Scale-Invariant Registration of Monocular Endoscopic Images to CT-Scans for Sinus Surgery

Darius Burschka[1], Ming Li[2], Russell Taylor[2], and Gregory D. Hager[1]

[1] Computational Interaction and Robotics Laboratory, CIRL
The Johns Hopkins University, Baltimore, USA
{burschka,hager}@cs.jhu.edu
[2] Computer Integrated Surgical Systems and Technology, CISST
The Johns Hopkins University, Baltimore, USA
{liming,rht}@cs.jhu.edu

Abstract. We present a scale-invariant registration method for 3D structures reconstructed from a monocular endoscopic camera to pre-operative CT-scans. The presented approach is based on a previously presented method [2] for reconstruction of a scaled 3D model of the environment from unknown camera motion. We use this scaleless reconstruction as input to a PCA-based algorithm that recovers the scale and pose parameters of the camera in the coordinate frame of the CT scan. The result is used in an ICP registration method to refine the registration estimates.
The presented approach is used for localization during sinus surgeries. It simplifies the navigation of the instrument by localizing it relative to the CT scan that was used for pre-operative procedure planning.
The details of our approach and the experimental results with a phantom of a human skull are presented in this paper.

1 Introduction

Surgery of the frontal sinus can be performed endonasally or through an external approach. In the external approach, with the patient under general anesthesia, the surgeon makes an incision behind the hairline or under the eyebrows. This approach requires large skin incisions and protracted long recovery time. The endonasal approach for surgical treatment of frontal Sinusitis has become increasingly established during the last few years. All information is provided primarily through the endoscope requiring from the surgeon a detailed knowledge of the anatomy. Therefore, Computer Integrated Surgery techniques have been employed in endonasal approach to simplify the procedure. After a registration process, the surgeon can point at a specific structure in 3D and then view the position of the instrument tip in the pre-operative CT-scan [3,4].

Patient motion, imperfections in the surgical instrument or shifts of the reference frame may cause a registration error to a pre-operative registration. Therefore, currently the operating surgeon must update the registration at several

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3217, pp. 413–421, 2004.

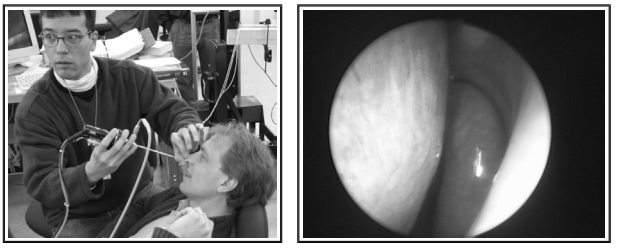

Fig. 1. Endoscopic inspection of the nasal sinus cavities depicting the limited information provided to the surgeon in the current procedures.

points in the operative field throughout the procedure. Typically, registration is verified by localization on known bony landmarks on the skull and in the nasal cavity. In image-guided sinus surgery, registration is performed based on fiducial points or based on the surface parameters [3]. In surface based methods, a probe is used to touch to contours to gather 3D surface data. Typical contours include the medial brow, nasal dorsum, and tragi. In fiducial point based method, radiopaque markers attached to both the patient's face and anatomic landmarks are used.

In this paper, we present a novel method for intra-operative registration directly from the endoscopic images without manual inputs from the surgeon. It is especially useful in revision cases, where the surgical landmarks are usually absent. The paper is structured as follows. In Section 2, we describe the underlying image processing that allows us to recover the 3D-structure and the motion from monocular images of an endoscopic camera and the way we perform the final alignment using a modified ICP approach. In Section 3, we present the experimental results on the phantom skull. We conclude in Section 4 with an evaluation of the presented approach and present our future research goals.

2 Approach

The two major problems that we address in this paper are: 3D reconstruction from monocular camera images and registration of the reconstructed 3D model to a pre-operative CT scan.

Our system reconstructs a scaled 3D model of the environment from a monocular camera. This reconstruction requires knowledge about the motion of the camera, which we assume to be unknown or at least uncertain. That means that, in parallel to model reconstruction, we need to estimate the motion of the camera as well. We discuss the implementation of the vision-based reconstruction in Section 2.1.

The 3D structure estimated from camera images is known up to scale. The correct scale needs to be recovered from the data to align the points roughly with the CT scan. The remaining alignment error between the CT scan data and the reconstructed model is corrected with our modified *Iterative Closest Point* (ICP) estimation with covariance tree optimization (Section 2.2).

