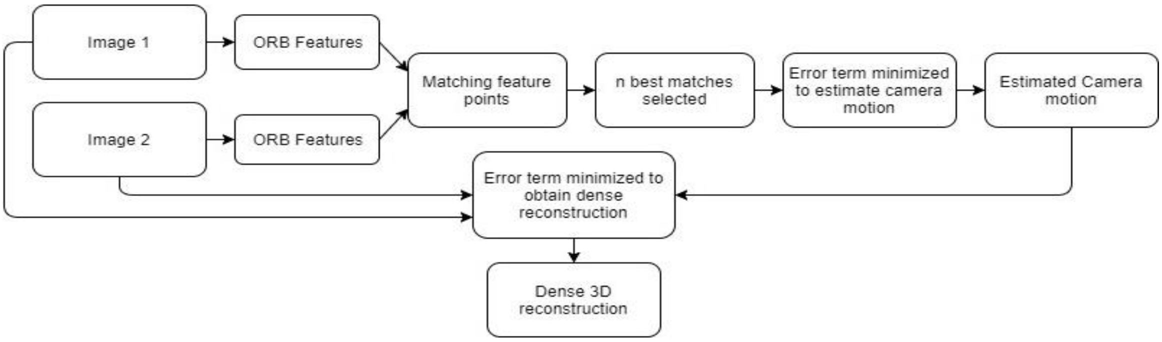

Figure 1: Block diagram of the proposed approach.

from one image to the other. The value of camera transformation when the mapping aligns with direct matching of feature points is supposed to be the correct camera transformation.

a) Mask is applied to the specular region and ORB (Oriented FAST and rotated BRIEF) Features are extracted from both the images (Im1 and Im2) in regions where specularity is absent. n best matching pairs of points are taken (Fig.1).

b) ¥ is initialized to:

$$[0 \quad 0 \quad 0 \quad d0 \quad d0 \quad . \; . \; . \quad d0]^{\mathrm{T}} \tag{1}$$

(¥ is (n+3)x1 dimensional)
That is, we initialized with no translation between the two images and uniform inverse depth of d at all n feature points.

c) At every feature point u1 in Im1, we can compute u2 in Im2 using (1). We also have u2_true, which is obtained by matching ORB feature points.

Thus we can define IE (Individual Error) at each (i$^{\mathrm{th}}$) point as:

IE(i) = ½(|u2-u2_true|2)2

Total Error (TE) is defined as:
TE = $\sum_{All\ Feature\ Points}$ IE

We wish to minimize TE by varying the entries in ¥.
d) We use gradient descent for m steps to minimize ¥.

e) The iteration step of gradient descent is:

$$\text{¥n+1} = \text{¥n} - \left\langle \mu \middle| \frac{\partial(TE)}{\partial(\text{¥})} \right\rangle \tag{2}$$

Where, µ is vector storing the gradient descent rates for each of the entries in ¥.
⟨ | ⟩ denotes dot product operation.

$$\text{f) } \frac{\partial(TE)}{\partial(\text{¥})} = \sum_{i=1}^{n} \frac{\partial(IE(i))}{\partial(\text{¥})}$$

$$\text{g) } \frac{\partial(TE)}{\partial(\mathbf{c_i})} = \sum_{i=1}^{n} \frac{\partial(IE(i))}{\partial(\mathbf{ci})} \tag{3}$$

$$= \sum_{i=1}^{n} \left(\mathrm{u_2(i)} - \mathrm{u_2true(i)}\right) \mathrm{x} \; \frac{\partial(u_2(i))}{\partial(\mathrm{c_i})} + (\mathrm{v_2(i)} - \mathrm{v_2true(i)}) \; \mathrm{x} \; \frac{\partial(v_2(i))}{\partial(\mathrm{c_i})} \tag{4}$$

Where:

$$\frac{\partial(u_2(i))}{\partial(\mathrm{c_i})} = \frac{\Upsilon_2 K_{1i} - \Upsilon_0 K_{3i}}{\Upsilon_2^2} d \tag{5}$$

$$\frac{\partial(v_2(i))}{\partial(\mathrm{c_i})} = \frac{\Upsilon_2 K_{2i} - \Upsilon_0 K_{3i}}{\Upsilon_2^2} d \tag{6}$$

$$\text{h) } \frac{\partial(TE)}{\partial(\mathrm{d_i})} = (\mathrm{u_2(i)} - \mathrm{u_2true(i)}) \; \mathrm{x} \; \frac{\partial(u2(i))}{\partial(\mathrm{d_i})} + (\mathrm{v_2(i)} - \mathrm{v_2true(i)}) \; \mathrm{x} \; \frac{\partial(v_2(i))}{\partial(\mathrm{d_i})} \tag{7}$$

