Stomach 3D Reconstruction Based on Virtual Chromoendoscopic Image Generation

Aji Resindra Widya 1, Yusuke Monno 1, Masatoshi Okutomi 1, Sho Suzuki 2, Takuji Gotoda 2, and Kenji Miki 3

Abstract—Gastric endoscopy is a standard clinical process that enables medical practitioners to diagnose various lesions inside a patient’s stomach. If any lesion is found, it is very important to perceive the location of the lesion relative to the global view of the stomach. Our previous research showed that this could be addressed by reconstructing the whole stomach shape from chromoendoscopic images using a structure-from-motion (SfM) pipeline, in which indigo carmine (IC) blue dye-sprayed images were used to increase feature matches for SfM by enhancing stomach surface’s textures. However, spraying the IC dye to the whole stomach requires additional time, labor, and cost, which is not desirable for patients and practitioners.

In this paper, we propose an alternative way to achieve whole stomach 3D reconstruction without the need of the IC dye by generating virtual IC-sprayed (VIC) images based on image-to-image style translation trained on unpaired real no-IC and IC-sprayed images. We have specifically investigated the effect of input and output color channel selection for generating the VIC images and found that translating no-IC green-channel images to IC-sprayed red-channel images gives the best SfM reconstruction result.

I. INTRODUCTION

Gastric endoscopy is a well-applied clinical process that enables medical practitioners to find a gastric lesion, such as an ulcer and cancer, from inside a patient’s stomach. Accurate localization of a found lesion is very important to decide the next clinical procedure. For example, if laparoscopic gastroectomy for early cancer needs to be done, the target cancer location relative to the global view of the stomach has to be known to decide the operative procedure. However, accurately recognizing the lesion’s 3D location only from 2D endoscopic images is difficult for gastric surgeons, especially when the images are captured by another endoscopist.

In our previous study, we tackled the problem of lesion localization by reconstructing the whole stomach 3D shape from endoscopic images based on SfM [1], [2]. Although the stomach 3D reconstruction by SfM is very challenging because of texture-less stomach surfaces, we found that a whole stomach shape could be reconstructed by using red-channel images of chromoendoscopy with indigo carmine (IC) blue dye, where the IC dye acts as an enhancement technique to bring up more textures to the stomach surface [3]. Compared with other approaches such as barium radiography [4] and 3D computed tomography gastrography [5], our SfM-based approach could provide the stomach 3D model with both geometric and color texture information, which is a great advantage for accurate lesion localization. However, though the IC dye is commonly used in gastroendoscopy, spraying it on the whole stomach surface requires additional procedure time, labor, and cost, which is not desirable for both patients and practitioners. Furthermore, the IC dye may also hinder the visibility for some lesions and reconstructed stomach 3D models because of its dark color tone.

In this paper, we propose a novel SfM-based approach for whole stomach 3D reconstruction that does not require chromoendoscopic image sequences. Instead of spraying the IC dye during endoscopy, we generate virtual IC-sprayed (VIC) images from no-IC images based on image-to-image style translation with CycleGAN [6]. The SfM pipeline is then applied using the generated VIC images.

With the rise of deep learning, image-to-image translation, in which the goal is to learn the mapping between one style of images to another, is attracting attention from researchers. The style translation has been proven to be useful for medical applications such as in colonoscopy depth estimation [7]. It is also reported that generating VIC images using CycleGAN improves the lesion detection and classification performance in colonoscopy [8]. Even though our VIC image generation is inspired by the study [8], we apply the generated VIC images for stomach 3D reconstruction, which is the first attempt to the best of our knowledge.

In our experiments, we trained CycleGAN for the style translation using no-IC single-color-channel images and IC-sprayed red-channel images and investigated the effect of input color channel selection. As a result, we found that the CycleGAN translating the no-IC green-channel images to the IC-sprayed red-channel images gives the best VIC images for SfM. Using those VIC images, we were able to reconstruct the whole stomach 3D model without the need of real IC-sprayed images. We can also localize an image frame including a gastric ulcer in the reconstructed 3D model.