2.1 Scaled 3D Reconstruction

Feature Extraction. The algorithm described below assumes that *point features* are extracted from the images. Possible features are: intersections of contours resulting from edge filters [7] or the areas themselves used for template matching in *Sum of Square Differences* (SSD) matching algorithms [5].

The problem in real endonasal images is the sparse density of points that actually can be used for a model reconstruction. Another problem is the moving light source, which is attached to the endoscope (Fig. 2). This violates the brightness constancy assumption used in most common stereo algorithms and thus forces us to switch to a brightness independent image representation.

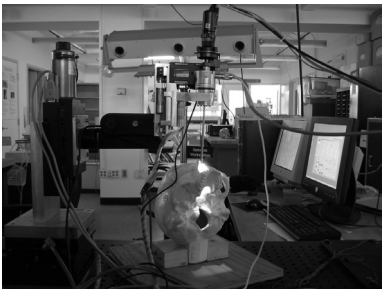

Fig. 2. Our experimental system.

Our current results are based on experiments with a phantom skull. This skull does not have any detectable texture. We added colored points on the surface that we segment in the hue space of the color representation. This way, we are able to identify and track the features in image sequences using a simple color blob tracker despite the changing lighting conditions (Fig. 3).

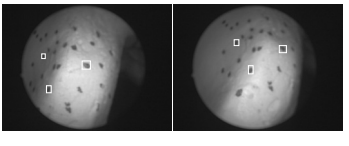

Fig. 3. Example of corresponding points on our phantom.

We obtained preliminary results with real endonasal images using our endoscope camera (Fig. 1). In real images, we compensate the brightness variations by running an edge detector on the original images and doing an SSD search on the resulting gradient images.

Localization and Mapping Step. Since the camera motion needs to be estimated in parallel to the reconstruction, the so called epipolar geometry from the motion between two camera frames needs to be recovered. An approach, commonly used in situations with at least eight point correspondences between images, is the *eight-point-algorithm*. The recovered *Essential Matrix* contains the information about the translation direction T' and rotation R between the images. The translation information can be recovered just up to a scale because of the way, how this matrix is constructed [7].

The number of corresponding (detectable) points between two camera frames varies significantly during the sinus surgery. There are situations, when less than eight points can be matched. The above approach fails in these cases, therefore, we apply here our method for camera localization and mapping requiring merely three point correspondences. We will sketch out the process below. The reader should consult [2] for details of the algorithm.

In this approach, we assume that each 3D point P_i imaged in a unifocal camera frame $p_i = (u_i v_i 1)^T$ can be represented as its direction vector $n_i = p_i/||p_i||$ and the distance to the real point D_i so that $P_i = D_i \cdot n_i$. Since, in typical applications, the scale m of the reconstruction may be unknown, the system works also with a scaled version of the distance $\lambda_i = D_i/m$. This approach calculates an estimate for the rotation $\tilde{\mathbf{R}}$ and the scaled translation $\mathbf{T}'^*$ between the points in the current frame $\{P_i\}$ and the next frame $\{P_i^*\}$ as

$$\begin{gathered}\bar{P} = \tfrac{1}{n}\sum_{i=1}^{n} P_i, \quad \bar{P}^* = \tfrac{1}{n}\sum_{i=1}^{n} P_i^*, \quad P_i' = P_i - \bar{P}, \quad P_i'^* = P_i^* - \bar{P}^*, \\ \tilde{\mathbf{M}} = \sum_{i=1}^{n} P_i'^* P_i'^{\mathbf{T}}, \quad [U\ D\ V^T] = \mathrm{svd}(\tilde{\mathbf{M}}), \\ \underline{\tilde{\mathbf{R}} = V \cdot U^T, \quad \mathbf{T}'^* = \bar{P}^* - \tilde{\mathbf{R}}^* \bar{P}.}\end{gathered} \tag{1}$$

The approach requires an initial knowledge of the values for λ_i for the first frame and it estimates a guess for translation $\mathbf{T}'^*$ and rotation $\tilde{\mathbf{R}}$. In the initial step, it assumes $\lambda_i' = \lambda_i$ and, afterwards, it iteratively converges to the true $\tilde{\mathbf{R}}, \mathbf{T}'^*$, and λ_i'. Details and simplifications of the algorithm are discussed in [1]. This algorithm requires only three corresponding points between both images to actually compute the pose difference between two camera frames $(\tilde{\mathbf{R}}, \mathbf{T}'^*)$, which makes it more suitable for the given application.

