Direct Endoscopic Video Registration for Sinus Surgery

Daniel Mirota*[a], Russell H. Taylor[a], Masaru Ishii[b], Gregory D. Hager[a]

[a]Department of Computer Science, Johns Hopkins University, Baltimore, MD, 21218;
[b]Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins Bayview Medical Center, Baltimore, MD, 21224

ABSTRACT

Advances in computer vision have made possible robust 3D reconstruction of monocular endoscopic video. These reconstructions accurately represent the visible anatomy and, once registered to pre-operative CT data, enable a navigation system to track directly through video eliminating the need for an external tracking system. Video registration provides the means for a direct interface between an endoscope and a navigation system and allows a shorter chain of rigid-body transformations to be used to solve the patient/navigation-system registration. To solve this registration step we propose a new 3D-3D registration algorithm based on Trimmed Iterative Closest Point (TrICP)[1] and the z-buffer algorithm.[2] The algorithm takes as input a 3D point cloud of relative scale with the origin at the camera center, an isosurface from the CT, and an initial guess of the scale and location. Our algorithm utilizes only the visible polygons of the isosurface from the current camera location during each iteration to minimize the search area of the target region and robustly reject outliers of the reconstruction. We present example registrations in the sinus passage applicable to both sinus surgery and transnasal surgery. To evaluate our algorithm's performance we compare it to registration via Optotrak and present closest distance point to surface error. We show our algorithm has a mean closest distance error of .2268mm.

Keywords: Registration, Endoscope Procedures, Localization and Tracking Technologies

1. INTRODUCTION

Sinus surgery and transnasal surgery demand high accuracy. Transnasal pituitary surgery is an procedure where the sinus passages are used to gain access to the pituitary gland.[3] Accuracy is crucial while operating in the sinus passages to ensure critical patient anatomy, such as the optic nerve and carotid artery, are preserved. Pituitary lesions transnasal surgery is shown to both reduce surgical time and recovery time.[4] Nasseri et al.[4] also drew attention to the need of navigation, especially to aide junior surgeons and for complex cases.

Today, the procedure uses both a navigation system and an endoscopic video system. The navigation system is used to perform real-time tracking of instruments and the endoscope. Endoscopic video system provides display of the surgical field.

Navigation systems provide the surgeon orientation and progress monitoring. The systems come in a variety of types including optical and electromagnetic. The optical and electromagnetic systems track via cameras and field generators, respectively. Optical navigation systems provide the highest level of accuracy available, but require clear line of sight to the instruments. In contrast electro-magnetic(EM) navigation systems require no line of sight, but the EM field can be distorted by metal present in the work area. Both navigation systems register to a pre-operative CT by identifying fiducial markers in the CT and on the patient. Optical navigation, specifically the Optotrak (Northern Digital, Waterloo, Ont.), is preferred in computer-integrated surgery applications for its high global accuracy and robustness.[5]

Endoscopic video systems simply display the surgical field. However, video data offer a rich amount of information, including geometric and photometric properties of the scene that can be used to enhance the surgeon's view. The computer vision literature addresses many different methods for processing video, including structure from motion,[6] that enables a 3D reconstruction to be created from video. Once reconstructed, the video data can be registered to pre-operative 3D data allowing both to be visualization in the same view.

*dan@cs.jhu.edu

Registration has been well studied in the fields of computer vision and medical image processing. Many variations on the classic Iterative Closest Point (ICP) method[7] have been developed; these methods offer robust registration techniques that work in the presence of large numbers of outliers. Trucco et al.[8] presented a Least Median Squares variation of ICP that ensured robustness with up to 50% outliers. Subsequently, Chetverikov et al.[1] described a Least Trimmed Squares variation of ICP that offered robustness with greater than 50% outliers. More recently, Estepar et al.[9] reported a solution to simultaneously solve the rigid-body transformation and model noise in the data that enables robustness to anisotropic noise and outliers.