Where:

$$\frac{\partial(u_2(i))}{\partial(\mathrm{di})} = \frac{\Upsilon_2(K_{11}c_0 + K_{12}c_1 + K_{13}c_2) - \Upsilon_2(K_{31}c_0 + K_{32}c_1 + K_{33}c_2)}{\Upsilon_2^2}$$
$$\frac{\partial(v_2(i))}{\partial(\mathrm{di})} = \frac{\Upsilon_2(K_{21}c_0 + K_{22}c_1 + K_{23}c_2) - \Upsilon_2(K_{31}c_0 + K_{32}c_1 + K_{33}c_2)}{\Upsilon_2^2} \tag{8}$$

$$\text{i) } \frac{\partial(TE)}{\partial(\text{¥})} = \left[\frac{\partial(TE)}{\partial(\mathrm{c_0})} \quad \frac{\partial(TE)}{\partial(\mathrm{c_1})} \quad \frac{\partial(TE)}{\partial(\mathrm{c_2})} \quad \frac{\partial(TE)}{\partial(\mathrm{d_1})} \quad \frac{\partial(TE)}{\partial(\mathrm{d_2})} \quad . \quad . \quad \frac{\partial(TE)}{\partial(\mathrm{d_n})}\right]^T \tag{9}$$

j) We use µ of dimension (n+3)x1 as:

$$\mu = [\mu_T \quad \mu_T \quad \mu_T \quad \mu_d \quad \mu_d \quad . \; . \; . \quad \mu_d]^T \tag{10}$$

There is several order of magnitude difference in gradient descent rates: µT and µd for c$_i$'s and d$_i$'s respectively. This is because of c$_i$'s are roughly near to 0 (Camera is not moved much in two nearby frames). However, initial values of d$_i$'s are assigned randomly to a constant d$_0$ and they can vary a lot from it.

3.2 Remaining Points

Photometric error at every pixel $\mathbf{u_1}$ in Im$_1$ is defined as:

ρ(Im$_{1,\mathbf{u1}}$,d) = | Im$_1$($\mathbf{u_1}$)-Im$_2$(Π(KTmrΠ^{-1}($\mathbf{u_1}$,d))) |
An exhaustive search is performed between d$_{min}$ and d$_{max}$ for every pixel and d$_{optimal}$ is chosen as the d that given minimum error.

$$d_{optimal} = argmin\_(\mathrm{d})\ \rho(Im_1, u, d)$$

d_{min} is chosen as smaller than all d_i's among the feature points. Similarly, d_{max} is chosen as larger than all d_i's among the feature points.

DTAM (Newcombe et al., 2011) uses pixel intensity values for both tracking and reconstruction. We chose to use feature points for tracking as pixel intensity values can give ambiguous tracking results in case of our target medical images where several similar intensity pixels can be found in the neighbourhood of each pixel.

4 PARAMETERS

We choose the parameters as follows:

n (number of feature points used)	= 30
d_0 (initial inverse depth value, Eq.1)	= 50
μ_T (gradient descent rate for translation, Eq.10)	= $9\text{x}10^{-14}$
μ_d (gradient descent rate for inverse depth, Eq.10)	= $9\text{x}10^{1}$
m (number of gradient descent steps)	= 100

The number of feature points n is considered low as in endoscopic image very less number of good distinct feature points can be found. A large number of feature points would lead to wrong matches and result in error in camera motion estimation. µT and µd have an order of 10^{15} difference. This is because translation value need not change much but the inverse depth value is initialized randomly. The number of gradient steps is considered sufficiently high at 100.

5 EVALUATION

We use Tsukuba dataset's [14, [15] stereo image pair to validate our approach. Though the images are taken from a stereo camera (1 DOF translation), our approach is designed to handle 3DOF translation.

The images taken are from stereo vision dataset. Thus we can verify results with Stereo vision formula for this particular dataset.

