Image-Based Navigation for Functional Endoscopic Sinus Surgery Using Structure From Motion

Simon Leonard, Austin Reiter, Ayushi Sinha, Masaru Ishii, Russel H. Taylor and Gregory D. Hager

The Johns Hopkins University

ABSTRACT

Functional Endoscopic Sinus Surgery (FESS) is a challenging procedure for otolaryngologists and is the main surgical approach for treating chronic sinusitis, to remove nasal polyps and open up passageways. To reach the source of the problem and to ultimately remove it, the surgeons must often remove several layers of cartilage and tissues. Often, the cartilage occludes or is within a few millimeters of critical anatomical structures such as nerves, arteries and ducts. To make FESS safer, surgeons use navigation systems that register a patient to his/her CT scan and track the position of the tools inside the patient. Current navigation systems, however, suffer from tracking errors greater than 1 mm, which is large when compared to the scale of the sinus cavities, and errors of this magnitude prevent from accurately overlaying virtual structures on the endoscope images. In this paper, we present a method to facilitate this task by 1) registering endoscopic images to CT data and 2) overlaying areas of interests on endoscope images to improve the safety of the procedure. First, our system uses structure from motion (SfM) to generate a small cloud of 3D points from a short video sequence. Then, it uses iterative closest point (ICP) algorithm to register the points to a 3D mesh that represents a section of a patients sinuses. The scale of the point cloud is approximated by measuring the magnitude of the endoscope's motion during the sequence. We have recorded several video sequences from five patients and, given a reasonable initial registration estimate, our results demonstrate an average registration error of 1.21 mm when the endoscope is viewing erectile tissues and an average registration error of 0.91 mm when the endoscope is viewing non-erectile tissues. Our implementation SfM + ICP can execute in less than 7 seconds and can use as few as 15 frames (0.5 second of video). Future work will involve clinical validation of our results and strengthening the robustness to initial guesses and erectile tissues.

1. INTRODUCTION

With over 250,000 Functional Endoscopic Sinus Surgery (FESS) performed annually in the United States,[1] FESS has become an effective procedure to treat common conditions such as chronic sinusitis. This challenging minimally invasive procedure involves inserting a slender endoscope and tools through the nostrils to enlarge sinus pathways by removing small bones and cartilage. FESS is also often used to remove polyps (polypectomy) and straightening the septum (septoplasty).

To reach the source of the problem and to ultimately remove it, the surgeon removes layers of cartilage and tissues. These cartilages are adjacent to critical anatomical structures such as optic nerves, anterior ethmoidal and carotid arteries and nasolacrimal ducts and often occlude them. Accidental damages to these structures are a cause of major complications that can result in cerebrospinal fluid (CSF) leaks, blindness, oculomotor deficits and perioperative hemorrhage.

A meta-analysis by Labruzzo et al.[2] demonstrates that experience and enhanced imaging technologies have contributed to significantly decrease the rate of FESS complications. Yet, recent studies report major complication of FESS ranging between 0.31-0.47% of cases and minor complications ranging between 1.37-5.6%. To improve the safety and efficiency of FESS, surgeons use navigation systems that register a patient to his/her CT scan and track the position of the tools inside the patient. These systems are reported to decrease intraopertive time, improve the surgical outcome and reduce the workload.[3] Current navigation systems, however, suffer from

Further author information: (Send correspondence to Simon Leonard.)
Simon Leonard: E-mail: sleonard@jhu.edu

tracking errors greater than 1 mm, which is large when compared to the scale of the sinus pathways[4] and errors of this magnitude prevent from accurately overlaying virtual structures on the endoscope images.

In this paper, we propose an image-based system for enhanced FESS navigation. Our system enables a surgeon to asynchronously register a sequence of endoscope image to a CT scan and to overlay 3D structures that are segmented from the CT scan. Other advantages of our system are that 1) it is invariant to bending of the endoscope shaft and 2) the invariant to rotation about the optical axis of the endoscope. A known disadvantage of our system is its sensitivity between a possible discrepancy between a pre-operative CT and intra-operative images caused by congestive variations. As demonstrated in this paper, this concern can be addressed by ensuring that patients are decongested and decreasing the duration between the times a patient is scanned and examined.