It would be ideal to have registration between the patient and pre-operative data without the encumbrance of the associated external tracking system. In this paper, we propose a method to directly register a 3D reconstructed structure from video to CT. This method does not require fiducial markers and can be performed within the standard practice for sinus and transnasal surgeries. Furthermore, there is evidence that such systems can potentially register with higher accuracy than current navigation technologies.[10]

2. METHODS

The proposed algorithm relies on Trimmed Iterative Closest Point (TrICP)[1] and the z-buffer algorithm.[2] Registration algorithms, such as ICP[7] and variations consider the entire target model at once. Here we take advantage of the fact that a reconstructed video sequence can only be made of visible points. Thus, we assume only visible points need to be considered in the registration process.

2.1 Algorithms

Our algorithm requires three inputs: a 3D point cloud of relative scale with the origin at the camera center, an isosurface from the CT, and an initial guess of the scale and location. First, the 3D point cloud of relative scale with origin at the camera center is the expected output of the 3D reconstruction process. Here the method by Wang et al.[11] is used to create a sparse reconstruction. We assume that 3D point cloud is of some uniform scale that need not be the same as the CT and that the origin be the camera center such that the rigid-body transformation aligning the 3D point cloud is the camera location in CT coordinates. Outliers typically contaminate the 3D reconstruction as a result of mismatched features and the ambiguity of sign in the epipolar constraint. Second, the isosurface from the CT is used to simplify the registration progress. While using only a surface does remove a large amount of data, there is sufficient data in the surface alone for registration. The isosurface is created by applying a threshold to the CT data at the air/surface boundary. Third, an initial guess of the location and scale, similar to ICP,[7] TrICP is prone to falling into local minima and an approximate solution is needed to start the algorithm.

The inputs are processed as follows. The isosurface from CT, represented as a polygons mesh, is loaded in the a renderer. The rendering is created from the initial camera location. The visible polygons are then fed into a kD-tree for TrICP. TrICP then solves for the rigid-body transformations and scale. The new camera location is fed back to the renderer and the process continues to convergence. Algorithm 1 shows a pseudo-code for the overall system. The algorithm uses the following variables: M to be the model polygon mesh, D to be the 3D reconstructed points and $R, \mathbf{t}$ and s to be the rotation, translation and scale relating the two, respectively.

TrICP was chosen for the algorithm's robustness to outliers and simple design. Our algorithm modifies the existing TrICP algorithm by adding scale. This modification is similar to the work of Du et al.[12] However, we assume a single uniform scale. The following is how the modification to TrICP is derived.

Let $X \in \Re^{3\times n}$ be the matrix of all the reconstructed points as column vectors. Let $B \in \Re^{3\times n}$ be the matrix of the corresponding closest points on the model as column vectors. Where n is the number of points. We then compute:

$$X_{zero} = X - mean(X) \quad B_{zero} = B - mean(B) \tag{1}$$

$$C_X = \frac{1}{n} X_{zero} X_{zero}^T \quad C_B = \frac{1}{n} B_{zero} B_{zero}^T \tag{2}$$

Algorithm 1 (R,t,s) = TrICPzbuffer($R_{init}, \mathbf{t}_{init}, s_{init}, M, D$)

```
$R \leftarrow R_{init}, \quad \mathbf{t} \leftarrow \mathbf{t}_{init}, \quad s \leftarrow s_{init}$
**while** Not Converged **do**
  $m = render(M, R, t)$
  Create kD-tree($m$)
  $x \leftarrow R * s * d + \mathbf{t}$
  $e_i \leftarrow \|b - x\|$
  **for all** $d \in D$ **do**
    $b \in B, b \leftarrow$kD-tree.$closest\_point(d)$
    $e_i \in e, e_i \leftarrow \|d - b\|$
  **end for**
  $e^{sorted} = sort(e)$
  $inliers = \underset{\alpha \in [0.4, 1.0]}{\operatorname{argmin}} \sum_{1}^{floor(\alpha * n)} e^{sorted} \alpha^{-5}$
  $K \leftarrow \forall i$ s.t. $i < inliers * n$
  $x \in X, x \leftarrow R * s * d + \mathbf{t}$
  $(R, \mathbf{t}, s) = registerPointSet(X(K), B(K))$
**end while**
```

Let λ_X and λ_B be vectors of the eigenvalues of C_X and C_B, respectively. We can interpret the eigenvalues of each data set as the diagonalized covariance matrix of the points. As a result, under perfect conditions, the following relationship holds between the data sets:

$$s^2 \lambda_X = \lambda_B \tag{3}$$

Thus, we can estimate the unknown scale factor as:

$$s = \sqrt{\frac{\lambda_X \cdot \lambda_B}{\lambda_X \cdot \lambda_X}} \tag{4}$$

The resulting modified pose computation is shown in Algorithm 2.

