Jan Mangulabnan
Project 12
Mentors: Dr. Roger Soberanis, Dr. Mathias Unberath

Background Reading Report:
Reconstructing Sinus Anatomy from Endoscopic Video –
Towards a Radiation-Free Approach for Quantitative Longitudinal Assessment

1. Introduction

The paper focused on for this report was titled Reconstructing Sinus Anatomy from Endoscopic Video – Towards a Radiation-Free Approach for Quantitative Longitudinal Assessment [1] published for the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference in 2020. The authors of this paper are: Xingtong Liu, Maia Stiber, Jindan Huang, Masaru Ishii, Gregory D. Hager, Russell H. Taylor, and Mathias Unberath. This paper discusses the pipeline that will be used for this project to generate patient-specific 3D reconstructions of the sinus anatomy from input endoscopic video sequences.

While the authors provided an evaluation of the pipeline, there is no direct comparison between the ground truth CT scans and the reconstruction. My project focuses on developing a framework for both global and local registration to evaluate the accuracy of the sinus reconstruction with respect to the CT. This will allow for a baseline evaluation which will serve as point of comparison for consequent changes to the pipeline.

2. Technical Summary

The pipeline has three main modules: Structure from Motion (SfM) with dense descriptor, depth estimation, and depth fusion and surface extraction, which can be seen in Figure 1.

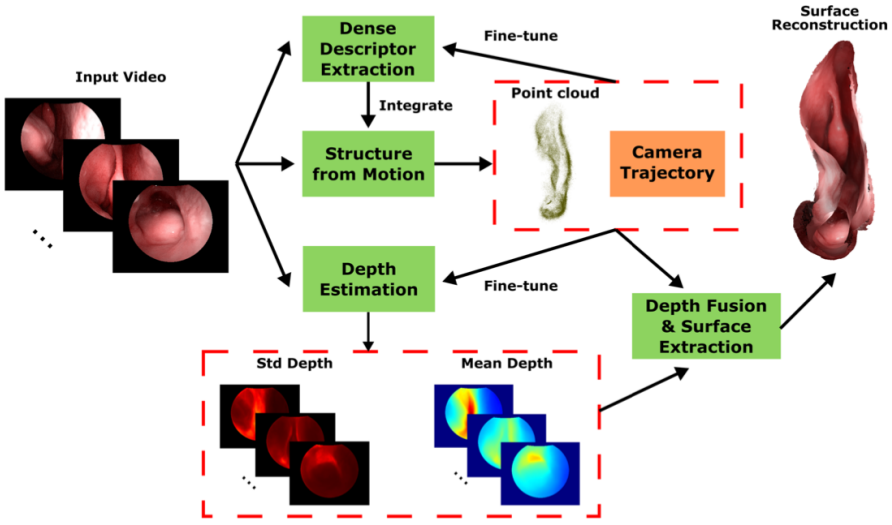

Figure 1. Dense reconstruction pipeline adapted from [1].

a. Structure from Motion with Dense Descriptor

The SfM with Dense Descriptor module was based off previous work [2] that uses a self-supervised learning method to extract dense descriptors from endoscopic video. This module uses an initial SfM with the SIFT descriptor, which is a standard feature descriptor used in computer vision for natural images. This descriptor is then used as the supervisory signal to fine-tune a pre-trained model to produce patient-specific dense descriptors. The resulting descriptors are then used in dense feature matching to produce correspondences between the feature points which are then integrated into the SfM algorithm. SfM then outputs a dense point cloud of the patient sinus anatomy and the camera trajectories of each input frame. The SfM point cloud is illustrated in Figure 2.

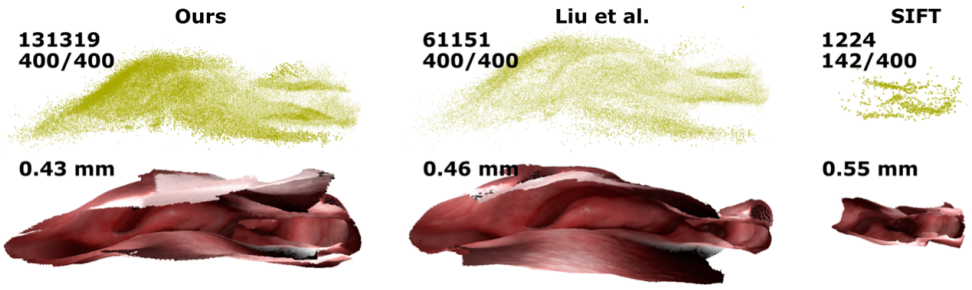

Figure 2. Adapted from [1]. *Comparison of resulting Structure from Motion point cloud using patient-specific descriptors, dense descriptors from pre-trained model, and SIFT descriptor with resulting reconstruction.*

b. Depth Estimation

The Depth Estimation module is also a self-supervised learning module based off the work in [3] used to estimate the distance from the camera to the surfaces in the input endoscopic video. This model uses guidance from dense descriptors generated from the SfM output. The depths are then represented as a probabilistic model based on a Gaussian distribution using the mean and standard deviation of the depth estimates. A sample of the depth maps generated by this module can be seen in Figure 3.

