CIS 2 Molly O'Brien
Seminar Critical Review 3/13/17
Team 4: 3D Tool Tracking in the Presence of Microscope Motion

Introduction:

The goal of my CIS 2 project is to accurately 3D track tool motion from stereo microscope video in the presence of microscope motion. Two main technical challenges of my project are tracking the background motion and finding the tool tip in the microscope video. The first paper I chose to review, "Image-Based Navigation for Functional Endoscopic Sinus Surgery Using Structure From Motion" [1] explains a method for extracting 3D points from endoscope video and computing the endoscope movement between frames. This is applicable to my project because to track the background motion I am triangulating 3D points from the stereo video and computing the rotation between the points. I also chose to review the paper "Visual Tracking of Laparoscopic Instruments in Standard Training Environments" [2]. This paper presents an algorithm to identify the tool tip in color images. I plan on implementing a similar algorithm to find the tool tip in my stereo video.

Paper 1: Image-Based Navigation for Functional Endoscopic Sinus Surgery Using Structure From Motion [1]

Motivation:

Functional Endoscopic Sinus Surgery (FESS) is a surgery that opens up the nasal airways to treat chronic sinusitis or remove polyps. The surgeon uses an endoscope for visualization and clears out small bones and cartilage from the nasal canal. FESS is a challenging procedure because the surgeon needs to operate in a long, narrow space near critical structures like nerves and arteries. Common complications of FESS are cerebrospinal fluid leaks, blindness, difficulty controlling eye movement, and excessive bleeding. Medical navigation systems can be used to register endoscope video to a preoperative CT scan to help the surgeon visual where he or she is with respect to critical structures, but current systems have tracking error over 1mm. To make FESS safer for patients and easier for surgeons to perform, this paper proposes a registration method that can track with 0.91mm error when there is no erectile tissue and 1.21mm when there is erectile tissue in the nose.

Technical Approach

To achieve submillimeter tracking this paper proposes

1. Generating a 3D point cloud using Structure from Motion and Bundle Adjustment with endoscope video
2. Registering the 3D point cloud to a preoperative CT scan with Trimmed-ICP

Once the registration has been performed, areas of interest segmented on the CT scan can be overlaid on the endoscope video for guidance.

Structure from Motion and Bundle Adjustment:

To register the endoscope video to a preoperative CT scan, 3D point clouds are extracted from the endoscope video. SURF features were extracted from each frame of the video. The Hierarchical Multi-Affine (HMA) algorithm was used to match features between

pairs of endoscope frames. HMA gives robust matches by matching clusters of features between images. This method can recover 3D geometry with corresponding feature matches from as few as 15 frames.

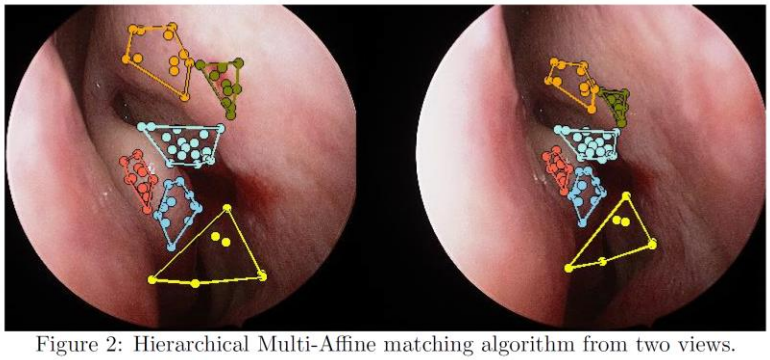


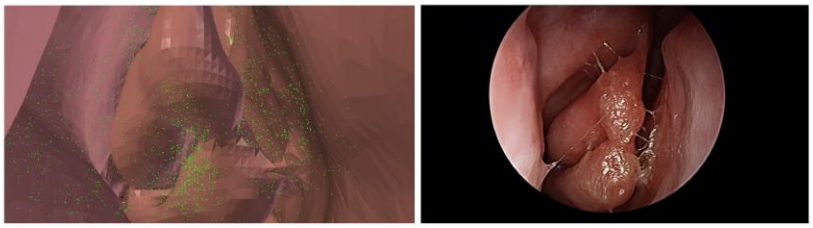
Figure 2: Hierarchical Multi-Affine matching algorithm from two views.