Algorithm 2 (R,t,s) = registerPointSet(X,B)

```
$X_{zero} \leftarrow X - mean(X), \quad B_{zero} \leftarrow B - mean(B)$
$C_X \leftarrow \frac{1}{n} X_{zero} X_{zero}^T, \quad C_B \leftarrow \frac{1}{n} B_{zero} B_{zero}^T$
$s \leftarrow \sqrt{\frac{\lambda_X \cdot \lambda_B}{\lambda_X \cdot \lambda_X}}$
$H \leftarrow \frac{1}{n} X_{zero} B_{zero}^T$
$U \Sigma V^T = H$
$R \leftarrow U V^T$
**if** det(R) = -1 **then**
  $V \leftarrow [\mathbf{v}_1, \mathbf{v}_2, -\mathbf{v}_3]$
**end if**
$R \leftarrow U V^T$
$t \leftarrow mean(B) - s * R * mean(X)$
```

The z-buffer algorithm efficiently determines the visible polygons that are the exact surface the reconstruction needs to be registered. The use of the z-buffer algorithm allows the entire sinus model to be loaded only once and enables registration to any section of the sinus.

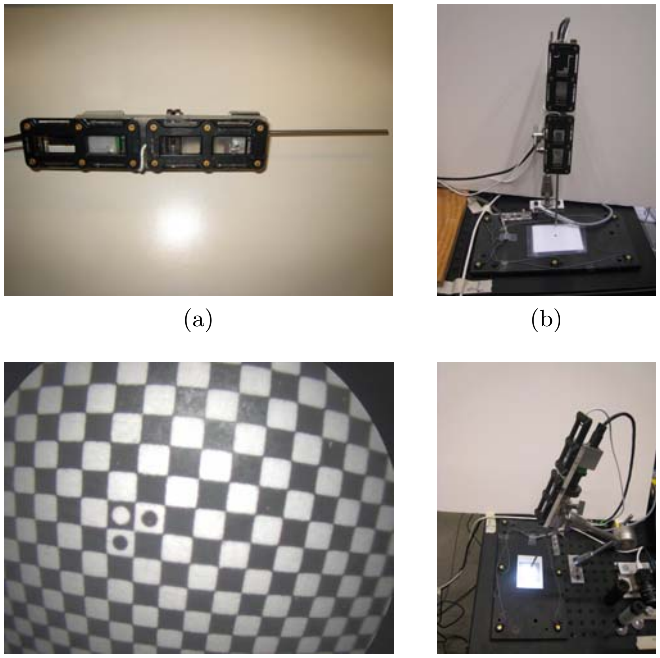
2.2 Experiments

We test our algorithm against noise and outliers with both simulated and real data from a skull phantom and porcine study. The simulated data consists of a mesh box and a random point sampling of the mesh. Added to the point cloud are both low level noise and large outliers. For the real data the following experimental protocol was used.

2.2.1 Calibration

The endoscope is fixed from rotation and is rigidly attached to camera. An Optotrak rigid-body is attached to the camera-endoscope pair. Figure 1a shows a picture of the completed assembly.

