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

Darius Burschka, Ming Li, Masaru Ishii [a,b,c]
Russell H. Taylor, and Gregory D. Hager [b,a]

[a] *Computational Interaction and Robotics Laboratory, CIRL*
[b] *Computer Integrated Surgical Systems and Technology, CISST*
[c] *Department of Otolaryngology,*

The Johns Hopkins University, Baltimore, USA

Abstract

In this paper, we present a novel method for intra-operative registration directly from monocular endoscopic images. This technique has the potential to provide a more accurate surface registration at the surgical site than existing methods. It can operate autonomously from as few as two images and can be particularly useful in revision cases where surgical landmarks may be absent. A by-product of video registration is an estimate of the local surface structure of the anatomy, thus providing the opportunity to dynamically update anatomical models as the surgery progresses.

Our approach is based on a previously presented method [12] for reconstruction of a scaled 3D model of the environment from unknown camera motion. We use this scaled reconstruction as input to a PCA-based algorithm that registers the reconstructed data to the CT data and 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 step to refine the registration estimates.

The details of our approach and the experimental results with a phantom of a human skull and a head of a pig cadaver are presented in this paper.

Key words: image registration, monocular reconstruction, vision-based SLAM, Sinus surgery

* Corresponding author: Darius Burschka,
Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218, USA
Email addresses: {burschka|liming}@cs.jhu.edu, mishii3@jhmi.edu,

1 Introduction

Sinus surgery is a procedure used to remove blockages in the sinuses (the spaces filled with air in some of the bones of the skull). These blockages cause sinusitis, a condition in which the sinuses swell and become clogged, causing pain and impaired breathing. The endonasal approach for surgical treatment of sinusitis has become increasingly established during the last few decades. In such a procedure, information is provided primarily through the endoscope. The limited information from the endoscopic view requires from the surgeon a detailed knowledge of the anatomy (Fig. 1).

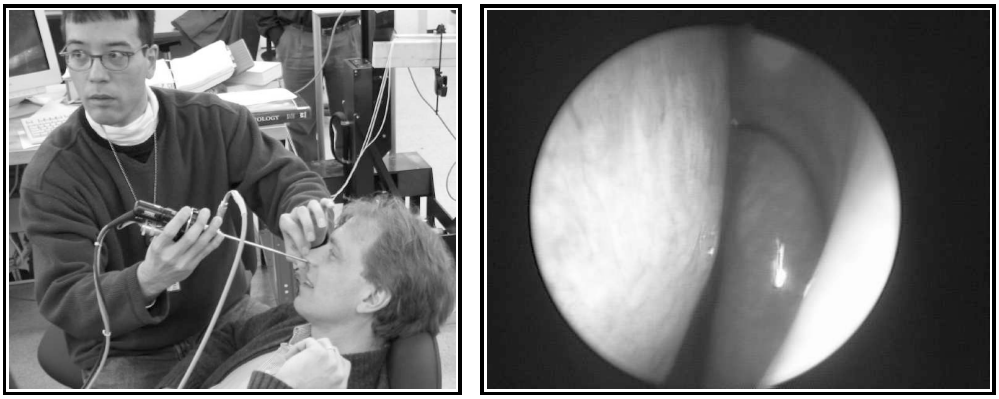

Figure 1. Endoscopic inspection of the nasal sinus cavities depicting the limited information provided to the surgeon in the current procedures.

The sinuses are physically close to the brain, the eye, and major arteries, which represent areas of concern when a fiber optic tube is inserted into the sinus region. The endonasal sinus surgery requires a high degree of precision, since minor misjudgments of anatomical relationships can lead to catastrophic consequences. These demands are particularly challenging when surgical landmarks used for navigation are distorted or obscured by extensive disease or previous surgeries. Surgical navigation systems, which allow for the real time tracking and localization of surgical instruments with respect to surrounding anatomical structures, are thus essential for safety and have been employed in endonasal approach to simplify the procedure. The evolution of this field is driven in part by technical advancements such as computer-aided surgical navigation. Figure 2 shows a typical system (here a VTI system from GE Medical Systems) for endoscopic sinus surgery.

This paper addresses two fundamental problems associated with the vision-based navigation of surgical tools in the human body: 1) the reconstruction of the 3D surface geometry within the view of the endoscopic camera and 2) the registration of the camera to a pre-operative CT scan that allows a

{rht|hager}@cs.jhu.edu (Russell H. Taylor, and Gregory D. Hager).

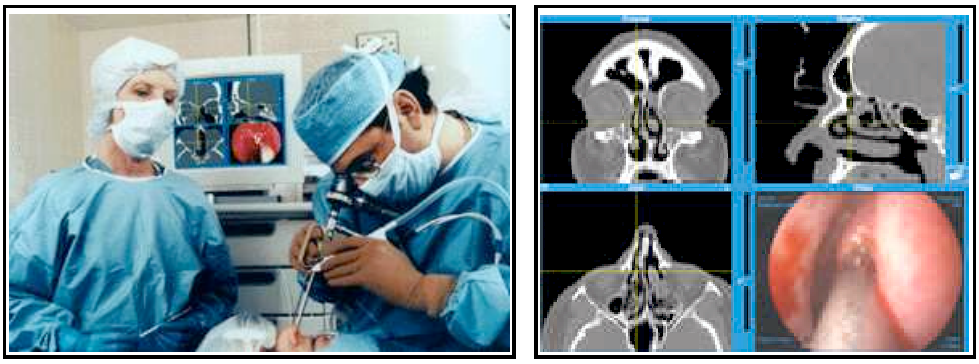

Figure 2. Typical surgical navigation system for endoscopic sinus surgery. (Left) operating room scene; (right) navigation screen shot. (Photos from VTI, Inc. — now part of GE Medical Systems)

verification of the pre-operative procedure planning. There exist a variety of surgical navigation systems that rely mostly on artificial optical and magnetic fiducials. We give a short overview over the existing approaches in the following section.

1.1 Related Work

Surgical Navigation Systems were first developed in the 1980's for neurosurgical applications (e.g., [53,31,43]). They have subsequently been applied in many surgical fields, including neurosurgery [48,26,35,23,5,56,44,46,19], craniofacial surgery [15,16,17,10,52], ENT [1,2,3,6], spine surgery [33,36,39,14], and orthopedic surgery generally [20,40,47,32,51,38]. Applications in other surgical disciplines, such as minimally-invasive hepatic surgery [28,49], are also being explored. Their primary purpose is to provide intra-operative information about the geometric relationships (e.g., between a surgical tool and anatomical structures) that cannot be readily observed using the surgeon's normal visual and haptic senses.