Depth = (Bxf)/Disparity

Where,
Baseline is the distance between optical centres of the two cameras in stereo vision. Disparity is the distance between the pixels of the same point in the two images.

Depth x Disparity = Bxf
= constant

In our case,

Depth = 1/d
∴ Disparity/d = constant

We used disparity values from Tsukuba Ground Truth and inverse depth from the implementation.

Disparity/d for the n (30) feature points are computed. The normalised standard deviation (ratio of standard deviation to mean) of the data is found to be 0.043<<1 implying the data is almost constant. Thus, the results are consistent with the stereo vision formula.

Matched feature points are shown in Fig.2. Linear interpolation of the obtained inverse depth map is shown Fig.3. The final dense inverse depth map is shown in Fig.4. Fig.5 shows the final inverse depth map (with median filter) alongside the ground truth disparity. The black strip on the left part of the reconstruction is because of the absence of the corresponding points in the 2nd image.

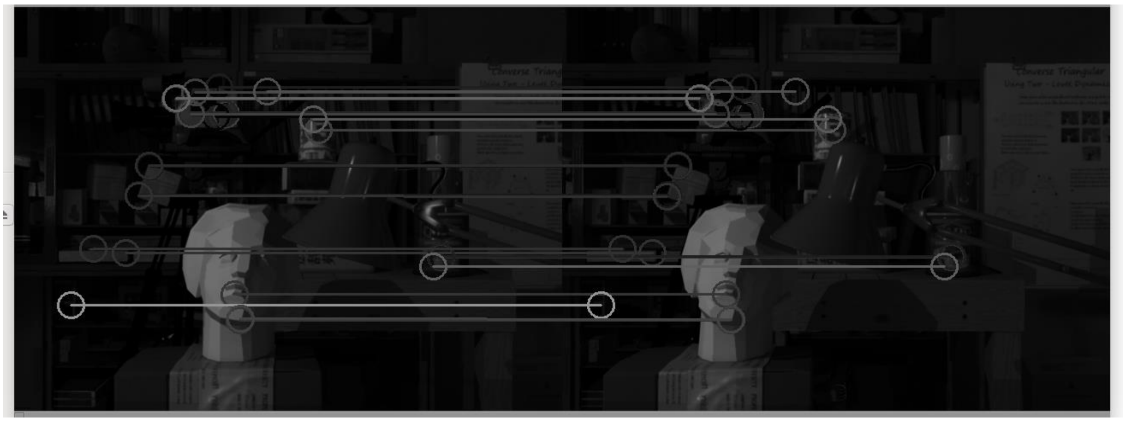

Figure 2: ORB features matched in the 2 input images.

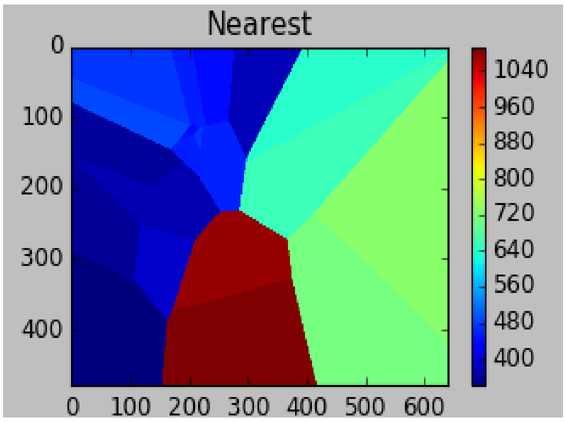

Figure 3: Nearest neighbour interpolation of inverse depth at feature points.

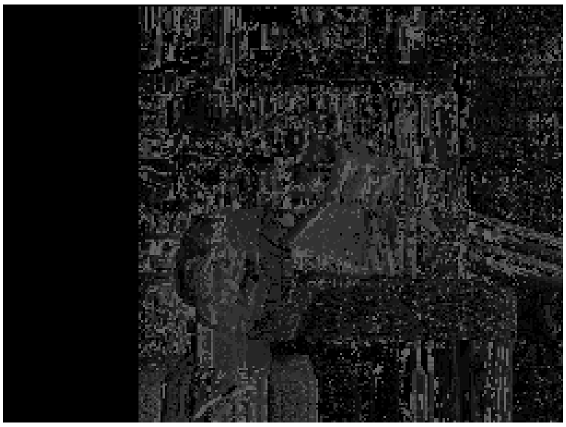

Figure 4: Final dense inverse depth map.