FESS became widely adopted during the 1980s after the pioneer work of Messerklinger and Kennedy.[5] Although progress in imaging and navigation systems have contributed to decrease complications from 8% to 0.31%,[2] Krings et al.[6] report that image-guided FESS (IG-FESS) are more likely to have complications and mention that possible reasons for the increased rate of complications include overconfidence in the technology and using these technologies to treat complex cases.

The maximum registration error for IG-FESS that is most commonly found in the literature is 2 mm[7,8] and accuracy of less than 1.5 mm have been reported for modern navigation systems.[9] Recently, an image-based registration method achieved reprojection error 0.7 mm[10] but this methods require an initial registration to function. Our work is closely related to,[11] where a sparse 3D point cloud is computed from a sequence of endoscopic images and then registered to a 3D geometry derived from a CT scan. The system presented in this paper, however, estimates the 3D geometry from a greater number of images and improves feature matching.

With these considerations in mind, our system achieves sub-millimeter error on non-erectile tissues by using as few as 30 frames (1 second of video). On erectile tissues, the registration error increases to 1.21 millimeter.

2. METHOD

Our video-CT system uses a small sequence of endoscope images, typically between 15 and 30, to compute a 3D geometry using structure from motion (SfM) algorithm with sparse bundle adjustment (SBA). Then, the resulting 3D point cloud and the sequence of 3D camera poses are registered to the 3D geometry of a patient's sinus cavity that is derived from the CT data using iterative closest point (ICP) with scale adjustment. Our system is implemented with the client/server illustration of Fig. 1. Once the registration is computed, the sequence of camera poses is used in rviz to overlay the CT scan, or part thereof, on the camera images. Contrary to commercial image-guidance systems where only the 3D position of a tool is displayed in the three anatomical planes, our video-CT method enables an enhanced augmented reality by overlaying virtual structures, visible or occluded, on top of video images.

Our system is implemented by Robot Operating System (ROS)[12] services on a server with 20 cores (dual Xeon E5-2690 v2, Intel, Santa Clara CA) and 3 GPUs (GeForce GTX Titan Black, NVdia, Santa Clara, CA). GPUs and CPUs are organized in pools that are available to process images in parallel.

Endoscopic images, such as those used during FESS, present a unique challenge for SfM algorithms. First, the lens of the endoscope occludes approximately 50% of the imaging area leaving a relatively small circular area to project the foreground data. Second, the difficult lighting condition such as specularities, high dynamic range and the complex environment that are common during minimally invasive surgeries make feature matching and 3D reconstruction very challenging.

Several features and matching algorithms, such as scale invariant feature transforms (SIFT)[13] and adaptive scale kernel consensus (ASKS),[14] have been used for SfM[11] but the difficulty to obtain a reliable point cloud depends on robust feature matching which constrains the motion to a specific operating range. Recent advances such as Hierarchical Multi-Affine (HMA) algorithm[15] have demonstrated superior robustness for surgical applications and were used in our system. More specifically, our system uses HMA with Speeded Up Robust Features (SURF)[16] between each possible pair of images (Figure 2). Our main argument for the choice of SURF is the availability to extract key points and descriptors using GPUs with OpenCV.[17] Initial matches between image