Eq. (1) updates the distance values for all tracked points P_i' for the new frame. New points can easily be added to the system using the rigid body assumption for the imaged points and solving (2)

$$\begin{pmatrix} \tilde{\mathbf{R}}\mathbf{n_x} & -\mathbf{n_x^*} \end{pmatrix} \begin{pmatrix} \lambda_x \\ \lambda_x^* \end{pmatrix} = \tilde{\mathbf{R}}\lambda_1 \mathbf{n_1} - \lambda_1^* \mathbf{n_1^*} \tag{2}$$

or in a more robust way from 3 frames to (3)

$$\begin{pmatrix} \tilde{\mathbf{R_1}}\mathbf{n_x} & -\mathbf{n_x^*} & 0 \\ \tilde{\mathbf{R_2}}\tilde{\mathbf{R_1}}\mathbf{n_x} & 0 & \mathbf{n_x^{**}} \end{pmatrix} \begin{pmatrix} \lambda_x \\ \lambda_x^* \\ \lambda_x^{**} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{R_1}}\lambda_1 \mathbf{n_1} - \lambda_1^* \mathbf{n_1^*} \\ \tilde{\mathbf{R_2}}\tilde{\mathbf{R_1}}\lambda_1 \mathbf{n_1} - \lambda_1^{**} \mathbf{n_1^{**}} \end{pmatrix}. \tag{3}$$

The pose change from image $1 \rightarrow 2$ is annotated here as $(\tilde{\mathbf{R_1}}, \mathbf{T_1})$ and the pose change between images $2 \rightarrow 3$ is annotated as $(\tilde{\mathbf{R_2}}, \mathbf{T_2})$. This equation estimates the distance λ_x to a new point P_x in the scale of an already known point P_1 from the currently tracked set of points. This way the newly added points are still measured with the same scaling factor m and the resulting 3D model has a uniform scale.

System Initialization. As mentioned above, our approach requires an initial guess about the structure of at least three landmarks. There are two possibilities for initialization of the surgical system:

- **the eight-point-algorithm** based on the estimation of the *Essential Matrix* of the system from 8 point correspondences that provides the necessary information about $(\tilde{\mathbf{R}}, \mathbf{T}'^*)$;

- **manual feature selection** in the endoscope image, where the surgeon selects three points with known correspondences to the CT-data and the system uses this information to build a map of the entire nose cavity.

The first alternative is completely unsupervised, but it requires a significant initial movement to get a well-conditioned *Essential Matrix*. The second alternative is similar to the current IGS procedure, but it is necessary just for the first frame of the sequence.

2.2 Registration of the Endoscope Data to CT Scan

Scale Recovery for 3D Reconstruction. The scaling factor m in Section 2.1 depends on the scale of the λ_i-values for the initial set of points P_i. In case of the unsupervised bootstrap using the *eight point algorithm* (Sec. 2.1) the resulting reconstruction has an arbitrary scale that depends on the scale of the translation vector $\mathbf{T}'^*$, which is usually assumed as a unit vector in this algorithm. Since the system is continuously updating the position information, it has a rough estimate about the current camera position. We use this estimate to carve out part of the CT data that fall into the expected visibility cone of the camera. This cone is slightly enlarged in all directions to compensate for the unknown camera motion.

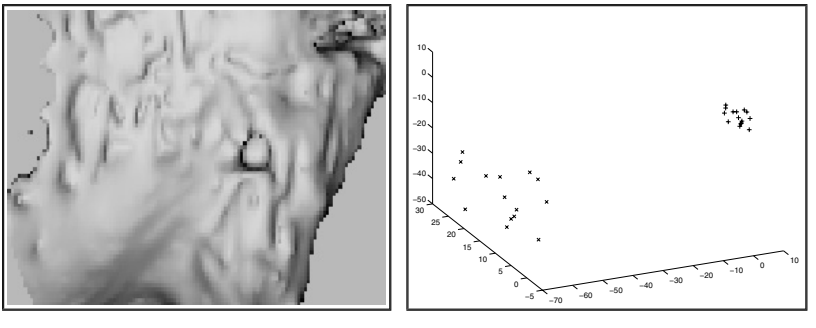

Fig. 4. Scaled reconstruction of surface points: (left) CT scan visualization of the area, (right) matched surface points with ICP {left point cloud}, scaled reconstructed points {right point cloud}.