Figure 1. HMA matching algorithm from two views. Figure taken from [1].

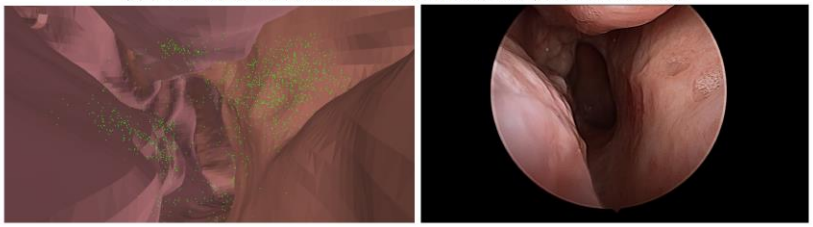
The 3D positions of the feature points and the endoscope motion between the frames was estimated up to a scale factor using structure from motion then optimized with sparse bundle adjustment. SBA looks for the camera motion between each frame that minimizes the overall error between the 3D points. A magnetic tracker was attached to the endoscope to find the magnitude of the endoscope motion and the scale factor in the 3D point clouds and camera motion.

Registration:

Next the 3D point clouds generated from the endoscope video were registered to the preoperative CT scan using Iterative Closest Point (ICP). ICP is an algorithm that find the correspondences transformation between two point clouds. Trimmed ICP is a variant of traditional ICP that is robust to noisy data [3]. Trimmed ICP is used to match the 3D point clouds generated using structure from motion to the preoperative CT scan.



(a) Point cloud registration with erectile tissues (middle turbinate).



(b) Point cloud registration with non-erectile tissues (nasopharynx).

Figure 6: Registration of two point clouds to CT scans.

Figure 2. (Left) 3D point clouds generated from endoscope video overlaid on the CT scan model. (Right) endoscope images from patients with and without erectile tissue. Figure taken from [1].

Experimental Setup

To test the tracking algorithm, endoscope videos were recorded and registered to a preoperative CT scan. The average endoscope video time was 90 seconds. The surgeon inserted the endoscope into both airways. Some subjects had significant congestion differences between when they had the CT scan and when the endoscope video was taken. This made the registration more difficult. An initial pose estimate for trimmed ICP was manually given.

Key Results

The key result of this work is submillimeter tracking accuracy. The average 70^{th} percentile registration error between the CT surface and the 3D point cloud was 0.91mm when there was no erectile tissue in the video and 1.21mm when there was erectile tissue. There was no ground truth camera pose to compare the endoscope tracking results to. The error was evaluated by comparing a binary image of a segmented structure in the endoscope frame and a binary image the same structure projected into the computed endoscope position from the CT model. There was an 86% overlap of these structures.

Conclusion

This paper presents state of the art tracking results with a near real-time algorithm. The future work includes clinical validation of the algorithm. Another goal is to improve the algorithm's robustness to initial guesses and erectile tissue.