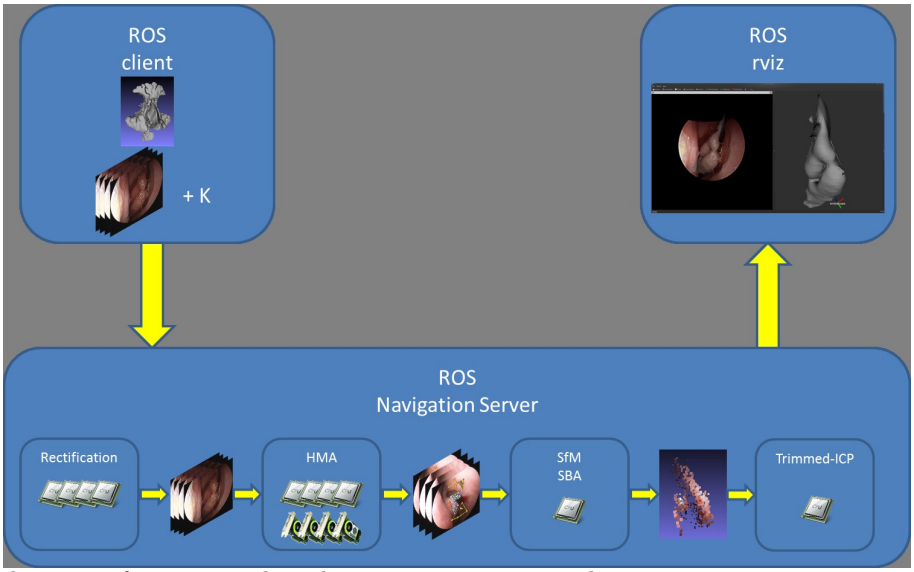


Figure 1: Block diagram of our image-based navigation system. The images are sent to a computing server to compute the structure from motion and to register the result to the patients CT scan.

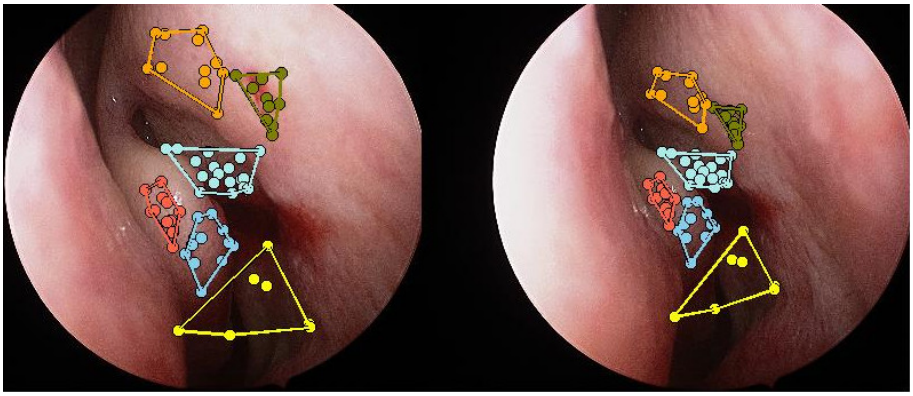


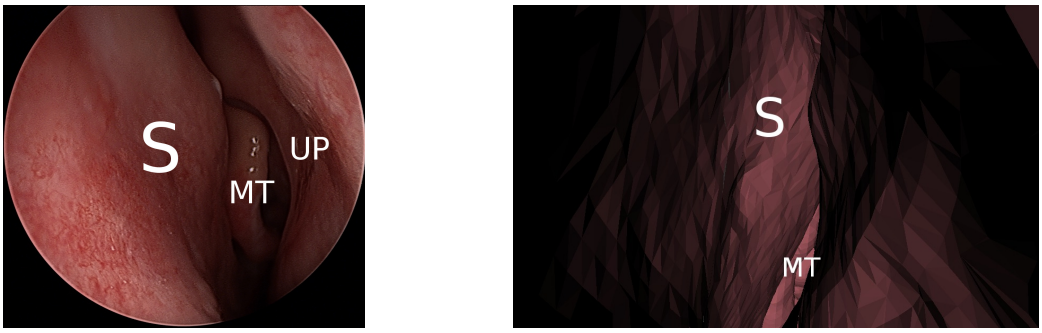
Figure 2: Hierarchical Multi-Affine matching algorithm from two views.

pairs key points and descriptors are extracted using the GPUs and the HMA matching algorithm is implemented in C++ without the recovery phase. The trimmed-ICP with scale uses non-linear Levenberg-Marquardt and is derived from the Point Cloud Library (PCL)[18] implementation.

