Deep Learning-Based Scene Reconstruction from Monoscopic Videos in Otologic Surgery

Andy S. Ding, MD MSE[1]; Zhaoshuo Li, MSE[2]; Manish Sahu, PhD[2]; Mathias Unberath, PhD[2]; Russell H. Taylor, PhD[2]; Francis X. Creighton, MD[1]

[1] Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2] Department of Computer Science, Johns Hopkins University Whiting School of Engineering, Baltimore, Maryland, USA

Introduction

- Recent advances in computer vision have enabled accurate reconstruction of 3D objects in monoscopic video
- 3D scene reconstruction from monoscopic video has the potential to augment endoscopic approaches in temporal bone surgery.
- This study presents a preliminary framework for recreating the surgical scene from monoscopic video using deep learning-based surface reconstruction techniques.

Objectives

1. To demonstrate feasibility of reconstructing a 3D surgical scene from a 2D video stream using state of the art structure-from-motion neural network.
2. To evaluate the accuracy of this 3D scene reconstruction technique against post-operative imaging.

Methods

- Monocular video of a PHACON temporal bone phantom was recorded after a completed cortical mastoidectomy.
- A corresponding post-operative cone-beam CT scan was obtained, and bony structures were manually segmented.
- Using the open-source structure-from-motion algorithm COLMAP, we obtained the 3D location and intrinsic parameters of the recording camera.
- Video images were then input to a neural surface reconstruction algorithm, which encodes positional information into corresponding pixels.
- A 3D mesh of the drilled phantom was extracted from the surface reconstruction output using the marching cubes algorithm.
- The extracted mesh and the segmented CT were then co-registered and compared by calculating the average Hausdorff distance between the mesh and CT.

Results

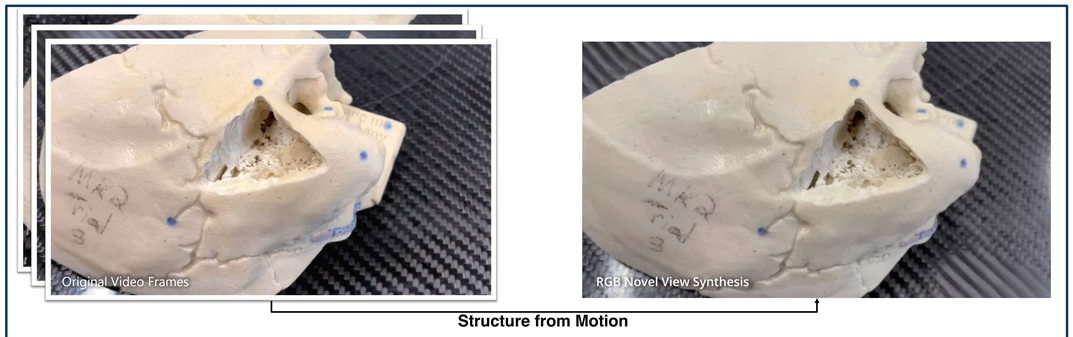


Figure 1 — Comparison of 3D Reconstruction to Original Video Frames
COLMAP was able to produce high-fidelity 3D scenes from multiple camera poses that look virtually identical to original video frames.