A typical surgical navigation system consists of a computer workstation, navigational tracking device, and associated tools with marker devices whose pose (position and orientation) is continuously measured relative to the navigational tracker (Fig. 3). Usually, one or more "reference" markers are affixed to the patient's anatomy in order to eliminate the effects of patient or tracker motion. The workstation typically imports preoperative CT, MRI, or other volumetric image data associated with the patient. After suitable calibration, typically relying on either artificial or anatomical landmarks, image processing, and registration steps are performed. This allows the system to determine the transformation between volumetric image coordinates and the patient reference coordinates. Typically, the system also computes an estimate of the accuracy of the registration, and allows for re-registration if

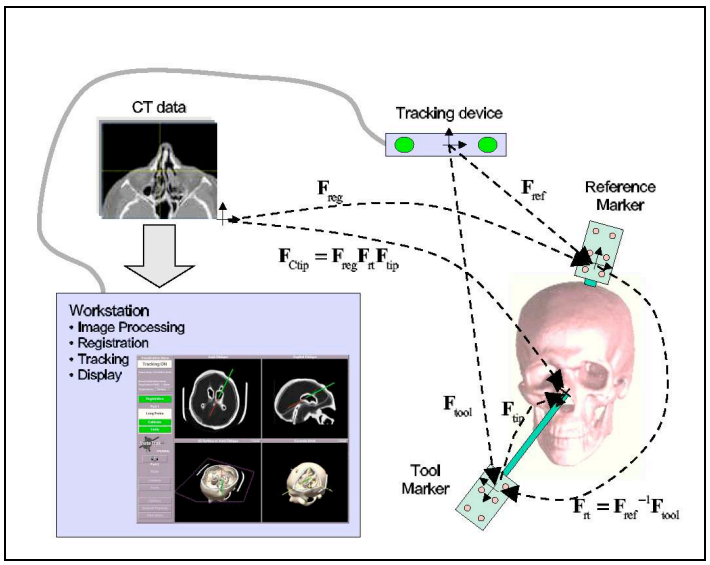

Figure 3. In the current surgical systems, the camera position is registered indirectly relative to external fiducials (markers) measured by a tracking device.

the initial solution is faulty or if re-registration is otherwise deemed necessary. Commonly used tracker technologies include specialized optical tracking systems [35,39,9,50,34,29], stereo vision with conventional cameras [27], electromagnetic sensors [54,41], acoustic sensors [44,9], and mechanical linkages [31,23,44]. Currently, systems based on specialized optical devices such as the Polaris^{TM} or OptoTrak^{TM} (Northern Digital, Waterloo, Ont.) are the most accurate, but each technology has advantages and disadvantages.

While surgical navigation has been a catalyst for advancement in this field, it suffers from a number of fundamental limitations that constrain further surgical innovation. These limitations include the inability to register the surgical site to anatomical landmarks, the inability to account for changes in anatomy brought about by surgery, and the inability to repetitively and autonomously register a patient. The first limitation is important, since it decreases the practical accuracy of surgical navigation. Current registration schemes rely on surface anatomy or fiducial points [21], away from the surgical site, to perform the registration, often leading to relatively large registration errors in the region of greatest interest. The second limitation means that, as the surgery progresses, the navigation system becomes less and less useful, as does the preoperative data. The latter leads to degradation of the surgical systems performance with time due to patient motion, shifts in the reference frame and so forth. Currently the operating surgeon must update the registration at several 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.

Most navigation systems today report position errors on the order of 2mm or

less [22,37]. However, the accuracy of registration can vary widely depending on the location of the surgical site relative to the landmarks. In particular, sites located far from a fiducial will generally have higher error than those near a fiducial. This is a consequence of the indirect nature of the tool-to-anatomy calculation.

1.2 Structure of the Paper

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: 1) 3D reconstruction from monocular camera images and 2) 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 monocular camera images is known only 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).

The architecture of our system is depicted in Figure 4. The system consists of two major components solving the problems stated above: 1) the monocular

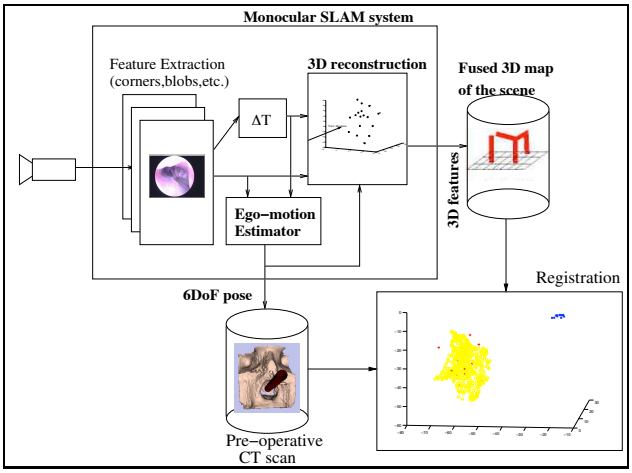

Figure 4. The architecture of our system.

VSLAM system that reconstructs the scaled version of the environment and 2) the registration module mapping the data reconstructed from the camera image to the CT scan.

In the following sections, we will describe the details of the system implementation.

2.1 Scaled 3D Reconstruction

As we mentioned in Section 2, an important component of our system is the 3D reconstruction that is depicted as the *Monocular SLAM system* in Figure 4. This module takes the monocular image data and extracts point features (e.g., corners or textured regions) in the *Feature Extraction* module and uses this pre-processed data to estimate the *Ego-Motion* of the camera relative to the environment in all 6 degrees of freedom. This motion data is then used to reconstruct 3D information of newly found features in the *3D reconstruction* module.

2.1.1 System Initialization

Our approach requires an initial guess about the 3D 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 $(\tilde{\mathbf{R}}, \mathbf{T}'^{*})$ [18]. A relation between the projections p_i, p_i^* in two camera images with known intrinsic calibration parameters can be expressed with the Essential Matrix $\tilde{\mathbf{E}}$ as $p_i^* \tilde{\mathbf{E}} p_i = 0$.

The Essential Matrix $\tilde{\mathbf{E}}$ consists of a product of two matrices

$$\tilde{\mathbf{E}} = \tilde{\mathbf{R}} \cdot \mathrm{sk}(\mathbf{T}), \text{ with } \quad \mathrm{sk}(\mathbf{T}) = \begin{pmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{pmatrix} \tag{1}$$

Note that, given a correspondence, we can form a linear constraint on $\tilde{\mathbf{E}}$. It is only unique up to scale, therefore, we need 8 matches to calculate E. This allows us to recover the rotation matrix $\tilde{\mathbf{R}}$ and the scaled version of the translation vector $\mathbf{T}'$ [18], which is sufficient for our approach as we describe later (Section 2.1.3).