The set of matches is then used to estimate the visible 3D structure and the camera motion[19] and then refine the estimate by computing a sparse bundle adjustment.[20] The resulting 3D structure and camera poses are defined up to an unknown scale but given that the motion of the endoscope is also tracked with a 6DOF magnetic tracker the unknown scale of the reconstruction is initially approximated from the magnitude of the endoscope's motion. Finally, trimmed iterative closest point (TriICP) algorithm with scale is used to align the cloud of points with a 3D mesh of the sinuses. The trimmed variant is necessary because the structure generally contains several outliers and only 70% to 80% of inliers are used in our experiments. The ICP implementation also adjust for the scale of the registration since the scale estimated by the magnetic tracker is inaccurate.

3. RESULTS

We collected data from several patients during a preoperative assessment (IRB NA_00074677).



(a) Decongested view of the middle turbinate.

(b) Rendering of the CT scan from a similar view. The middle turbinate is almost completely occluded by the septum.

Figure 4: Difference between a congested and decongested view of the middle turbinate. The middle turbinate is severely occluded by the septum on the CT scan. S — nasal septum, MT — middle turbinate, UP — uncinate process.

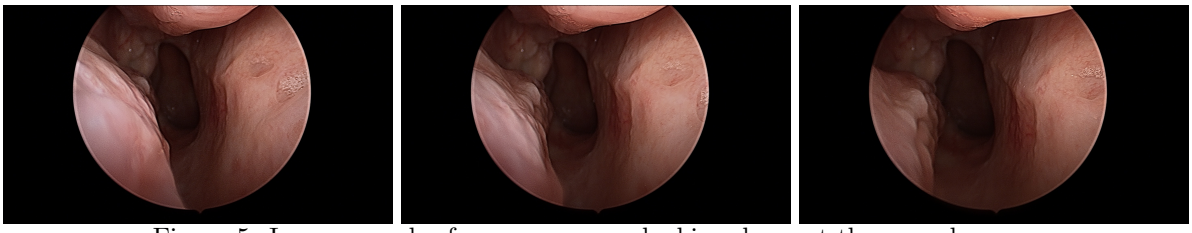


Figure 5: Image samples from a sequence looking down at the nasopharynx.

Figure 3: Data collection cart.

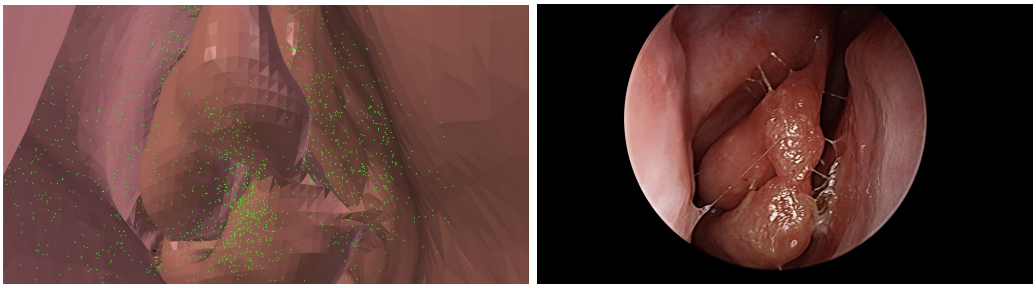
We designed a small cart (Figure 3) that holds a laptop, a DVI2USB 3.0 (Epiphan Video, Ottawa Canada) to collect 1920x1080 images at 30 frames per second, an Aurora magnetic tracking system (NDI Waterloo, Canada) and an isolation transformer. Within 60 seconds, the cart can be wheeled into a room, the video input connected to a 1288HD endoscopic camera (Stryker Kalamazoo, MI), the magnetic tracker clipped to the endoscope and the software initialized. During the data collection, the raw video images and the position/orientation of the magnetic reference are saved in ROS bag files.