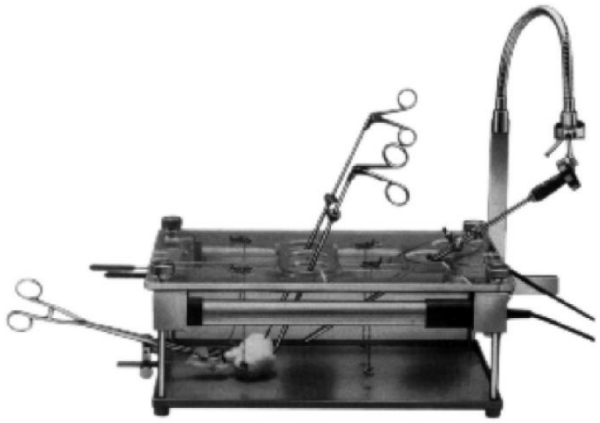
This paper is interesting because it uses multiple views from a single camera to estimate 3D feature positions instead of using stereo views. One weakness in the paper is that the true registration error is unclear. The error was computed between the 3D triangulated points and the CT mesh, but we do not know how accurate the 3D triangulated points are. Similarly, it is unclear how an 86% overlap of segmented structures translates into the 3D endoscope pose error.

This paper is relevant to my research because I am also triangulating 3D points from video. My problem is easier though because I can use stereo video from the microscope to directly compute 3D point clouds in each frame. I may also implement bundle adjustment to compute the camera motion between each frame depending on how well simple ICP between frames works.

Paper 2: Visual Tracking of Laparoscopic Instruments in Standard Training Environments

Motivation:

Laparoscopic surgery requires unique motions to manipulate anatomy with long instruments through small incisions. While surgical residents are training to do laparoscopic surgery, it is important to have a good measure of their skill. But, "training surgeons have little ability to self-assess" [2]. The Fundamental of Laparoscopic Surgery (FLS) Toolbox is a standard training tool for laparoscopic surgery that lets surgeons practice doing laparoscopic tasks.



(b) A standard FLS box trainer.

Figure 3. An image of the FLS box trainer. Taken from [2].

Simple metrics can be used to evaluate a resident's performance like the time required to do a task or the number of times the resident makes a mistake. These sorts of metrics do not capture information like the quality of the path the surgeon chose or how economical the surgeon's movements were. More sophisticated magnetic or mechanical tracking systems can get detailed tool movement information but they require cumbersome physical attachments. Virtual reality surgical environments can be used in training but these environments train with unrealistic joysticks movements. The goal of this paper is to get accurate tool tracking information from video of the FLS Toolbox to gauge surgeon skill.

Technical Approach

This paper proposes a tool tracking algorithm from video by

1. Segmenting color images to find the tool in each video frame
2. Estimating the tool shaft direction
3. Identifying the tool tip in the image
4. Computing the tool tip position in 3D coordinates

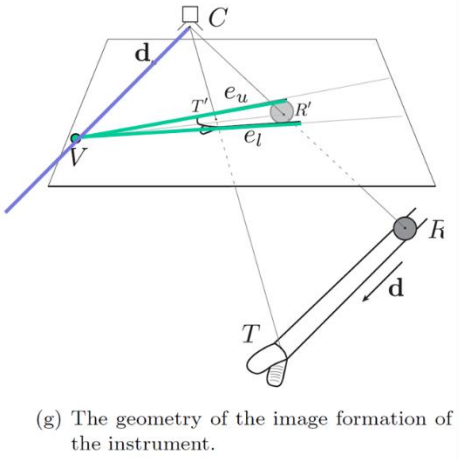
Tool Segmentation

The video frames were segmented to find the black tools. First, the video frames were thresholded to find the black pixels. Several erosion and dilations of the black regions were performed to find the tools in the image. The tool tips were identified as points along the tool shaft where there was an abrupt change in the color image and the image gradient. How closely the tool tip positions found in the color space and gradient space agreed was used as a confidence measure. A low confidence in the tool tip position in the image can be used to identify poor tracking due to tool occlusion or low image quality.