- **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 bootstrap the processing.

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.

2.1.2 Feature Extraction.

The 3D reconstruction algorithm assumes that *point features* are extracted from the images. Possible features are: intersections of contours resulting from edge filters [18] or the areas themselves used for template matching in *Sum of Square Differences* (SSD) matching algorithms [25]. There are several possible matching algorithms to use in this system [8]; some using local measures [4] and some incorporating global surface constraints [45,42].

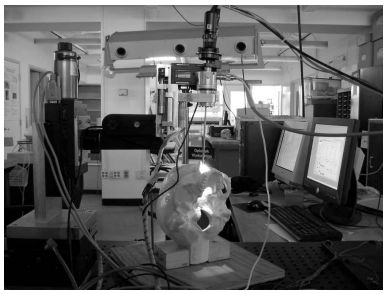

Figure 5. Our experimental system.

The problem in real endonasal images is the sparse density of points that actually can be reliably placed in correspondence to be used for a model reconstruction. Another problem is the moving light source, which is attached to the endoscope (Fig. 5). This changes the shadow cast between the images due to different illumination of small structures on the reconstructed surfaces. This forces us to switch to a brightness independent image representation.

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

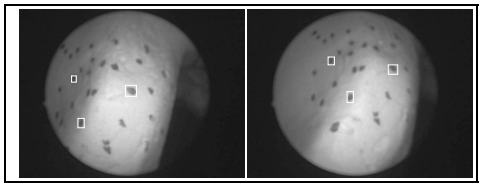

Figure 6. Example of corresponding points on our phantom.

are able to identify and track the features in image sequences using a simple color blob tracker despite the changing lighting conditions (Fig. 6).

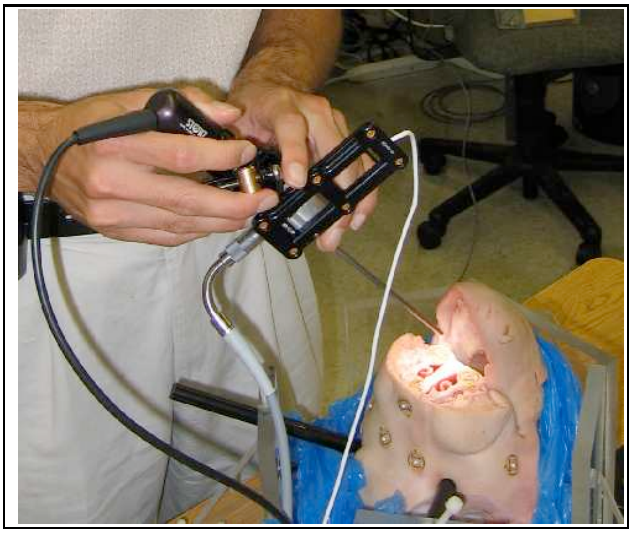

Figure 7. Validation of the approach on a head of a pig cadaver with anatomical structures similar to the human sinuses. OptoTrak data is used as a ground truth to validate the motion estimation of our algorithm.

We obtained preliminary results with real endonasal images using an endoscope camera (Fig. 7). We used an image tracker based on the minimization of the *Sum of Square Differences* (SSD) between a reference image template and its current projection in the image. This tracker follows an image template compensating for rotation, translation, illumination changes and occlusions. The details of the implementation are presented in [25]. The tracker was able to track point features in real sinuses of the head as depicted in (Fig. 8). The estimated positions of the centers of the tracked regions are used in the following step to estimate the motion of the camera and to recover the 3D information for the newly established correspondences, which allows us to add new points to the model to compensate for loss of visible points due to occlusions and coming out of the field of view.

2.1.3 Localization and Mapping Step.

Since the camera motion needs to be estimated simultaneously with the reconstruction, the so called epipolar geometry from the motion between two cam-

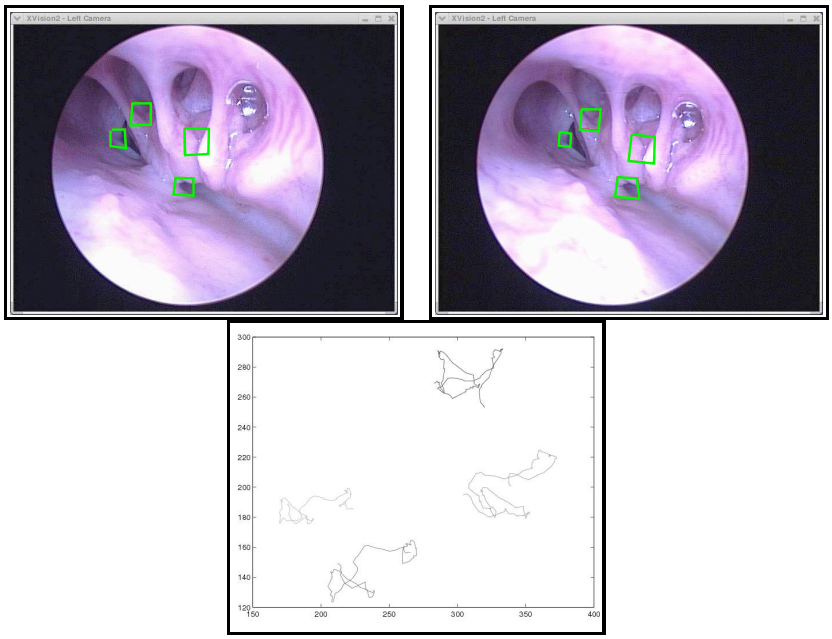

Figure 8. Tracking result in real sinus images: (top images) begin and end of a sequence with four point features being tracked; (bottom) tracking result in image coordinates.

era 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 of the structure of the epipolar constraint [18].

The number of corresponding (detectable) points between two camera frames varies significantly during the sinus surgery. There are situations, when only 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 [12] for details of the algorithm.

In this approach, we assume that each 3D point P_i imaged in a normalized 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 is supposed to work 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

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i, \quad \bar{P^*} = \frac{1}{n}\sum_{i=1}^{n} P_i^*, \quad P_i' = P_i - \bar{P}, \quad P_i'^* = P_i^* - \bar{P^*},$$

$$\tilde{\mathbf{M}} = \sum_{i=1}^{n} P_i'^* P_i'^{\mathbf{T}}, \quad [U\ D\ V^T] = \mathrm{svd}(\tilde{\mathbf{M}}), \tag{2}$$

$$\tilde{\mathbf{R}} = V \cdot U^T, \quad \mathbf{T}'^* = \bar{P^*} - \tilde{\mathbf{R}}^* \bar{P}.$$

The approach requires an initial guess for the values for λ_i for the first frame and it estimates a guess for translation $\mathbf{T}'^*$ and rotation $\tilde{\mathbf{R}}$ in an iterative process. In the initial step, it assumes the distances to the points in the new image identical to the previous one $\lambda_i' = \lambda_i$ and, afterwards, it iteratively converges to the true $\tilde{\mathbf{R}}$, $\mathbf{T}'^*$, andλ_i'. Details and simplifications of the algorithm are discussed in [11]. 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. (2) updates the distance values λ_i' for all tracked points P_i' for the new frame. New points $P_x = \lambda_x n_x$ can easily be added to the system using the rigid body assumption for the imaged points that requires the translation between the observed points to be identical for P_1 and P_x. The parameters of P_x can be calculated by solving (3).