On average, the data collection lasted 90 seconds during which the surgeon inserted the endoscope in both airways. Camera calibration was executed after each examination by using CALTag.[21]The middle turbinate of each patient was examined and in some cases the endoscope was pushed to the nasopharynx. Erectile tissues are found in several observed structures, such as the middle turbinate, nasal septum and the uncinate process and these structures can swell for various reasons. As illustrated in Figure 4, certain patients have significant swelling discrepancies between their CT scan and the video data. One explanation to this observation is the delay between the appointments for radiology and otorhinolaryngology and the use of decongestant in one but not the other. Each patient was examined in both airways resulting in 10 to 15 short video sequences (lasting between 0.5 and 1 second). Samples images are illustrated in Figure 5. These sequences were processed with the algorithms described in Section 2. The sequences generated an average of 910 3D points and the initial pose estimate for the TriICP was manually given.

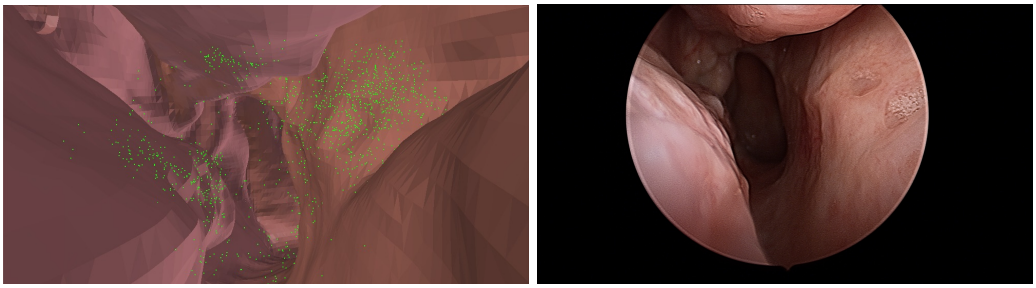
We selected five sequences with erectile tissues and five sequences without erectile tissues. No ground truth or reference is available to evaluate the accuracy of the registration for our *in vivo* data. Therefore, no target registration error (TRE)[22] can be computed for our experiments. The first result reported is the registration error e_{ICP} provided by TriICP. This error corresponds to the 70th percentile of the distance between a 3D point from the point cloud and the 3D mesh of the airways. These results are reported in Table 1. The results

Table 1: Average 70th percentile registration error

	Non-Erectile Tissues	Erectile Tissues
$\bar{e}_{ICP}$	0.91 mm (0.2 mm)	1.21 mm (0.3 mm)



(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.

demonstrate that our algorithm is able to register the non-erectile structures with less than a millimeter error and the erectile structures with an error slightly above one millimeter which is comparable to state-of-the-art navigation systems. Illustrations of two registered point clouds are shown in Figures 6a and 6b.

The main drawback of the previous registration result is that, although an accurate 3D registration is essential, 3D registration only represents an intermediate step in estimating the position of the endoscope. Since we do not have fiducial markers to estimate the error of the camera pose, we evaluate the error of our system by projecting a 3D structure that is easily segmented from the CT data on a binary image I_A. Then, we manually segment the corresponding structure in the image to obtain a second binary image I_B and we compute percentage of pixels that are incorrectly labeled by the operation $\sum(I_A \oplus I_B)/\sum I_B$. An example of overlay is illustrated in Figure 7. Since the middle turbinate is easily segmented and is visible in all the video sequences, we use it to evaluate the error and we obtained an average of 86% with a standard deviation of 3%.