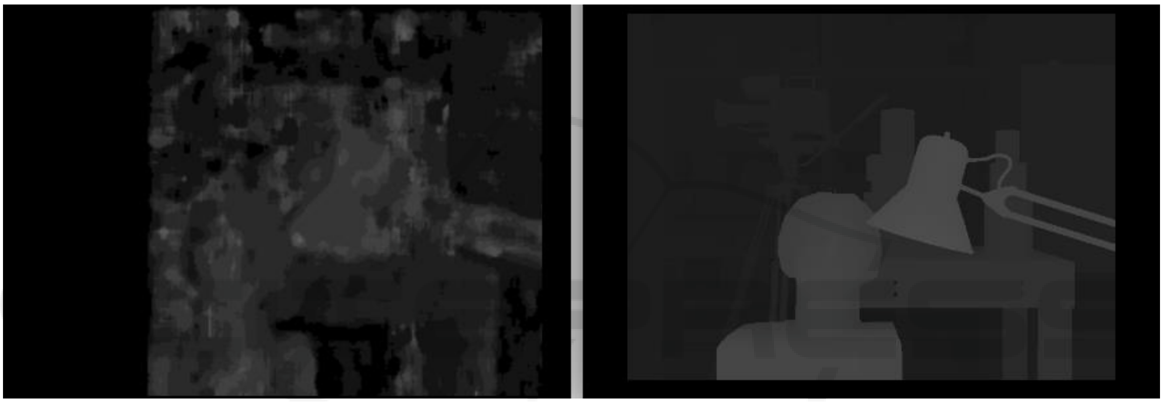

Figure 5: Left- Median filter applied to dense inverse depth reconstruction, Right- Ground Truth disparity. In stereo vision disparity is proportional to inverse depth.

6 RESULTS

We tested the algorithm on endoscopic video which we collected from Aichi Medical University. We carefully selected two nearby video frames which had translation only motion. The feature tracking fails without specularity removal leads to matching points in specular region (Fig.6). This would result in erroneous camera tracking as specular regions are not static. Specularity removal is done using 2 methods of separately as shown in Fig.7 (Bertalmio et al., 2001; Telea, 2004). We finally use the method proposed by Telea which gave better result. Even after specularity removal the matching gives poor results (Fig.8) This is because specularity removal is not perfect. Moreover, there is information loss wherever specularity is removed. We, therefore, use a mask to extract feature points only in regions where there was no specularity in the original image. This results in almost perfect matching (Fig.9). The final dense reconstruction is shown in Fig.10.

7 CONCLUSIONS

Our approach effectively applies feature matching for camera motion estimation and performs pixelwise operations to compute dense reconstruction. Even though ground truth 3D structure is not available for the endoscope images, we performed a check on our method using the Tsukuba dataset. The tracking results are highly accurate, and the dense reconstruction closely resembles the ground truth.

Further developments could be to include rotation into the model. There is also the possibility to improve reconstruction by imposing smoothness constraint. Computational efficiency can also be improved.