To ensure an accurate comparison between our algorithm versus the Optotrak, calibration is crucial. First, the video signal and Optotrak signal need to synchronized. To synchronize the video and Optotrak, the two are calibrated for their phase difference. This phase difference calibration is measured but sending a periodic signal through the system. The periodic signal is created by setting the endoscope on a pendulum and recording the motion of a spot. The waveform of the both the motion in the video and the Optotrak are compared to measure the phase difference. Figure 1b shows the experimental setup. Note the light source cable was removed to reduce cable drag during the calibration. The calibration was repeated three times and the phase difference each computed. Phase difference was found to be different in each trial. Since the phase difference are not consistent the phase is recalibrated just before entering the nose by finding the phase difference that minimizes the error between the Optotrak and the 2D-3D registration. Second, the endoscope requires calibration of the lens distortion. The endoscope is camera calibrated with the German Aerospace Center (DLR) Camera Calibration Toolbox[13] which offers an optimal solution to correct the radial distortion of endoscopes. In figure 1c we see the three circles that define the coordinate frame of the calibration grid and allow corners to be detected in the entire view. Another advantage of the toolbox is it's implementation of an optimal hand-eye calibration solution. To avoid any potential difference in phase between the endoscope and optotrak, the endoscope was fixed in a passive arm during calibration. Figure 1d shows the calibration setup.



(a)

(b)

(c)

(d)

Figure 1: Endoscope setup: 1a complete endoscope/Optotrak assembly, 1b phase calibration setup, 1c an example calibration image, 1d passive arm for calibration.

2.2.2 Setup

For each data collection the Optotrak was positioned two meters from the working area. The porcine head is secured in the workspace with a rigid-body attached. Similarly, the phantom is secured in the workspace. Then the skull or phantom is registered to the tracker by recording the locations of each fiducial marker.

Next, the endoscopic video is recorded. At least four fiducial markers outside of the head are imaged for use with the 2D-3D registration algorithm. The recording proceeds then into the nose. Once at the back of the nasal passage a side to side motion is recorded. This set of motions is similar to that of actual sinus surgery, where surgeons first survey the patient's anatomy.

3. RESULTS

3.1 Simulation

From the simulated data we found that our registration algorithm tolerates both noise and outliers. Figure 2 from left to right shows the effect of different levels of noise on the translation and the first and last iteration of an experiment with 30% low level noise and 30% outliers. The algorithm was run for 30 trials with random noise and outliers and recovered the transform within 0.0073 (STD=0.0122) degrees rotation and 0.0501 (STD=0.0564) units translation. The simulated data was without metric units. In figure 3 we show the convergence basins of with respect to rotation, translation and scale when tested independently. The graphs show the result of 100 samples each. Figure 3a shows rotation about a random axis of magnitudes from $-\pi$ to π. Figure 3b shows translation in a random direction of magnitudes from 0 to -10. The error is the product of the actual transformation and the estimated transformation. Let $F_{actual} = \begin{bmatrix} R_{actual} * s_{actual} & \mathbf{t}_{actual} \\ \mathbf{0} & 1 \end{bmatrix}$ and the estimate $F_{est} = \begin{bmatrix} R_{est} * s_{est} & \mathbf{t}_{est} \\ \mathbf{0} & 1 \end{bmatrix}$