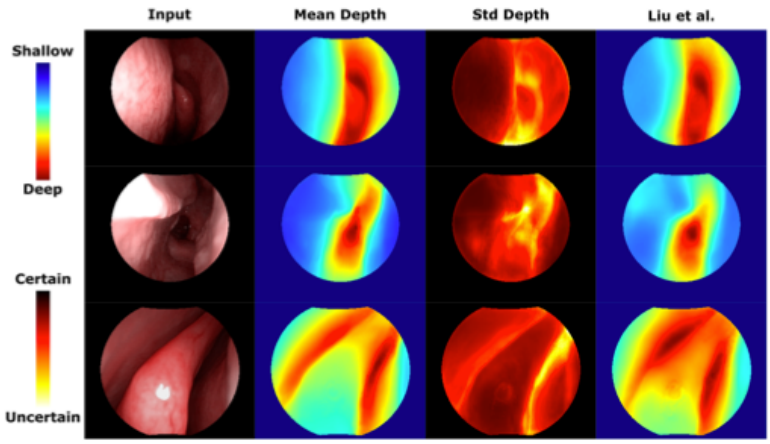

Figure 3. Adapted from [1]. *Heat map of resulting patient-specific depth estimates represented as mean and standard deviation in comparison to original input video and depth estimated from pre-trained model.*

c. Depth Fusion and Surface Extraction

The final step in the pipeline is the Depth Fusion and Surface Extraction which uses the dense point cloud and camera trajectories generated by the SfM with Dense Descriptors step, and the mean and standard deviation produced by the Depth Estimation. These outputs are used in a fusion method based on truncated signed distance functions [4] where a ray is cast from the SfM camera trajectories, and the mean estimated depth is used to establish the length and direction of the ray. The original SfM point cloud is then used to rescale the point cloud after depth fusion and the marching cubes method [5] is applied to generate a watertight mesh from the resulting point cloud.

3. Results

a. Comparison with SfM

The authors first compared the resulting reconstruction with the sparse SfM point cloud generated without fine-tuning the model. Since the sparse SfM results were used as a self-supervisory signal for training, this comparison was done to evaluate consistency throughout the pipeline to ensure that using these signals were accurate towards the reconstruction. The proposed reconstruction was rescaled based on the results of SfM and to unify scale, they also rescaled the sparse SfM results based on ground truth CT surface models. A qualitative comparison between the sparse SfM point cloud and the resulting reconstruction can be seen in Figure 4. Quantitatively, they were able to observe minimal discrepancies between the reconstruction and SfM output achieving an average point-to-mesh distance of 0.34 (± 0.14) mm.

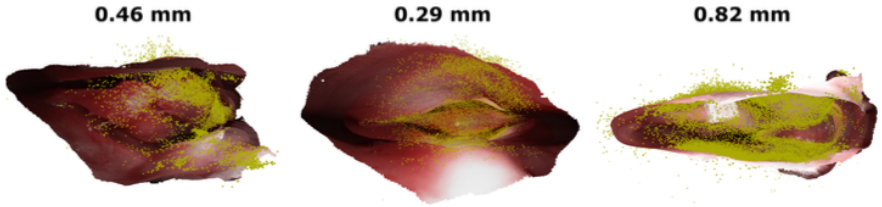

Figure 4. Overlay of sparse SfM point cloud and surface from reconstruction pipeline adapted from [1].

b. Consistency against Video Variation

The authors also provided an analysis of the consistency of the reconstruction against video variation. When capturing endoscopic video, there may be variable factors such as the camera speed of the collected frames as the clinician scans the patient anatomy. Considering the endoscopic video is the only input for this pipeline, it is important that the reconstruction is robust to these variations. In order to test this, they subsampled frames from the input video by randomly selecting 7 frames for every 10 consecutive frames. The reconstruction produced by this subset was then compared to the reconstruction from the entire input sequence. The structures were then registered with each other to achieve an average residual error of 0.21 (± 0.10) mm.

c. Comparison with COLMAP Methods

The reconstruction method was also evaluated in comparison with the ball-pivoting method [6] used to create a triangular mesh from a point cloud. The qualitative results can be seen in Figure 5. The authors also mention that the Poisson method is a popular choice for this application but does not produce reasonable results which is why it is not used in the comparison.