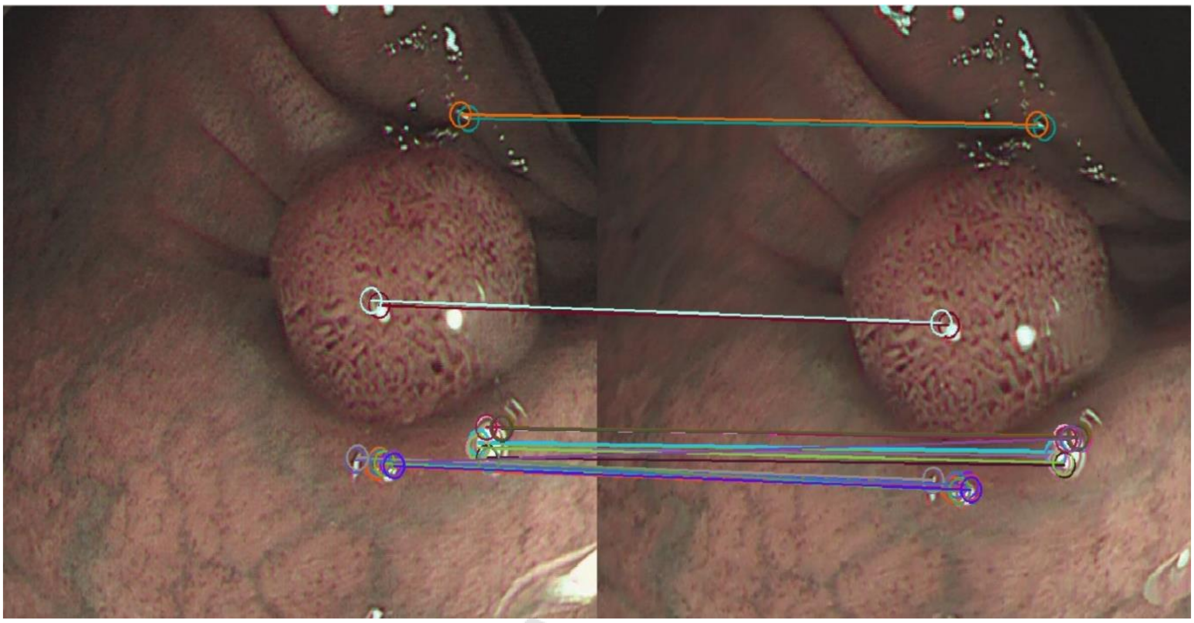

Figure 6: Matching feature points without specularity removal. Feature points are selected in specular region which would result in wrong estimation of tracking.

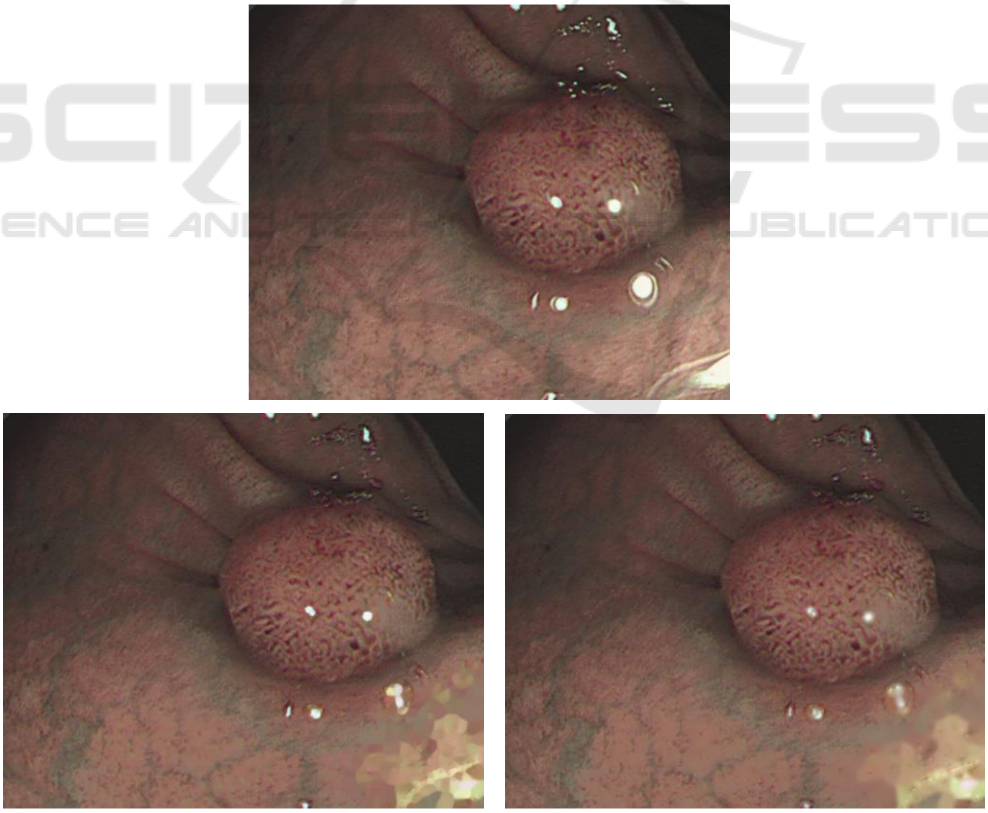

Figure 7: Top- Original Image. Bottom Left- Specularity Removal using method of Bertalmio et al., 2001. Bottom Right- Specularity removal using method of Telea, 2004.