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

Alternatively, a more robust estimation can be computed from 3 frames by solving (4).

$$\begin{pmatrix} \tilde{\mathbf{R_1}}\mathbf{n_x} & -\mathbf{n_x^*} & 0 \\ \tilde{\mathbf{R_2}}\tilde{\mathbf{R_1}}\mathbf{n_x} & 0 & \mathbf{n_x^{**}} \end{pmatrix} \begin{pmatrix} \lambda_x \\ \\ \lambda_x^* \\ \\ \lambda_x^{**} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{R_1}}\lambda_1 \mathbf{n_1} - \lambda_1^* \mathbf{n_1^*} \\ \tilde{\mathbf{R_2}}\tilde{\mathbf{R_1}}\lambda_1 \mathbf{n_1} - \lambda_1^{**} \mathbf{n_1^{**}} \end{pmatrix}. \tag{4}$$

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.

2.2 Registration of the Endoscope Data to CT Scan

The other important module in our system depicted in (Fig. 4) is the registration of the scaled data extracted in the previous section. The extracted patch needs to be localized correctly within the CT scan and registered to it. This is a two step process in which we recover back the scale from a PCA based method and later use ICP (Iterative Closest Point) method to align it correctly.

2.2.1 Scale Recovery for 3D Reconstruction

The scaling factor m in Section 2.1.3 depends on the scale of the λ_i-values for the initial set of points. In case of an unsupervised bootstrap using the *eight point algorithm* (Sec. 2.1.1) the resulting reconstruction has an arbitrary scale that depends on the scale of the translation vector $\mathbf{T}'$, which is usually assumed to be a unit vector.

The system has usually a rough estimate of the current camera position. We use this estimate to carve out part of the CT surface data that falls into the expected visibility cone of the camera. This cone is slightly enlarged in all directions to compensate for the unknown camera motion. The size of the extracted patch depends on the uncertainty about the camera position.

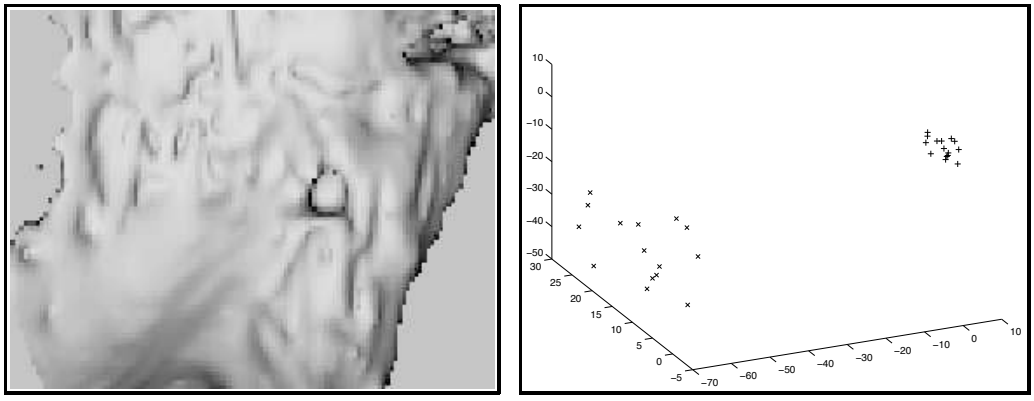

Figure 9. Scaled reconstruction of surface points: (left) CT scan visualization of the area, (right) matched surface points with ICP {left point cloud}, scaled reconstructed points {right point cloud}.

The visible regions are usually surfaces with two dominant directions of the cavity walls with a third dimension representing the surface structure or combinations of such walls. We assume for now that the CT data patch consists of a single surface with some surface structure on it. We will discuss later, how to split more complex structures into simpler planar patches.

We use the property of two dominant surface directions for our scale recov-

ery by calculating the covariance matrix $\widetilde{C}_k$ of the point cloud in the selected CT scan region and in the current camera reconstruction. The eigenvalues and eigenvectors of $\widetilde{C}_k$ define a new coordinate system with the two eigenvectors calculated from the larger eigenvalues defining the supporting plane in the cloud and the third eigenvector describing the depth variation in the measurement.

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 (5). 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 (5)). The rotation around the normal vector cannot be restored in this way.

$$
\begin{aligned}
m = \frac{\sqrt{E_{ct}}}{\sqrt{E_{3d}}}, \quad V_{p\in\{CT,3D\}} = (V_{px}\ V_{py}\ V_{pz})^T, \quad V_{n-p} = \frac{(0\ V_{pz}\ -V_{py})^T}{||(0\ V_{pz}\ -V_{py})^T||} \\
\tilde{R}_{ct} = \Big((V_{n-ct} \times V_{ct})\ V_{n-ct}\ V_{ct} \Big), \\
\tilde{R}_{3d} = \Big((V_{n-3d} \times V_{3d})\ V_{n-3d}\ V_{3d} \Big), \\
\tilde{R}_{tot} = \tilde{R}_{3d} \cdot \tilde{R}_{ct}^T
\end{aligned}
\tag{5}
$$

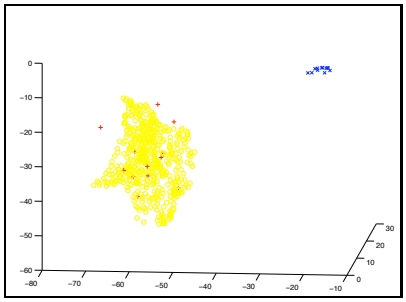

Figure 10. After the alignment along the normal vector to the supporting plane the scale is roughly recovered, but rotation around the normal vector is possible.

We apply the scaling m and rotation $\tilde{R}_{tot}$ 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. Fig. 10 depicts an alignment result of the scaled point cloud in the top right corner to the 3D surface from the CT scan. The transformed points may have a significant rotational error around the normal vector visible in the above figure as crosses ('+'). We correct this error in the next step.