4. CONCLUSION

This paper introduced an enhanced navigation for endoscopic sinus surgery. First, the method is based on obtaining a sparse 3D reconstruction of the airways from a few images. Second, the sparse 3D point cloud is registered to the 3D model of the airways. For non-erectile tissues, the method is able to register 70% of the points within 0.91 mm of the CT scan data and 1.21 mm for erectile tissues. On average, the method is able to overlay 86% of the middle turbinate on manually segmented images of the airways. Using 15 frames, the computation time for the registration is less than 7 seconds and 10 seconds for 30 frames. In comparaison, we observed that a surgeon can take as much as 30 seconds to use a 3D pointer for localization since this procedure requires to 1) remove a tool from the airways 2) insert a tracked pointer 3) localize the pointer tip in the images and 4) insert a tool.

This work has not been submitted or presented elsewhere.

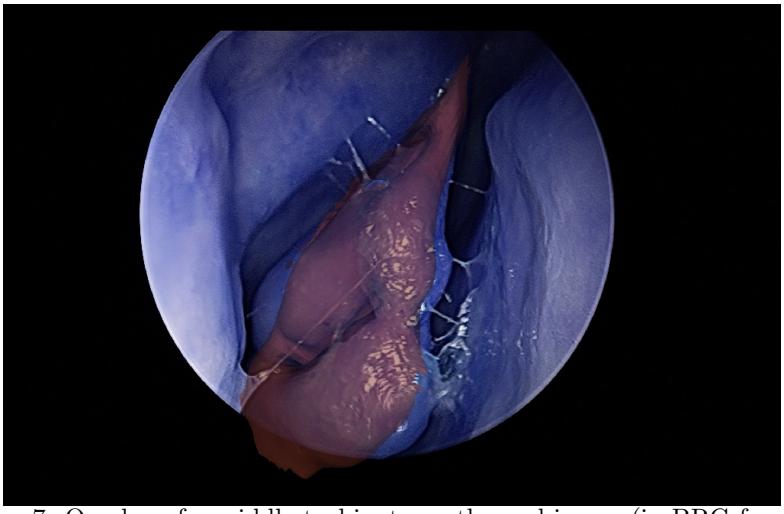


Figure 7: Overlay of a middle turbinate on the real image (in BRG format).

REFERENCES

[1] Rosenfeld, R. M., Piccirillo, J. F., Chandrasekhar, S. S., Brook, I., Ashok Kumar, K., Kramper, M., Orlandi, R. R., Palmer, J. N., Patel, Z. M., Peters, A., Walsh, S. A., and Corrigan, M. D., "Clinical practice guideline (update): adult sinusitis," *Otolaryngology–Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery* **152**, S1–S39 (Apr. 2015).

[2] Labruzzo, S. V., Aygun, N., and Zinreich, S. J., "Imaging of the Paranasal Sinuses: Mitigation, Identification, and Workup of Functional Endoscopic Surgery Complications," *Otolaryngologic Clinics of North America* **48**, 805–815 (Oct. 2015).

[3] Strauss, G., Limpert, E., Strauss, M., Hofer, M., Dittrich, E., Nowatschin, S., and Lth, T., "[Evaluation of a daily used navigation system for FESS]," *Laryngo- rhino- otologie* **88**, 776781 (December 2009).

[4] Lorenz, K., Frhwald, S., and Maier, H., "The use of the BrainLAB Kolibri navigation system in endoscopic paranasal sinus surgery under local anaesthesia. An analysis of 35 cases," *HNO* **54**(11), 851–860 (2006).

[5] Kennedy, D. W., "Functional endoscopic sinus surgery. Technique," *Archives of Otolaryngology (Chicago, Ill.: 1960)* **111**, 643–649 (Oct. 1985).

[6] Krings, J., Kallogjeri, D., Wineland, A., Nepple, K., Piccirillo, J., and Getz, A., "Complications of primary and revision functional endoscopic sinus surgery for chronic rhinosinusitis," *Laryngoscope* **124**(4), 838–845 (2014). cited By 1.