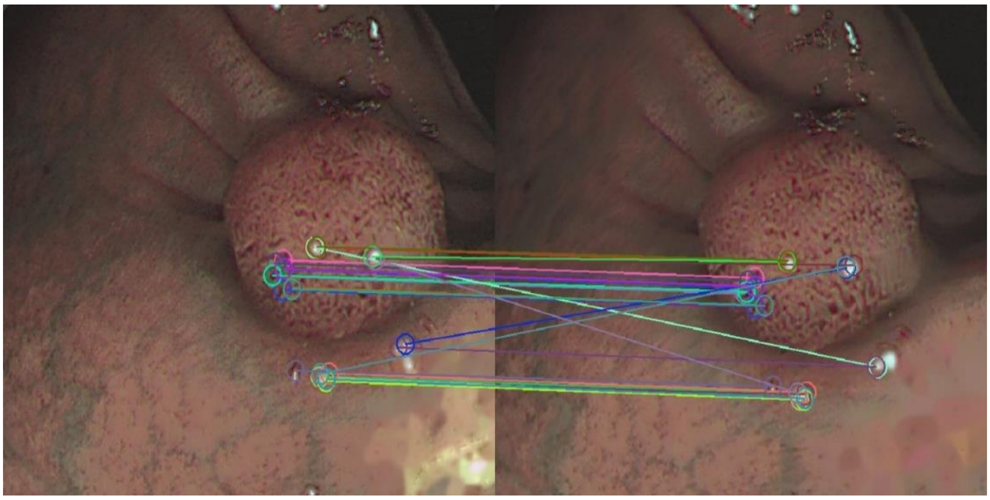

Figure 8: Matched feature points after removing specularity using method Proposed by Telea (2004).

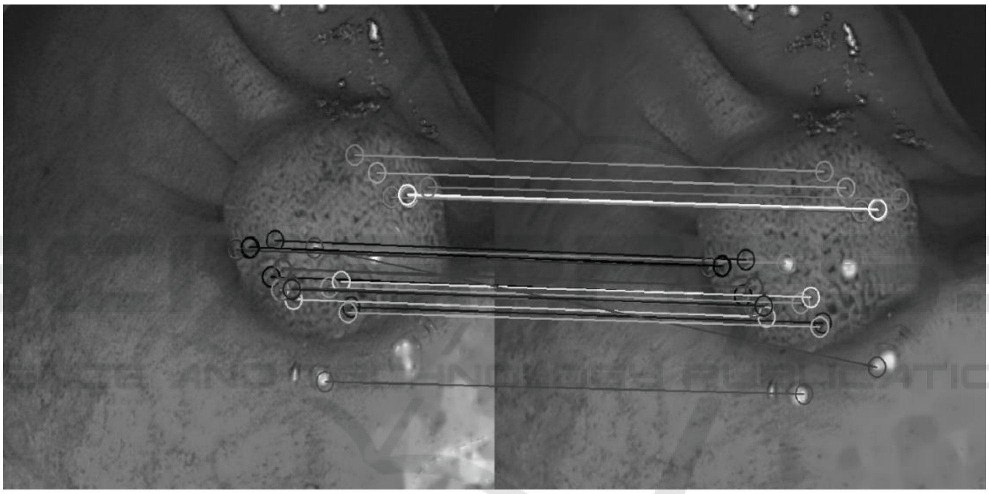

Figure 9: Matched feature points after removing specularity and applying mask to specular region.

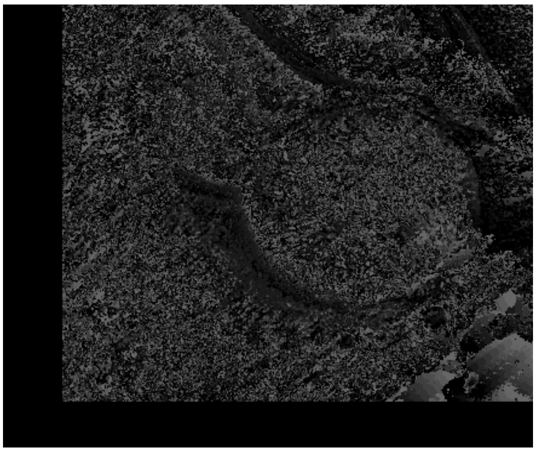

Figure 10: Final dense reconstruction of endoscopic image.