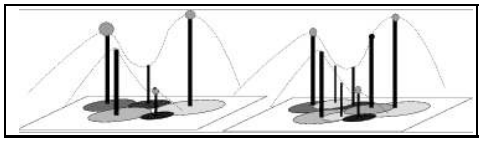

Figure 11. Distance to the *supporting plane* calculated in (5) is used as a pseudo-image representation to match the sparse reconstruction (left) to the dense point cloud (right).

We consider now both rotated point clouds as sparse "images", where each "pixel" is represented by its distance to the supporting plane calculated from the covariance matrix. We use the reconstructed 3D structure from the current view as a template that is matched to the "image" constructed from the CT scan data using the SSD matching approach described in [25]. In this case we can allow only an optimization in in-the-image rotations to find the missing parameter for the alignment and we treat the missing information in the sparse image of the reconstruction as "occlusions." This allows us to find the rotation between the two point sets and to localize the current view within the larger region of the CT scan to model the uncertainty in the registration of the camera relative to the CT scan. The "search window" can be adjusted dependent on the current knowledge about the relative position of the two coordinate frames to each other.

The physical positions of the sampling points, especially in the 3D reconstruction, do not necessarily correspond to the extreme values of the surface hull. Therefore, the above match trial can fail for a specific set of three points. The 3D reconstruction may not have reconstructed the peak value but some random value along the slope instead.

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 bootstrap phase and in cases when the system was not able to maintain the minimum number of three features and needs to re-initialize the distance measurements.

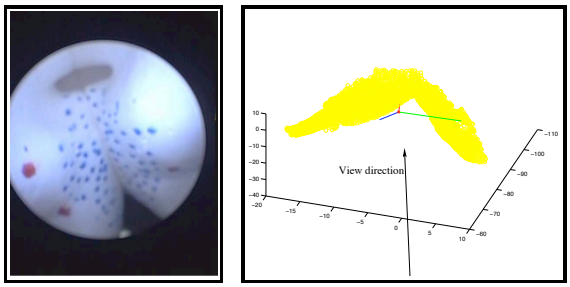

Figure 12. The two surfaces need to be separated first, before our PCA-based scale recovery can be applied.

The above scaling step relies on the fact that the data contains only one supporting plane with depth variations that are used to calculate the scaling factor m (5). In case, when the CT or the reconstructed data set contains two or more surfaces in a corner arrangement (Fig. 12), the structure needs to be split into single surfaces before we apply the above scaling. We use a modified version of a split-and-merge algorithm here. We establish for the entire data-set the plane equation of the supporting plane A based on $\tilde{C}_k$ and (5). For each point P_i of the data-set, we calculate the distance d_i from the plane

$$d_i = (P_i - \bar{P}) \cdot V_n \tag{6}$$

with V_n being the normal vector of the estimated plane A.

Point with the largest deviation from the original plane A and border points on both sides of the data set are used to split the original surface consecutively into sub-surfaces defined by these points. Each three points define one sub-surface. We split the points of the original surface into sub-surfaces depending on the distances to the resulting sub-planes. The above evaluation is repeated on the sub-surfaces until the maximum deviation $\max\{\mathrm{d_i}\}$ is smaller than the expected depth structure in the surfaces.

2.2.2 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.

We have chosen to perform the rigid registration between CT images and physical data reconstructed from endoscopic images 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. The anatomy of the nasal and sinus cavity is mostly bony tissue. To this end, ICP is considered the right or suitable registration solution for this application. An alternative approach for the alignment, that performs the point alignment and the registration with reduced sensitivity to deformation and outliers is described in [13]. This approach uses a thin-plate spline assumption to iteratively match sets of points and recover the alignment scale. The advantage in our case is that for the initial alignment and scale recovery only the region properties of the CT scan are considered. We have a strong mismatch in the number of points reconstructed by the camera and recovered from the CT scan that makes the application of the algorithm in [13] difficult.

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 [30].

3 Results

The system is implemented in C++ using the image processing library XVision2 [24] and the Vision-Based SLAM (*Simultaneous Localization and Mapping*) library VGPS [12]. It was tested on a laptop computer with a Pentium-M processor @1.2GHz running Linux OS.

3.1 Geometric Investigations

In this section, we first review the essential geometric concepts necessary for our algorithms, and calculate the theoretical accuracy with which geometry can be computed in several representative cases. We then describe initial registration results we have achieved using prototype algorithms on phantom data. This data validates the concept that 3D-3D surface registration from video data is possible given good image information. Finally, we describe preliminary results on video tracking and reconstruction from porcine cadaver data to demonstrate that the image-level matching problem can be solved on realistic data.

3.1.1 Point Reconstruction Accuracy

We performed a standard calibration procedure [55,7] on a Storz Telecam 20212113U NTSC equipped with a zero degree lens. Using the resulting parameters, we then calculated the expected accuracy of point location reconstruction assuming image point localization to 1 pixel. This is generally considered conservative — many systems report point localization accuracy to 1/4 pixel or better [8].

(Fig. 13) shows the expected reconstruction accuracy for several cases. The x axis is the distance to the target point, and the y axis is the reconstruction error due to a 1 pixel mislocalization. Three cases are considered: 1) a point directly in front of the endoscope and a motion parallel to the image plane of

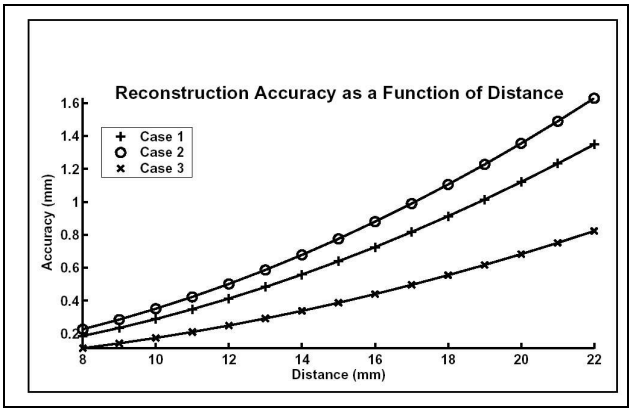

Figure 13. Expected reconstruction accuracy for several cases. Case 1: a point located in different distances along the optical axis estimated using 1mm lateral motion. Case 2: A point located along a 15 degrees off-axis estimated using a 3mm forward motion. Case 3: a point located 15 degrees off-axis estimated using combined 1 mm lateral and 3 mm forward motion.

1mm, 2) a point 15 degrees to the side and motion of 3mm in the direction of the optical axis, and 3) a point 15 degrees to the side and a motion consisting of 1 mm side to side motion with 3 mm forward motion. As can be seen from the figure, for targets within 2 cm of the endoscope, reconstruction accuracies on the order of 1.2mm or less can be expected. Indeed, for targets 1 cm from the endoscope, accuracies are 0.3mm or better.