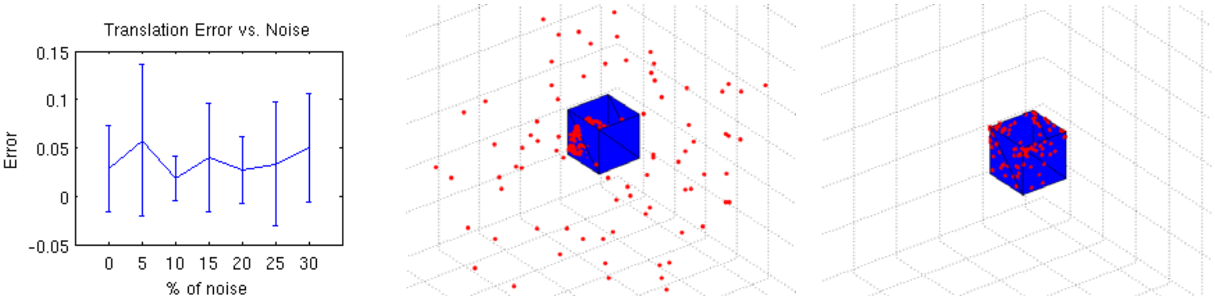


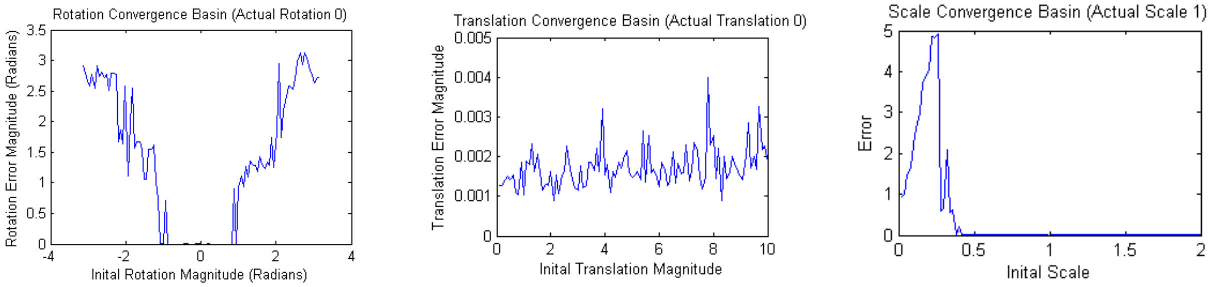
Figure 2: The plot of translation error versus noise (left), first (middle) and last (right) iteration of simulated data.

The error is therefore $F_{error} = F_{actual} * F_{est}$. The algorithm estimates the inverse of the actual transformation. The rotation error is the l_2-norm of the Rodrigues vector of the rotation component of F_{error}. The translation error the l_2-norm of the translation component of F_{error}. The scale error is defined as $\left| \frac{1}{s_{est}} - s_{actual} \right|$.

Figure 3a shows if the rotation is initialized within [-50.4, 46.8] degrees of the actual rotation the algorithm converges to the true rotation. When the model match the data well, figure 3b shows the translation can be found with any initial guess. Note the scale of the y-axis figure 3b is in thousandths. Similarly, in figure 3c we see that if the scale is initialized within .38 of the actual scale or over estimates the scale the algorithm converges to the true scale.

3.2 Real Data

Table 1 shows the absolute registration difference of the example registration presented in figures 4 and 5. Baseline/scale difference in table 1 is the absolute difference of the estimated scale and the baseline of the images measured by the Optotrak. Since the true distance between image pairs is only recovered up to scale in monocular images, the scale and the distance between the image pair should be the same. Our result shows that the relative scale was accurately recovered. Table 2 shows the closest distance to surface error. Figures 4 and 5 we present



(a) (b) (c)

Figure 3: The plot of initial rotation versus rotation error (left), initial translation versus translation error (middle) and initial scale versus scale error (right)

the results of endoscopic images and the 3D mesh overlay compared to the Optotrak from the skull phantom study and from the porcine study, respectively. Though there is a discrepancy between our algorithm and the Optotrak, visually our algorithm more closely reflects the actually camera location.

Table 1: Absolute registration difference of video registration and Optotrak

	Rotation (degrees)	Translation (mm)	Baseline/scale difference
Phantom study	5.4679	1.8204	0.2724
Porcine study	3.8174	2.0281	0.0675

Table 2: Closest distance to surface error

	Median (mm)	Mean (mm)	Inliers	Number of points
Phantom study	0.4078	0.5623	99%	106
Porcine study	0.2176	0.2268	63%	230

4. CONCLUSIONS

We have shown that our proposed modified version of TrICP registers and finds the scale of a 3D reconstruction from endoscopic video data of the sinuses. After registration the rendered results are similar to that of the original images. However, is it clear when the number of reconstructed point is small the registration is poorly constrained. While the absolute minimum number of points is three points a large number of reconstructed points is needed to have an accurate representation of the view. Dense reconstruction answers this need for more reconstruction data and is one future direction.

Our registration algorithm provides the initial step for navigation. After the 3D-3D correspondence is established a more efficient 2D-3D tracking algorithm to robustly maintain images feature would enable real-time performance. The efficient 2D-3D tracking algorithm would switch the current offline processing to online processing and is another future direction. The results here are the first steps toward a system that interfaces endoscopic video directly with a navigation system. The ultimate goal is to provide a system that will give surgeons a streamlined easy-to-use tool that will provide access to the data they need, thereby improving surgical outcomes and reducing cost.