ACKNOWLEDGEMENTS

Iwahori's research is supported by JSPS Grant-in-Aid for Scientific Research (C) (17K00252) and Chubu University Grant. The authors would like to thank Iwahori Lab. member for their useful discussions.

REFERENCES

Nakatani, H., Abe, K., Miyakawa, A., and Terakawa, S. (2007). Three-dimensional measurement endoscope system with virtual rulers. *Journal of Biomedical Optics*, vol. 12, no. 5, Article ID 051803-1.

Hayashibe, Mitsuhiro, Suzuki, Naoki, and Nakamura, Yoshihiko (2005). Laser-scan endoscope system for intraoperative geometry acquisition and surgical robot safety management. *Special Issue on Functional Imaging and Modelling of the Heart (FIMH 2005)*, Volume 10, Issue 4, August 2006, Pages 509-519.

Chang, PL., Handa, A., Davison, A.J., Stoyanov, D., Edwards, P.. (2014). Robust Real-Time Visual Odometry for Stereo Endoscopy Using Dense Quadrifocal Tracking. *International Conference on Information Processing in Computer-Assisted Interventions (IPCAI 2014),* pp 11-20.

Stoyanov, Danail, Scarzanella, Marco Visentini, Pratt, Philip, Yang, Guang-Zhong (2010). Real-Time Stereo Reconstruction in Robotically Assisted Minimally Invasive Surgery. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010,* pp 275-282.

Mourgues, F., Devemay, F., and Coste-Maniere, E. (2001). 3D reconstruction of the operating field for image overlay in 3D-endoscopic surgery. *In Proceedings of the IEEE and ACM International Symposium on Augmented Reality (ISAR '01),* pp. 191–192, New York, NY, USA.

Schmalz, Christoph, Forster, Frank, Schick, Anton, Angelopoulou, Elli (2012). An endoscopic 3D scanner based on structured light. *Medical Image Analysis*, Volume 16, Issue 5, July 2012, Pages 1063-1072.

Iwahori, Yuji, Sugie, Hidezumi, Ishii, Naohiro (1990). Reconstructing shape from shading images under point light source illumination. *International Conference on Pattern Recognition (ICPR),* i. 83 - 87 vol.1. 10.1109/ICPR.1990.118069.

Wang, Guo-hui, Han, Jiu-qiang and Zhang, Xin-man (2009). Three-dimensional reconstruction of endoscope images by a fast shape from shading method. *Measurement Science and Technology*, Volume 20, Number 12.

Wu, Chenyu, Narasimhan, Srinivasa G., Jaramaz, Branislav (2010). A Multi-Image Shape-from-Shading Framework for Near-Lighting Perspective Endoscopes. *International Journal of Computer Vision,* January 2010, Volume 86, Issue 2–3, pp 211–228.

Grasa, O.G., Civera, J., Guemes, A., Munoz, V. and Montiel, J.M.M. (2009). EKF monocular SLAM 3D modeling, measuring and augmented reality from endoscope image sequences. *Medical image computing and computer-assisted intervention (MICCAI)* (Vol. 2).

Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J. and Montiel, J.M.M. (2016). Orbslam-based endoscope tracking and 3d reconstruction. *International Workshop on Computer-Assisted and Robotic Endoscopy* (pp. 72-83). Springer, Cham.

Newcombe, R.A., Lovegrove, S.J. and Davison, A.J., 2011, November. DTAM: Dense tracking and mapping in real-time. *Computer Vision (ICCV), 2011 IEEE International Conference* (pp. 2320-2327). IEEE.

Peris, M., Martull, S., Maki, A., Ohkawa, Y. and Fukui, K. (2012). Towards a simulation driven stereo vision system. In *Pattern Recognition (ICPR), 2012 21st International Conference* on (pp. 1038-1042). IEEE.

Martull, S., Peris, M. and Fukui, K. (2012). Realistic CG stereo image dataset with ground truth disparity maps. In *ICPR workshop TrakMark2012* (Vol. 111, No. 430, pp. 117-118).

Bertalmio, M., Bertozzi, A.L. and Sapiro, G. (2001). Navier-stokes, fluid dynamics, and image and video inpainting. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference* on (Vol. 1, pp. I-I). IEEE.

Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1), pp.23-34.