3.1.2 Viewpoint Registration Accuracy

Consider a set of points $p_i = (x_i, y_i, z_i)^t$ lying on a surface, let c_i denote the optical center of the camera, and let $q_i = (u_i, v_i)$ denote the image observations of the points. Our objective is to determine the accuracy with which the observer position c_i can be calculated from noisy observations. We note that we can consider the case when $c_i = 0$, as any other problem can be transformed to this case by a rigid body motion. Also, note that, in general, we need at least 3 points in general position to compute the full rigid body pose of the camera.

Under these assumptions, the first order term in a Taylor series expansion of the camera projection equations yields

$$\triangle q_i = \frac{1}{z_i} \begin{bmatrix} f & 0 & u_i \\ 0 & f & v_i \end{bmatrix} \triangle c = J_i \triangle c \tag{7}$$

If we now assume that we have a covariance matrix $\Lambda_q = \sigma^2 I_2$ modeling the accuracy of each observation q_i, we can compute the expected covariance on

c_i as:

$$\lambda_c = \sigma^2(\sum_i J_i^t J_i)^{-1} \tag{8}$$

Table 1 shows the expected accuracy of registration as a function of distance and point spread for 4 points arrayed symmetrically about the optical center. This data is again for the previously described endoscope. It can be seen that the expected registration accuracy is more than an order of magnitude better than that stated for commonly used navigation systems [22,37].

3.2 Tracking

To verify the feasibility of video feature tracking on sinus tissue, we acquired data ex-vivo from a porcine cadaver. The model was acquired within 24 hours of slaughter. The nose was shortened to make the anatomy more consistent with the human sinuses. Video and tracking data was acquired in a manner identical to the skull phantom with the exception that the endoscope was manually (rather than robotically) manipulated.

In this case, tracking was performed using gradient optimization of a local correlation measure with respect to the four parameters: 2 locations, scale and orientation. Figure 14 shows tracking results in two of five test cases where different types of features were tracked. The tracking uses a variety of anatomical structures in the images varying from significant vessel structures visible in the images to small structures on the actual surfaces that allow a robust tracking. A validation method described in [12] allows a robust filtering of reliable features in the set of tracked landmarks. We need to ensure that the feature position is stable to ensure an accurate calculation of the camera ego-motion from the collected image data.

In this case, the features themselves were chosen by hand rather than automatically as it was the case for the phantom skull. Manual inspection of the

angle	5 degrees		10 degrees		15 degrees	
distance	10mm	20mm	10mm	20mm	10mm	20mm
accuracy parallel to image plane	0.0148	0.0296	0.0146	0.0292	0.0143	0.0287
accuracy along optical axis	0.1195	0.2389	0.0586	0.1172	0.0378	0.0756

Table 1
Expected registration accuracy for several representative cases. It is assumed that 4 points are observed. The points are located at the given distance from the image plane and their rays of projection form the stated angle with the optical axis.

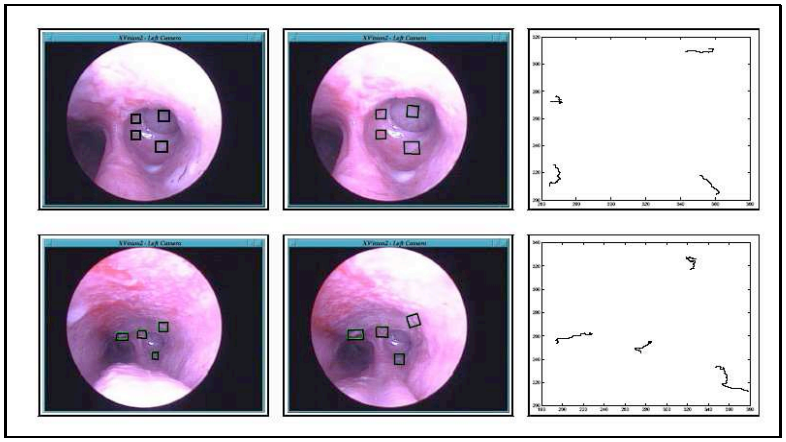

Figure 14. The first column shows the originally chosen regions, the second the locations at a later point in the sequence and the final column shows the change in location of the features through the sequence.

tracking data shows it to be quite accurate, however no external reference validating these tracking results is yet available.

The experiments on the porcine head reveal that the type of useful features for tracking depends on the region where the navigation is performed. Along the nasal area, useful features seem to be vascular structures that provide the necessary gradient information to uniquely define the tracking region over a sequence of images. In deeper regions of the skull, small surface structures visible in Fig. 14 provide a good gradient information for tracking. Finally, the Sinus structures themselves define robust structures to be tracked in the actual Sinus area.

3.3 Localization and Mapping Subsystem

The system is capable of tracking and reconstruction of 4-6 point features with 3D structure recovery in frame-rate on a Pentium-M processor @1.2GHz running Linux OS. The actual iterative structure calculation with the reported accuracy of 0.5mm takes approx. 10ms on the above CPU. This is part of the Ego-Motion Estimator depicted in (Fig. 4).

3.4 Registration to the CT Scan

The experimental validation of our approach is carried out on the setup depicted in Fig. 5. We track the position of the endoscope with the $OptoTrak^{TM}$ system in the background to verify the motion estimation results from our system.

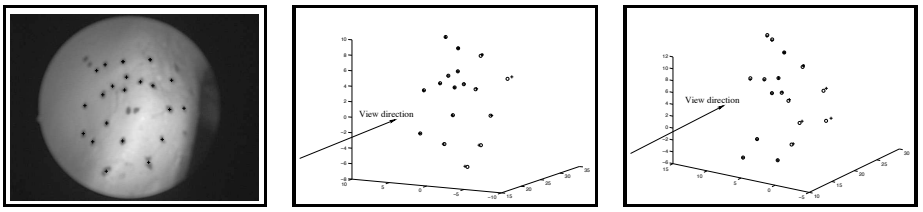

Figure 15. 3D reconstruction results in camera coordinate frame from 2 consecutive reconstructions: (left) camera view (middle, right) reconstructed points '+', ground-truth from OptoTrak 'o'.

Fig. 15 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 Rodrigues 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.