II. MATERIALS AND METHODS

A. Endoscope video dataset

In this work, we used exactly the same endoscope video dataset from our previous work [2]. In the dataset, there are seven videos captured from seven subjects undergoing
general gastroendoscopy procedure. Each video contains two different image type sequences: no-IC and IC-sprayed sequences. We extracted the image frames from each video and divided them based on its sequence type to obtain training image data for VIC generation and also test no-IC sequences for 3D reconstruction. The experimental protocol was approved by the research ethics committees of Tokyo Institute of Technology and Nihon University Hospital.

B. Cycle-consistent image-to-image translation (CycleGAN)

Since the exact pair of no-IC and IC-sprayed images is impossible to obtain, we decided to use CycleGAN [6] as our image-to-image translator that supports unsupervised and unpaired training data. Let $A$ and $B$ be two different image domains. CycleGAN consists of two sets of generator and discriminator pair, $(G_A, D_A)$ and $(G_B, D_B)$. The generator’s task is to generate a virtual image by translating an input image from one domain to another and fool its opposite domain’s discriminator. In the other hand, the discriminator’s task is to distinguish between the generated and real images.

The loss of CycleGAN consists of discriminator, generator and cycle consistency loss. Let $b' = G_A(a)$ and $a' = G_B(b)$ describe the output of a generator from a random input image $a$, $b \sim p_{data}$, respectively. Each respective discriminator, $D_A$ and $D_B$, should give high scores for real input images, $a$ and $b$, and low scores for generated input images, $a'$ and $b'$, respectively. The consistency loss makes sure that CycleGAN generates a close image with the original input image when translating it circularly, e.g., $a \approx a_{cyc}$, where $a_{cyc} = G_B(G_A(a))$. The consistency loss enables CycleGAN to be trained on the unpaired set of images for image-to-image style translation.

C. Generating VIC images using CycleGAN

Figure 1 shows our CycleGAN training and application overview (the RGB color image case is shown for better visualization, though we actually use single-channel images as explained below). We train CycleGAN to translate the styles between no-IC and IC-sprayed images using the previously mentioned image dataset.

In our previous research [2], we observed that there is color channel misalignment in the RGB data. Because of that, we used single-channel images for the 3D reconstruction and investigated which color channel gives the best 3D reconstruction result. It was found that the whole stomach could be reconstructed using IC-sprayed red-channel images. This is because the red channel of IC-sprayed images has the best contrast and the most visible textures. We also found that, for the case of no-IC images, the green channel gave the best 3D reconstruction result, though only partial stomach could be reconstructed. The blue channel was not preferable for 3D reconstruction due to low contrasts.

Based on those findings, we set the target domain to the IC-sprayed red-channel, since it is the best channel for stomach reconstruction. We then conducted two separated CycleGANs training using the no-IC red and green channels to investigate which color channel is better for the input to generate VIC images for SfM. Specifically, we set the domain $A$ and $B$ for each CycleGAN to (i) no-IC red-channel and IC-sprayed red-channel and (ii) no-IC green-channel and IC-sprayed red-channel. For the rest of this paper, we will refer to each CycleGAN as cGAN_{r2r} and cGAN_{g2r}, respectively. We then applied each trained CycleGAN to generate the VIC images from no-IC images.

D. 3D reconstruction using the generated VIC images

We followed the 3D reconstruction pipeline presented in our previous research [2]. It consists of point cloud reconstruction, outlier removal, and mesh and texture generation. The point cloud reconstruction follows the general flow of SfM [9]. It starts with extracting scale invariant feature transform (SIFT) features [10] from all of the input images followed by exhaustive feature matching. Those steps are then followed by features triangulation, poses estimation, and bundle adjustment [11] in parallel. It is then followed by random sample consensus (RANSAC)-based plane-fitting outlier removal to remove apparent outlier 3D points. Meshing and texturing is also performed to obtain the textured 3D mesh model.

III. RESULTS AND DISCUSSION

A. Implementation details

We trained CycleGAN using a single NVIDIA GeForce GTX 1080Ti GPU. We used the same network setting as presented in [6]. The network was trained for 100 epochs for each channel setting, i.e., cGAN_{r2r} and cGAN_{g2r}, with 7978 no-IC images and 7453 IC-sprayed images. The training took approximately 28 hours to complete. Due to the GPU memory limitation, we resized the original $1155 \times 1003$ images to $600 \times 524$ pixels and trained the CycleGANs with randomly cropped image patches of $510 \times 510$ pixels. For the 3D reconstruction pipeline, we used exactly the same setup and implementation with our previous research [2].