The visible regions are usually surfaces with two dominant directions of the cavity walls with a third dimension representing the surface structure. We use this for our scale recovery by calculating the covariance matrices of the point clouds in the selected CT scan region and the current reconstruction. In both cases, the smallest eigenvalue (E_{ct}, E_{3d}) represents a metrics for the depth variations in the surface of the CT scan and in the reconstructed point cloud. The normalized eigen-vectors $\{V_{ctx}\}$ and $\{V_{3dx}\}$ and the eigenvalues allow us to calculate the scale m and the rotation $\tilde{R}_{tot}$ between the two data sets to (4). The rotation matrix $\tilde{R}_{tot}$ aligns both dominant surfaces along their normal vectors,

which are represented by the eigenvector calculated from the smallest eigenvalue (last column in each of the rotation matrices in (4)). The rotation around the normal vector cannot be restored in this way.

$$m = \frac{\sqrt{E_{ct}}}{\sqrt{E_{3d}}}, \quad V_{p\in\{CT,3D\}} = \begin{pmatrix} V_{px} \\ V_{py} \\ V_{pz} \end{pmatrix}, \quad V_{n-p} = \begin{pmatrix} 0 \\ V_{pz} \\ -V_{py} \end{pmatrix}$$
$$\tilde{R}_{ct} = \left((V_{n-ct} \times V_{ct})\, V_{n-ct}\, V_{ct} \right), \quad \tilde{R}_{3d} = \left((V_{n-3d} \times V_{3d})\, V_{n-3d}\, V_{3d} \right), \tag{4}$$
$$\tilde{R}_{tot} = \tilde{R}_{3d} \cdot \tilde{R}_{ct}^{T}$$

We apply the scaling and rotation to the zero mean point clouds that were used to calculate the covariance matrices above. This way, we obtain two point clouds with the same alignment, but the CT-scan represents a much larger area because of the expected unpredictable camera movement. Both clouds have a similar scale and alignment of the supporting plane.

We consider now both rotated clouds as sparse "images", where each "pixel" is represented by its distance to the plane calculated from the covariance matrix. We use the reconstructed 3D structure from the current view as template that is matched to the *image* constructed from the CT scan data using standard coarse-to-fine pattern matching techniques. Significant points with large deviation from the supporting plane are identified and matched in the *images* with significantly different resolutions (Fig. 5) first. This match is verified and refined based on the remaining points from the reconstruction. The physical position of the sampling points, especially in the 3D reconstruction, does not necessarily correspond to extreme values of the surface hull. We use interpolations between known values as estimates for matching.

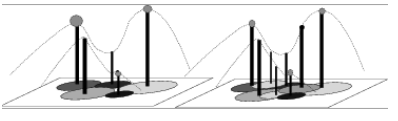

Fig. 5. Three significant points in both images are identified and matched: (left) sparse from the camera, (right) dense from CT scan.

The resulting match is used to align the two data sets. The residual error is due to imperfect sampling and the coarse structure of the point sets, especially in the case of the reconstructed data from the phantom skull. The 3D scaling step needs to be performed just in the initial step and in the cases when the system was not able to maintain the minimum number of three features and needs to re-initialize the distance measurements.

ICP. Now, we have reconstructed and localized a 3D dataset with endoscopic images, which has right scale and similar orientation and translation in the coordinate frame of the CT scan.

Rigid registration between CT images and physical data reconstructed by endoscopic images is achieved using the Iterative Closest Point (ICP) algorithm. For some applications in the endoscopic surgery, a deformable registration method can be further applied based on the results of the ICP.

We use a covariance tree data structure to search for the closest point for ICP. A covariance tree is a variant of a k-dimensional binary tree (k-D tree). The traditional k-D tree structure partitions space recursively along principal coordinate axes. In our covariance tree each sub-space is defined in the orthogonal coordinate system of the eigenvectors centered at the center of mass of the point set, and is recursively partitioned along this local coordinate frame. An important advantage of covariance trees is that the bounding boxes tend to be much tighter than those found in conventional k-D trees and tend to align with surfaces, thus producing a more efficient search [6].

3 Experimental Results

The experimental validation of our approach is carried out on the setup depicted in Fig. 2. We track the position of the endoscope with the $OptoTrack^{TM}$ system in the background to verify the motion estimation results from our system.

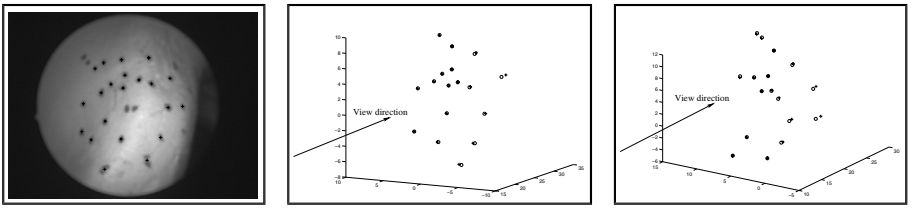

Fig. 6. 3D reconstruction results in camera coordinate frame from 2 consecutive reconstructions: (left) camera view (middle, right) reconstructed points '+', ground-truth from OptoTrack 'o'.