ACKNOWLEDGMENTS

This work has been supported by the National Institutes of Health under grant number 1R21EB005201 - 01A1.

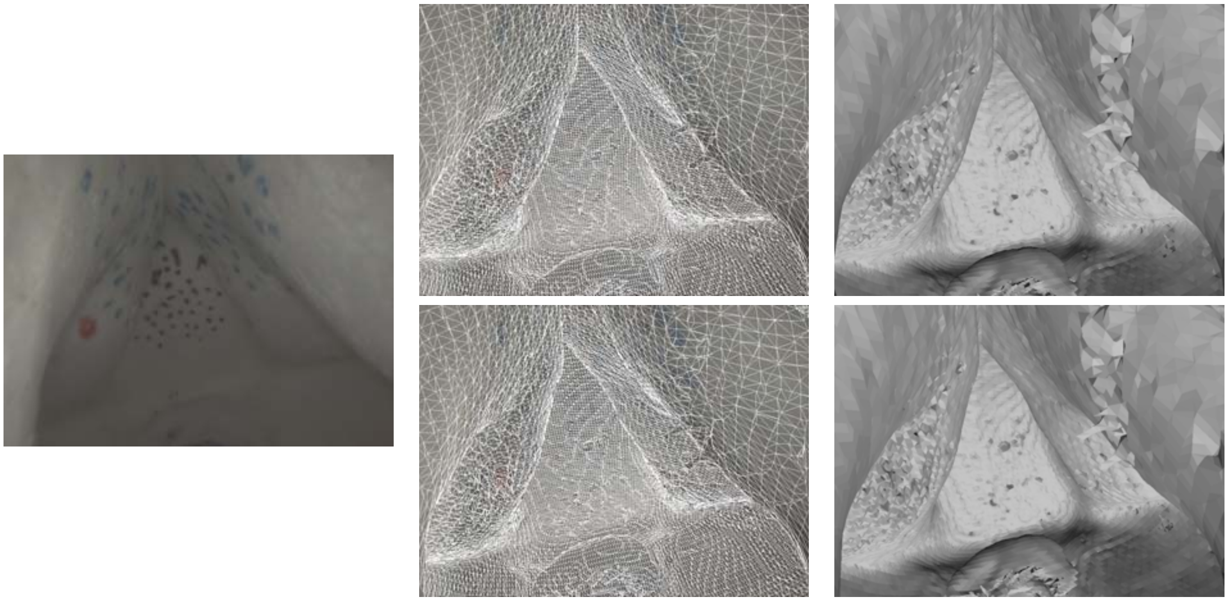


Figure 4: Phantom study example registration comparison. The first column is the undistorted endoscopic image. The second column shows the result of both our registration and the Optotrak as a mesh overlay, top and bottom respectively. The third column shows the rendered CT.