B. VIC image generation results

We first show the example results of generated VIC images using cGAN_{r2r} and cGAN_{g2r}. Figure 2 shows the comparison between the input no-IC images and the generated VIC images.
(a) No-IC red-channel
(b) VIC from (a) with cGAN\textsubscript{r2r}
(c) No-IC green-channel
(d) VIC from (c) with cGAN\textsubscript{g2r}

Fig. 2: Example results of generated VIC images.

As we can see from the results, both cGAN\textsubscript{r2r} and cGAN\textsubscript{g2r} can transfer the style of the IC-sprayed image to the input no-IC image, not only by transferring the IC pattern, but also either tuning up or down the overall surface brightness. However, if we see the no-IC red-channel image of (a), we can observe that the surface looks fairly texture-less. It is hard even for convolutional neural networks (CNN) to extract features from this kind of texture-less images. In other hand, the no-IC green-channel image of (c) shows more textures, enabling slightly better style transfer. We will discuss the effect of the input channel in more detail in the following subsection.

C. Feature matching results

Using the generated VIC images for Subject B, we calculated the average number of extracted SIFT features per image. The VIC images from cGAN\textsubscript{r2r} have the average of 3003.05 features, while the VIC images from cGAN\textsubscript{g2r} have the average of 3511.49 features. As the baselines, we also calculated the average numbers of extracted SIFT features of no-IC red-channel and green-channel images, which are 633.75 and 743.18 features, respectively. It is clear that the VIC images have more extracted features compared to the no-IC images by more than four times.

However, solely increasing the number of features is not sufficient. Since SfM relies on the consistency of features across multiple images, we also tested the feature matching performance of the generated VIC images. For this purpose, we took 11 consecutive images from a sequence. We then chose the first image as an anchor, \( t \), and performed feature matching to all of its consecutive images, \( t+1, t+2, ..., t+10 \).

Figure 3 shows the average number of feature matches between the anchor frame and its 10 consecutive frames. The x-axis represents the relative time stamp and the y-axis represents the average number of matches calculated using 43 samples. It is clearly shown that the VIC images from cGAN\textsubscript{g2r} has a higher number of matches across frames.

Fig. 3: Comparison of the average number of feature matches between the anchor frame and its 10 consecutive frames.

Fig. 4: The example feature matching result for two frames \((t \text{ and } t + 9)\). The red marks represent the locations of extracted SIFT features. The color lines represent the matched features. It is clear that the number of feature matches in (b) is much fewer than that in (c), even though the number of extracted features in (b) significantly increases from (a). This result implies that the generated VIC images from cGAN\textsubscript{g2r} have better pattern consistency between the frames.
between the anchor frame and each of its consecutive frames taken from 43 group-of-11-consecutive-images samples, which were extracted from the sequence of Subject B. It can be seen that the VIC images from cGANr2r has a higher number of matches across frames compared to the other three image types. It implicitly means that the VIC images from cGANr2r has better temporal pattern consistency between frames. Figure 4 shows the example feature matching results.

D. 3D reconstruction results

Figure 5 shows the SfM reconstruction results for Subject B and D using three different image types, i.e., no-IC green-channel images, VIC images from cGANr2r, and VIC images from cGANg2r. Please note that all three types of images were extracted or generated from the same source RGB sequence and thus the comparison can be fairly performed. Using those types of images, 49.43%, 88.84%, and 99.77% images of Subject B and 34.74%, 35.94%, and 99.27% images of Subject D were reconstructed, respectively. In (a) and (d), the stomach shape cannot be reconstructed using no-IC green-channel images. In (b) and (e), the results using VIC images from cGANr2r only show partially reconstructed stomach shapes. In (c) and (f), we can confirm that the results using VIC images from cGANg2r achieve the best point cloud quality and completeness.

Table I shows the objective evaluation of SfM reconstruction results on all seven subjects. It shows that the generated VIC images from cGANg2r achieve better results on all subjects compared to the baseline no-IC green-channel images. Using the VIC images for SfM significantly improves the number of reconstructed images, especially for Subject B to F. All reconstruction results using the VIC images achieve more than 95% of reconstructed images. The triangulated 3D points also demonstrate significant improvement. It is because an increased number of feature matches across multiple images could be obtained from the VIC images. This led to the increase of features that could be triangulated, as shown by “Avg. observation” in the table, which indicates the average number of triangulated feature points per image.