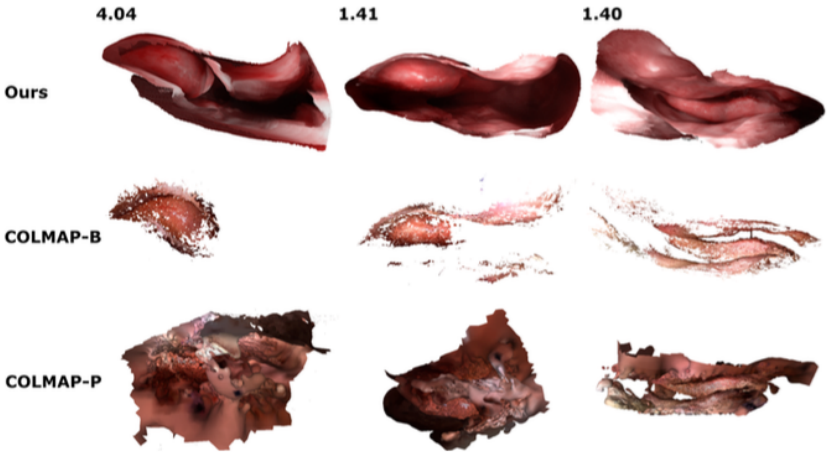

Figure 5. Adapted from [1]. *Comparison of surface reconstruction using proposed method, ball-pivoting method (COLMAP-B) and Poisson method (COLMAP-P).*

The dense reconstruction was registered with the mesh produced by the ball-pivoting method and achieved an average residual distance of 0.24 (± 0.08) mm. This reconstruction method was also reported to have a runtime of 127 minutes in comparison to COLMAP which had a runtime of 778 minutes.

d. Comparison with CT

Lastly, the reconstruction was evaluated by comparing the results to the ground-truth CT as the main goal of this paper was to evaluate the pipeline as a CT alternative to provide a 3D structure of sinus anatomy. They registered the camera trajectories from SfM to the CT which was used as the origin to isolate cross-sectional areas of the volume. These cross-sectional areas were compared to report relative differences. The registration between the cross-sectional areas was reported to have an average residual error of 0.69 (± 0.14) mm and a qualitative comparison can be seen in Figure 6.

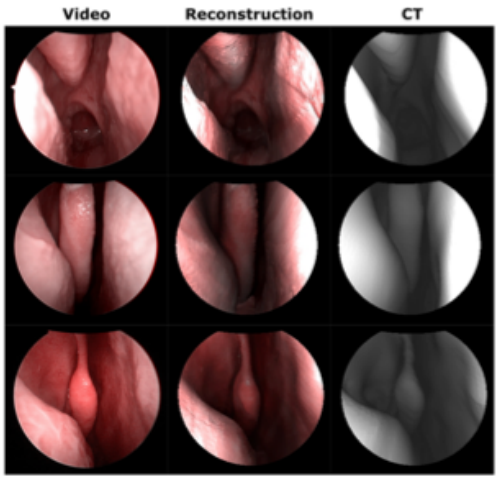

Figure 6. Adapted from [1]. *Comparison of input video frame to rendering of reconstruction and CT at estimated camera trajectory.*

4. Discussion

This report demonstrates the strengths of the proposed dense reconstruction pipeline as the various experiments were able to achieve sub-millimeter errors when compared to the sparse SfM reconstruction, subsampled video frames, COLMAP methods for reconstruction, and the ground-truth CT. The paper was able to evaluate the consistency of the modules throughout the pipeline and report metrics relevant towards use in a clinical setting. However, it is unclear how the relative differences were computed to compare the reconstruction to the CT and cross-sectional planes were used to evaluate rather than the entire CT volume. This project will improve on this comparison by implementing a registration framework to directly evaluate the reconstruction with respect to the CT scan. This will provide a stronger analysis of the accuracy of the dense reconstruction to serve as a baseline for further improvements of the pipeline.

References

[1] Liu, X., Stiber, M., Huang, J., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Reconstructing sinus anatomy from endoscopic video – towards a radiation-free approach for quantitative longitudinal assessment. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 3–13. Springer, Cham (2020)

[2] Liu, X., et al.: Extremely dense point correspondences using a learned feature descriptor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4847–4856 (2020)

[3] Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M.: Dense depth estimation in monocular endoscopy with self- supervised learning methods. IEEE transactions on medical imaging 39(5), 1438– 1447 (2019)

[4] 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, pp. 303–312 (1996)

[5] Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. ACM SIGGRAPH Comput. Graph. 21, 163–169 (1987)

[6] Bernardini, F., Mittleman, J., Rushmeier, H., Silva, C., Taubin, G.: The ball- pivoting algorithm for surface reconstruction. IEEE Trans. Vis. Comput. Graph. 5(4), 349–359 (1999)