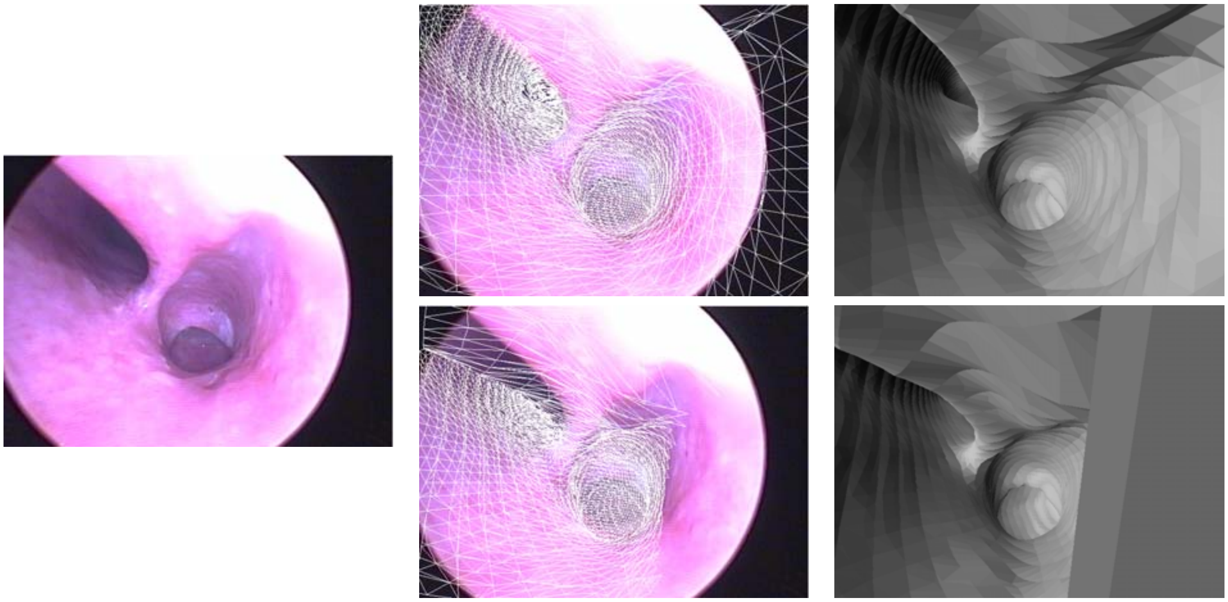


Figure 5: Porcine study example registration comparison. The first column is the undistorted endoscopic image. The second column shows the result of both our registration and the Optotrak as a mesh overlay, top and bottom respectively. The third column shows the rendered CT.

REFERENCES

[1] Chetverikov, D., Svirko, D., Stepanov, D., and Krsek, P., "The trimmed iterative closest point algorithm," *Pattern Recognition, 2002. Proceedings. 16th International Conference on* **3**, 545–548 vol.3 (2002).

[2] Catmull, E. E., *A subdivision algorithm for computer display of curved surfaces.*, PhD thesis, The University of Utah (1974).

[3] Carrau, R. L., Jho, H.-D., and Ko, Y., "Transnasal-transsphenoidal endoscopic surgery of the pituitary gland," *The Laryngoscope* **106(7)**, 914–918 (1996).

[4] Nasseri, S. S., Kasperbauer, J. L., Strome, S. E., McCaffrey, T. V., Atkinson, J. L., and Meyer, F. B., "Endoscopic transnasal pituitary surgery: Report on 180 cases," *American Journal of Rhinology* **15**, 281–287(7) (July-August 2001).

[5] Chassat, F. and Lavalle, S., "Experimental protocol of accuracy evaluation of 6-d localizers for computer-integrated surgery: Application to four optical localizers," in [*Medical Image Computing and Computer-Assisted Interventation MICCAI98*], 277 – 284 (1998).

[6] Longuet-Higgins, H., "A computer algorithm for reconstructing a scene from two projections," *Nature* **293**, 133–135 (September 1981).

[7] Besl, P. and McKay, H., "A method for registration of 3-d shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **14**, 239–256 (Feb. 1992).

[8] Trucco, E., Fusiello, A., and Roberto, V., "Robust motion and correspondence of noisy 3-d point sets with missing data," *Pattern Recognition Letters* **20**, 889–898 (September 1999).

[9] Estepar, R. S. J., Brun, A., and Westin, C.-F., "Robust generalized total least squares iterative closest point registration," in [*Seventh International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'04)*], *Lecture Notes in Computer Science* (September 2004).

[10] Burschka, D., Li, M., Ishii, M., Taylor, R. H., and Hager, G. D., "Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery," *Medical Image Analysis* **9**, 413–426 (October 2005).

[11] Wang, H., Mirota, D., Ishii, M., and Hager, G., "Robust motion estimation and structure recovery from endoscopic image sequences with an adaptive scale kernel consensus estimator," in [*Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*], *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* , 1–7 (June 2008).

[12] Du, S., Zheng, N., Ying, S., You, Q., and Wu, Y., "An extension of the icp algorithm considering scale factor," *Image Processing, 2007. ICIP 2007. IEEE International Conference on* **5**, 193–196 (2007).

[13] Strobl, K. H. and Hirzinger, G., "Optimal hand-eye calibration," in [*Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*], *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on* , 4647–4653 (Oct. 2006).