Figure 6(a)–6(d) show the meshing and texturing results using the VIC images from cGANg2r. We can see that the
resulting meshes are well reconstructed and well textured. We can also confirm that the resulting meshes resemble the shape of the stomach. In Fig. 6(e), we also show the localization result of the gastric ulcer in Subject G. To localize the ulcer, we chose a reconstructed VIC image containing the ulcer and then projected the corresponding RGB image onto the reconstructed mesh using the estimated camera pose. We believe that our SfM-based approach could become a useful tool for the lesion localization.

IV. Conclusion

In this paper, we have presented a new approach to reconstruct a whole stomach 3D shape from a gastroendoscopy video without the need of IC dye spraying. We have applied CycleGAN as an image-to-image style translator to generate VIC images from no-IC images for the stomach 3D reconstruction and showed that the generated VIC images significantly increase the number of extracted SIFT feature points. Furthermore, we have found that input color channel selection for the style translation affects the feature matching performance of the VIC images. Based on the investigation, we have found that translating from no-IC green-channel images to IC-sprayed red-channel images gives significant improvements to the SfM reconstruction quality. We have experimentally demonstrated that our new approach can reconstruct the whole stomach shapes of all seven subjects and showed that the estimated camera poses can be used for the lesion localization purpose. The reconstruction result videos can be accessed from the following link (http://www.ok.sc.e.titech.ac.jp/Res/Stomach3D). In future work, we plan to investigate on how to increase the temporal pattern consistency of the VIC images to further improve the feature matching performance.

References

[1] A. R. Widya, Y. Monno, K. Imahori, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, “3D reconstruction of whole stomach from endoscopy video using structure-from-motion,” in Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3900–3904, 2019.
[2] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, “Whole stomach 3D reconstruction and frame localization from monocular endoscopy video,” IEEE Journal of Translational Engineering in Health and Medicine, vol. 7, no. 3300310, pp. 1–10, 2019.
[3] P. F. Alcantarilla, A. Bartoli, F. Chadebecq, C. Tilmant, and V. Lepillieze, “Enhanced imaging colonoscopy facilitates dense motion-based 3D reconstruction,” in Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 7346–7349, 2013.
[4] N. Yamamichi, C. Hirano, Y. Takahashi, C. Minatsuki, C. Nakayama, R. Matsuda, T. Shimamoto, C. Takeuchi, S. Kodashima, S. Ono, Y. Tsuji, M. Fujishiro, R. Wada, T. Mitsushi, and M. Koike, “Comparative analysis of upper gastrointestinal endoscopy, double-contrast upper gastrointestinal barium X-ray radiography, and the titer of serum anti-Helicobacter pylori IgG focusing on the diagnosis of atrophic gastritis,” Gastric cancer, vol. 19, no. 2, pp. 670–675, 2016.
[5] J. W. Kim, S. S. Shin, S. H. Heo, H. S. Lim, N. Y. Lim, Y. K. Park, Y. Y. Jeong, and H. K. Kang, “The role of three-dimensional multilayer CT gastroscopy in the preoperative imaging of stomach cancer: Emphasis on detection and localization of the tumor,” Korean Journal of Radiology, vol. 16, no. 1, pp. 80–89, 2015.
[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. of IEEE Int. Conf. on Computer Vision (ICCV), pp. 2223–2232, 2017.
[7] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, “Implicit domain adaption with conditional generative adversarial networks for depth prediction in endoscopy,” Int. Journal of Computer Assisted Radiology and Surgery, vol. 14, no. 7, pp. 1167-1176, 2019.
[8] A. Fukuda, T. Miyamoto, S. Kamba, and K. Sumiyama, “Generating virtual chromoendoscopic images and improving detectability and classification performance of endoscopic lesions,” in Proc. of MICCAI Workshop on Domain Adaptation and Representation Transfer (DART), pp. 99–107, 2019.
[9] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4104–4113, 2016.
[10] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.
[11] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, “Bundle adjustment – A modern synthesis,” in Proc. of Int. Workshop on Vision Algorithms, pp. 298–372, 1999.