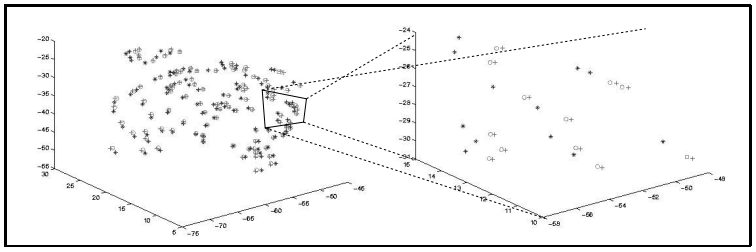

Figure 16. The relative displacements of the sparse samples (+), their initial position recovered by VGPS(*) and their final position after alignment by ICP (o). Left is the global view of the sample data for a patch. Right is a closer look.

	Translation offset range	Rotation offset range	noise level	Average error
Patch1	$\{\Delta X, \Delta Y, \Delta Z\} = \pm 10mm$	$\{\Delta\alpha_{k\in\{X,Y,Z\}} = \pm 10^\circ$	0.5mm	0.65mm
Patch2	$\{\Delta X, \Delta Y, \Delta Z\} = \pm 10mm$	$\{\Delta\alpha_{k\in\{X,Y,Z\}} = \pm 10^\circ$	0.5mm	0.48mm

Table 2
The results of ICP on phantom. Patch 1 contains around 2500 triangles, and 100 sample points are used for test. Patch 2 contains around 900 triangles, and 50 sample points are used for test.

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. We plan to enhance the quality of the final registration using ICP by reconstructing dense surface models around the points reconstructed from the vision-based SLAM system. The reconstructed points will be used as seeds for a dense disparity reconstruction under a smoothness assumption in the local area around the point. This increased description of the surface properties around the reconstructed point promises additional matching criterions that can be used in the alignment process.

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 an animal cadaver.

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.

References

[1] L. Adams, J. M. Gilsbach, W. Krybus, D. Meyer-Ebrecht, R. Mosges, and G. Schlondorff. Cas - a navigation support for surgery. In *3d Imaging in Medicine*, pages 411–423. Springer-Verlag, 1990.

[2] L. Adams, A. Knepper, W. Krybus, D. Meyer-Ebrecht, G. Pfeiffer, R. Rueger, , and M. Witte. Navigation support for surgery by means of optical position detection and real-time 3d display. In *Proceedings Computer Aided Radiology*. Springer Verlag, 1991.

[3] L. Adams, A. Knepper, W. Krybus, D. Meyer-Ebrecht, G. Pfeiffer, R. Ruger, , and M. Witte. Orientation aid for head and neck surgeons. *Innovation et Technologie en Biologie et Medicine*, 14(4):409–424, 1992.

[4] J. Banks and P. Corke. Quantitative evaluation of matching methods and validity measures for stereo vision. *International Journal of Robotics Research*, 20(7):512–532, 2001.

[5] G. H. Barnett, D. W. Kormos, D. Piraino C. P. Steiner, J. Weisenberger, F. Hajjar, C. Wood, , and J. McNally. Frameless stereotaxy using a sonic digitizing wand: Development and adaptation to the picker vistar midical imaging system. In Robert J. Maciunas, editor, *Interactive Image-Guided Neurosurgery*, chapter 10. American Association of Neurological Surgeons, 1993.

[6] D. Bartz, O. Gurvit, D. Freudenstein, H. Schiffbauer, and J. Hoffman. Integration of navigation, optical and virtual endoscopy in neurosurgery and oral and maxillofacial surgery. In *3rd Caesarium – Computer Aided Medicine*, 2001.

[7] Jean-Yves Bouget. The matlab camera calibration toolkit. `http://www.vision.caltech.edu/bouguetj/calib_doc/`.

[8] Myron Z. Brown, Darius Burschka, and Gregory D. Hager. Advances in Computational Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003.

[9] R. Bucholz and K. Smith. A comparison of sonic digitizers versus light emitting diode-based localization. *Interactive Image-Guided Neurosurgery*, pages 179–200, 1993.

[10] C. Burghart, R. Krempien, T. Redlich, A. Pernozzoli, H. Grabowsky, J. Muncherberg, S. Hassfeld J. Albers, C. Vahl, U. Rembold, and H. Worn. Robot assisted craniofacial surgery: first clinical evaluation. *Computer Assisted Radiology and Surgery*, pages 828–833, 1999.

[11] Darius Burschka and Gregory D. Hager. V-GPS – Image-Based Control for 3D Guidance Systems. In *Proc. of IROS*, pages 1789–1795, October 2003.

[12] Darius Burschka and Gregory D. Hager. V-GPS(SLAM): – Vision-Based Inertial System for Mobile Robots. In *Proc. of ICRA*, pages 409–415, April 2004.

[13] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *CVPR*, pages 44–51, 2000.

[14] K. Cleary. Workshop report: Technical requirements for image-guided spine procedures, 1999.

[15] C. Cutting, R. Taylor, R. Bookstein, D. Khorramabadi, B. Haddad, A. Kalvin, H. Kim, and M. Nox. Computer aided planning and execution of cranofacial surgical procedures. In *Proc. IEEE Engineering in Medicine and Biology Conference*, 1992.

[16] C. B. Cutting, F. L. Bookstein, , and R. H. Taylor. Applications of simulation and morphometrics, robotics in craniofacial surgery. In R. H. Taylor, S. Lavallee, G. Burdea, , and R. Mosges, editors, *Computer-Integrated Surgery*, pages 541–544. MIT Press, 1997.

[17] C. B. Cutting, B. Grayson, and H. C. Kim. Precision multi-segment bone positioning using computer aided methods in craniofacial surgery applicationa. In *12'th IEEE Engineering in Medicine and Biology Conference*. IEEE, 1990.

[18] Christopher M. Brown Dana H. Ballard. *Computer Vision*. Department of Computer Science University of Rochester, Prentice-Hall, Inc., 1982.

[19] D. Dey, D. Gobbi, P. Slomka, K. Surry, and T. Peters. Mixed reality merging of endoscopic images and 3d surfaces. *IEEE Transactions on Medical Imaging*, 21(1):23–30, 2002.

[20] A. M. DiGioia, D. A. Simon, B. Jaramaz, F. Morgan M. Blackwell, R. V. O'Toole, B. Colgan, , and E. Kischell. Hipnav: Pre-operative planning and intra-operative navigational guidance for acetabular implant placement in total hip replacement surgery. *Computer Assisted Orthopedic Surgery*, 1996.

[21] Kennedy D.W., Bolger W.E., Zinreich S.J., and Zinreich J. *Diseases of the Sinuses: Diagnosis and Management*. 2001.

[22] M. Fried, J. Kleefield, H. Gopal, E. Reardon, B. Ho, and F. Kuhn. Image-guided endoscopic surgery: Results of accuracy and performance in a multicenter clinical study using an electromagnetic tracking system. *Laryngoscope*, 107(5):594–601, 1997.