Estimating the Tool Shaft Direction

Next the tool shaft direction was found in 3D coordinates. Lines were fit to the tools segmented in the previous step. The intersection of the tool edges in the image is the vanishing point, V, of the tool. The vector from the camera to the tool vanishing point, $\vec{d}$,

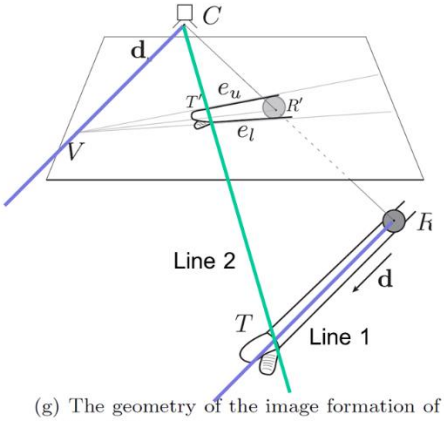
gives the direction of the tool in 3D coordinates. $\vec{d}$ can be computed as the unit vector of V: $\frac{V}{\|V\|}$.



(g) The geometry of the image formation of the instrument.

Figure 4. A drawing of the image of the tool and the tool in 3D coordinates. The green lines are along the tool shaft in the image, they intersect at the tool's vanishing point V. $\vec{d}$ points from the camera to V. Figure taken from [2] and annotated.

The trocar position is computed as the point closest to all the tool edge points in the video. The 3D tool tip position is the intersection of the vector from the camera to the tool tip in the image and the vector from the trocar in the direction $\vec{d}$. If the two vectors do not exactly intersect the tool tip is computed as the midpoint of the shortest line between the vectors.



(g) The geometry of the image formation of the instrument.

Figure 5. The tool tip is shown as the intersection between the line from the camera through the tool tip in the image and the line from the trocar (R) in the direction of $\vec{d}$. Figure taken from [2] and annotated.

Key Results

This tracking algorithm computes accurate 3D tool tip tracking using only one video. This tool tracking algorithm can be used as a foundation for a quantitative measure of laparoscopic skill.

Conclusion

This paper presents and interesting geometric method to find the 3D tool tip position from only one video frame. One weakness is there was no quantitative evaluation of the tracking algorithm. The paper merely says the results were "validated visually" [3]. I think quantitative results would be very useful for the 3D tool tracking results because their method is unique and only uses one video frame. It is unclear to me why some of the geometry they do is valid (why should the line from the camera to the tool vanishing point be parallel to the tool in the world?) so quantitative results would be valuable.

Another limitation of this work is that the tracking algorithm is only for a training environment. The algorithm does not address tracking problems in real surgical data like tracking in the presence of blood and water. A possible next step for this work would be to apply their tracking algorithm to real surgical video.

Relevance

This paper is very relevant to my project. I plan on implementing a similar tool tracking algorithm by segmenting the tool in the color image, finding the tool tip in each image, and triangulating the tool tip position. It is easier for me to find the 3D tool tip position because I can triangulate it using the stereo images. In the paper, the FLS has a fixed camera. My project has the added challenge of tracking the camera motion then tracking the tool motion.

Conclusion:

Both papers presented are very relevant to my CIS 2 project. The first paper relates to how I will find the background motion. The second paper presents an algorithm I will implement to find the tool tip positions in the microscope video. Both papers present interesting ways to find 3D geometry from one camera. A weakness in both papers is that the quantitative error between the tracked points and the real 3D points is unclear. Since an important aspect of my project is the tracking accuracy, I should give detailed error analysis comparing the triangulated points to the actual 3D points, the computed camera motion to the real motion, and the computed tool tip to the real tool tip.

References:

[1] S. Leonard, A. Reiter, A. Sinha, M. Ishii, R. Taylor, and G. Hager, "Image-Based Navigation for Functional Endoscopic Sinus Surgery Using Structure From Motion," in *SPIE*, San Diego, 2016.

[2] B. Allen, F. Kasper, G. Nataneli, E. Dutson, and P. Faloutos, "Visual Tracking of Laparoscopic Instruments in Standard Training Environments," in *MMVR,* Newport Beach, 2011.

[3] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek, "The Trimmed Iterative Closest Point algorithm," in *Pattern Recognition*, Québec City, 2002.