[7] Otake, Y., Leonard, S., Reiter, A., Rajan, P., Siewerdsen, J. H., Gallia, G. L., Ishii, M., Taylor, R. H., and Hager, G. D., "Rendering-Based Video-CT Registration with Physical Constraints for Image-Guided Endoscopic Sinus Surgery," *Proceedings of SPIE–the International Society for Optical Engineering* **9415** (Feb. 2015).

[8] Al-Swiahb, J. N. and Al Dousary, S. H., "Computer-aided endoscopic sinus surgery: a retrospective comparative study," *Annals of Saudi Medicine* **30**(2), 149–152 (2010).

[9] Paraskevopoulos, D., Unterberg, A., Metzner, R., Dreyhaupt, J., Eggers, G., and Wirtz, C. R., "Comparative study of application accuracy of two frameless neuronavigation systems: experimental error assessment quantifying registration methods and clinically influencing factors," *Neurosurgical Review* **34**, 217–228 (Apr. 2010).

[10] Mirota, D., Wang, H., Taylor, R., Ishii, M., Gallia, G., and Hager, G., "A System for Video-Based Navigation for Endoscopic Endonasal Skull Base Surgery," *IEEE Transactions on Medical Imaging* **31**, 963–976 (Apr. 2012).

[11] Mirota, D. J., Uneri., A., Schafer, S., Nithiananthan, S., Reh, D. D., Ishii, M., Gallia, G. L., Taylor, R. H., Hager, G. D., and Siewerdsen, J. H., "Evaluation of a System for High-Accuracy 3d Image-Based Registration of Endoscopic Video to C-Arm Cone-Beam CT for Image-Guided Skull Base Surgery," *IEEE Transactions on Medical Imaging* **32** (July 2013).

[12] Quigley, M., Conley, K., Gerkey, B. P., Faust, J., Foote, T., Leibs, J., Wheeler, R., and Ng, A. Y., "Ros: an open-source robot operating system," in [*ICRA Workshop on Open Source Software*], (2009).

[13] Lowe, D. G., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision* **60**, 91–110 (Nov. 2004).

[14] Wang, H., Mirota, D., and Hager, G. D., "A Generalized Kernel Consensus-Based Robust Estimator," *IEEE transactions on pattern analysis and machine intelligence* **32** (Jan. 2010).

[15] Puerto-Souza, G. and Mariottini, G.-L., "A fast and accurate feature-matching algorithm for minimally-invasive endoscopic images," *Medical Imaging, IEEE Transactions on* **32**, 1201–1214 (July 2013).

[16] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L., "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.* **110**, 346–359 (June 2008).

[17] Bradski, G. *Dr. Dobb's Journal of Software Tools* (2000).

[18] Rusu, R. B. and Cousins, S., "3D is here: Point Cloud Library (PCL)," in [*IEEE International Conference on Robotics and Automation (ICRA)*], (May 9-13 2011).

[19] Hartley, R. I. and Zisserman, A., [*Multiple View Geometry in Computer Vision*], Cambridge University Press, ISBN: 0521540518, second ed. (2004).

[20] Triggs, B., McLauchlan, P. F., Hartley, R. I., and Fitzgibbon, A. W., "Bundle adjustment - a modern synthesis," in [*Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*], *ICCV '99*, 298–372, Springer-Verlag, London, UK, UK (2000).

[21] Atcheson, B., Heide, F., and Heidrich, W., "CALTag: High precision fiducial markers for camera calibration," in [*15th International Workshop on Vision, Modeling and Visualization*], (November 2010).

[22] Mirota, D. J., Uneri, A., Schafer, S., Nithiananthan, S., Reh, D., Ishii, M., Gallia, G., Taylor, R., Hager, G., and Siewerdsen, J., "Evaluation of a system for high-accuracy 3d image-based registration of endoscopic video to c-arm cone-beam ct for image-guided skull base surgery," *Medical Imaging, IEEE Transactions on* **PP**(99), 1 (2013).