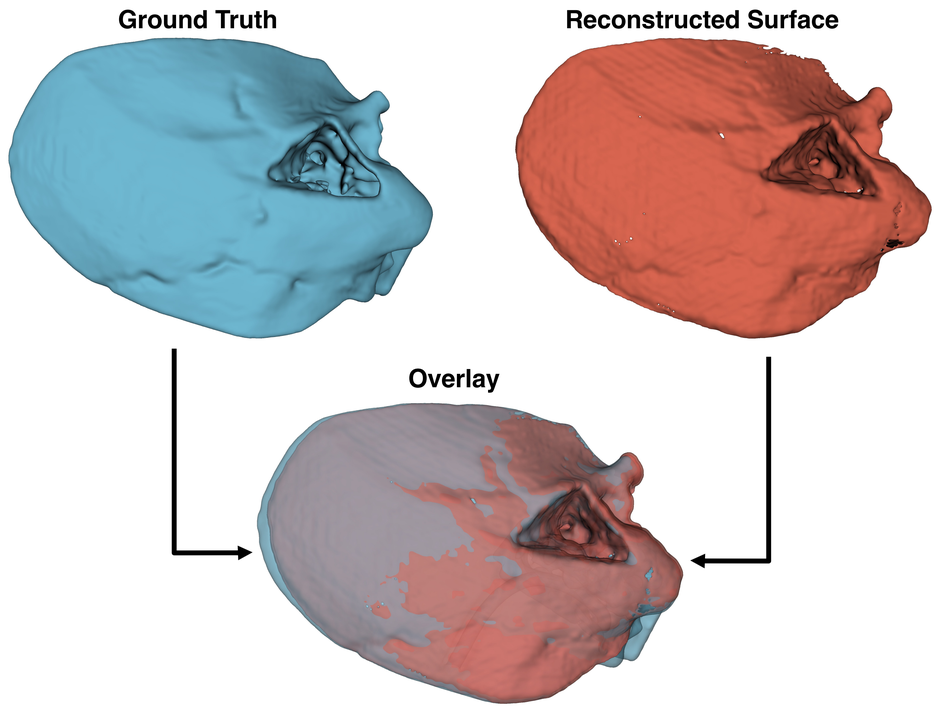
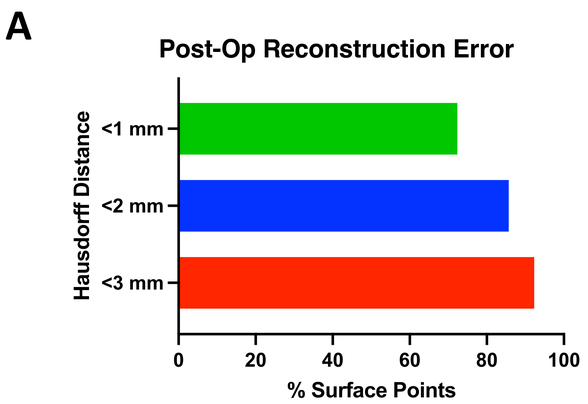


Figure 2 — 3D Scene Reconstruction Compared to Post-Operative Imaging Registration
Top Left) Ground truth segmentation of the temporal bone from post-operative CT scan. Top Right) 3D reconstruction of the surgical scene after mastoidectomy. Bottom) Rigid registration of the 3D reconstruction onto the ground truth.



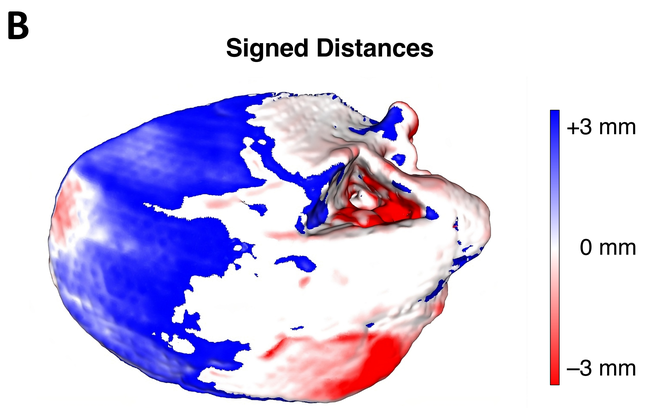


Figure 3 — 3D Scene Reconstruction Accuracy Analyses
A) Percentage of surface points from the 3D scene reconstruction within 1, 2, or 3 mm from the post-operative 3D mesh. Mean Hausdorff distances for the entire scene (0.98 ± 1.76 mm).
B) Signed distances overlaid onto the 3D scene reconstruction.

Conclusion

- We demonstrate submillimeter accuracy for surgical scene reconstruction from monoscopic video.
- Coupled with deep learning-based instrument tracking, this depth estimation algorithm has the potential to provide real-time image-guided navigation in microscopic or endoscopic procedures.