Fig. 6 shows two reconstruction results from a camera motion of $(4.8, 0.2, 5.2)[mm]$ with small and significant rotation between the consecutive frames. The resulting reconstruction errors had a standard deviation of (0.62, 0.3382) for each of the cases. The minimal rotational error expressed as Rodigues vector was r=(0.0017, 0.0032, 0.0004), (-0.0123, -0.0117, -0.0052). The error in the estimate of the translation vector was $\Delta T = (0.05, -0.398, 0.2172)^T, (-0.29, 0.423 - 0.4027)^T[mm]$

We tested our registration with different reconstruction results (patches) that were registered to CT skull images. Because the 3D surface data reconstructed by monocular camera may not cover the whole surface patch, we were interested in the sensitivity to drop-outs. We purposely removed parts of the data from the reconstructed patch. Our experiments with the phantom show that the ICP can accommodate noise levels in the data up to 0.6mm, combined with translational offsets of up to 10mm, and rotational offsets within 10 degrees. The vision-based reconstruction gives us errors an order of magnitude below these limits.

After ICP alignment the average distance error for the sample points is around 0.65mm. By comparison, the fiducial based registration residual error is around 0.40mm for four fiducials that are attached to the surface of the skull.

However, our method directly tells the registration error of the target region for the surgery.

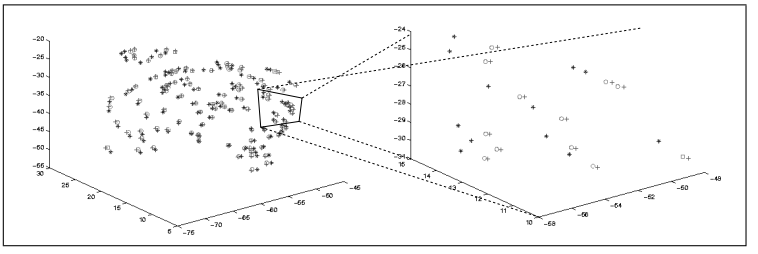

Fig. 7. The relative displacements of the sparse samples (+), their initial position recovered by VGPS(*) and their final position after alignment by ICP (o). Left is the global view of the sample data for a patch. Right is a closer look.

4 Conclusions and Future Work

The presented system performs accurate reconstruction of 3D surface points based on images from an endoscopic camera. The points are successfully aligned with CT scans of our phantom skull in the sinus area. Our major goal is to more extensively test our system in different parts of the skull and on other range images to better evaluate the performance of the system. We are currently investigating the feature type that can be used for a robust estimation and tracking of our *point features* in real endonasal images obtained in a preliminary experiment on a human subject.

Acknowledgments. Partial funding of this research was provided by the National Science Foundation under grants EEC9731748 (CISST ERC), IIS9801684, IIS0099770,and IIS0205318. This work was also partially funded by the DARPA Mars grant. The authors want to thank Dr. Masaru Ishii for his help in obtaining the preliminary data set of real endonasal images.

References

1. Darius Burschka and Gregory D. Hager. V-GPS – Image-Based Control for 3D Guidance Systems. In *Proc. of IROS*, pages 1789–1795, October 2003.
2. Darius Burschka and Gregory D. Hager. V-GPS(SLAM): – Vision-Based Inertial System for Mobile Robots. In *Proc. of ICRA*, April 2004. to appear.
3. Kennedy D.W., Bolger W.E., Zinreich S.J., and Zinreich J. *Diseases of the Sinuses: Diagnosis and Management.* 2001.
4. Olson G. and Citardi M.J. Image-guided Functional Endoscopic Sinus Surgery. *Otolaryngology-Head and Neck Surgery*, 123(3):188–194, 2000.

5. G.D. Hager and P. Belhumeur. Real-Time Tracking of Image Regions with Changes in Geometry and Illumination. *Proceedings of the IEEE Conference on Computer Vision and, Pattern Recognition*, pages 403–410, 1996.
6. Williams J.P., Taylor R.H., and Wolff L.B. Augmented k-D Techniques for Accelerated Registration and Distance Measurement of Surfaces. In *Computer Aided Surgery: Computer-Integrated Surgery of the Head and Spine*, pages 1–21, 1997.
7. E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998.