[23] R. L. Galloway, C. A. Edwards, S. Schreiner J. G. Thomas, , and R. J. Maciunas. A new device for interactive, image guided surgery. In *Proc. SPIE Medical Imaging V*, 1991.

[24] G. Hager and K. Toyama. The XVision System: A General-Purpose Substrate for Portable Real-Time Vision Applications. *Computer Vision and Image Understanding*, 69(1):23–37, 1995.

[25] 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.

[26] M. P. Heilbrun, S. Koehler, P. McDonald, V. Sieminov W. Peters, , and C. Wiker. Implementation of a machine vision method for stereotactic localization and guidance. In R. Maciunas, editor, *Interactive Image-Guided Neurosurgery*, pages 169–177. AANS, 1993.

[27] M. P. Heilbrun, P. McDonald, C. Wiker, S. Koehler, and W. Peters. Stereotactic localization and guidance using a machine vision technique. *Stereotact Funct Neurosurg*, 58:94–98, 1991.

[28] A. J. Herline, J. D. Stefansic, S. L. Hartmann J. P. Debelak, C. W. Pinson, R. L. Galloway, and W. C. Chapman. Image-guided surgery: Preliminary feasibility studies of frameless stereotactic liver surgery. *Arch. Surg.*, 134:644–650, 1999.

[29] R. Hofstetter, M. Slomczycowski, M. Sati, , and L. P. Nolte. Principles of precise fluoroscopy based surgical navigation. *in 4th International Symposium on CAOS*, page 28, 1999.

[30] 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.

[31] Y. Kosugi, E. Watanabe, J. Goto, T. Watanabe, S. Yoshimoto, K. Takakura, , and J. Ikebe. An articulated neurosurgical navigation system using mri and ct images. *IEEE Transactions on Biomedical Engineering*, pages 147–152, February 1988.

[32] M. Kunz, F. Langlotz, J. Strauss, W. Ruther, , and L.-P. Nolte. Development and verification of an non-ct based total knee arthroplasty system for the lcs prosthesis. *in First Annual Meeting of CAOS International. Davos*, page 131, 2001.

[33] S. Lavallee, P. Sautot, J. Troccaz, P. Cinquin, , and P. Merloz. Computer assisted spine surgery: a technique for accurate transpedicular screw fixation using ct data and a 3-d optical localizer. *Medical Robotics and Computer-Assisted Surgery*, 2:315–32, 1994.

[34] S. Lavallee, J. Troccaz, P. Sautot, B. Mazier, P. Cinquin, P. Merloz, and J.-P. Chirossel. Computer-assisted spinal surgery using anatomy-based registration. In *Computer-Integrated Surgery*, pages 425–449. MIT Press, 1996.

[35] R. J. Maciunas. *Interactive Image-Guided Neurosurgery.* American Association of Neurological Surgeons, 1993.

[36] P. Merloz, J. Tonetti, A. Eid, C. Faure, L. Pittet, M. Coulomb, P. Sautot, , and O. Raoult. Computer-assisted versus manual spine surgery: clinical report. In E. Grimson and R. Mosges, editors, *Proc. First Joint Conference of CVRMed and MRCAS*, volume 1205, pages 541–544. Springer, 1997.

[37] R. Metson, R. Gliklich, and M. Cosenza. A comparison of image guidance systems for sinus surgery. *Laryngoscope*, 108(8):1164–1170, 1998.

[38] L. P. Nolte, H. Visarius, and et al. *Computer Assisted Orthopaedic Surgery.* Hofgrefe & Huber, 1996.

[39] L. P. Nolte, J. Zamorano, Jiang, F. Langlotz G. Want, E. Arm, , and H. Visurius. A novel approach to computer assisted spine surgery. *in First Int. Symp. on Medical Robotics and Computer Assisted Surgery (MRCAS 94). Pittsburgh: Shadyside Hospital*, pages 323–328, 1994.

[40] F. Picard, A. Digioia, D. Sell, B. Jaramaz J. Moody, C. Nikou, R. LaBarca, , and T. Levison. Computer-assosted navigation for knee arthroplasty: intra-operative measurements of alignment and soft tissue balancing. *in First Annual Meeting of CAOS International. Davos*, page 114., 2001.

[41] F. Poulin and L. Amiot. Electromagnetic tracking in the or: Accuracy and sources of intervention. In *Proc. CAOS USA*, pages 233–235, 2001.

[42] Nicholas A. Ramey, Jason J. Corso, William W. Lau, Darius Burschka, and Gregory D. Hager. Real Time 3D Surface Tracking and Its Applications. In *Proceedings of Workshop on Real-time 3D Sensors and Their Use (at CVPR 2004)*, 2004.

[43] H. Reinhardt, H. Meyer, and A. Amrein. A computer-assisted device for intraoperative ct-correlated localization of brain tumors. *Eur. Surg. Res*, 20:51–58, 1988.

[44] H. F. Reinhardt. Neuronagivation: A ten years review. In R. Taylor, S. LAvallee, G. Burdea, and R. Moegses, editors, *Computer-Integrated Surgery*, pages 329–342. MIT Press, 1996.

[45] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, April 2002.

[46] M. Scholtz, W. Konen, S. Tombrock, L. Adams B. Fricke, M. v. During, A. Hentsch, L. Heuser, , and A. Harders. Development of an endoscopic navigating system based on digital image processing. *Journal of Computer Aided Surgery*, 3(3):134–143, 1998.

[47] D. A. Simon, B. Jaramaz, M. Blackwell, A. M. Digioia F. Morgan, M. D., E. Kischell, B. Colgan, , and T. Kanade. Development and validation of a navigational guidance system for acetabular implant placement. In J. Troccaz, E. Grimson, , and R. Mosges, editors, *Proc. First Joint Conference of CVRMed and MRCAS*, volume 1205, pages 583–592. Springer, 1997.

[48] K. R. Smith, K. J. Frank, and R. D. Bucholz. The neurostation - a highly accurate minimally invasive solution to frameless stereotactic neurosurgery. *Comput. Med. Imaging Graph.*, 18:247–256, 1994.

[49] J. Stefansic, A. Herline, Y. Shyr, W. Chapman, J. Fitzpatrick, B. Dawant, and R. J. Galloway. Registration of physical space to laparoscopic image space for use in minimally invasive hepatic surgery. *IEEE Trans Med Imaging*, 19(10):1012–1023, 2000.

[50] R. H. Taylor, H. A. Paul, B. Mittelstadt C. B. Cutting, W. Hanson, P. Kazanzides, B. Musits, A. Kalvin Y.-Y. Kim, B. Haddad, D. Khoramabadi, , and D. Larose. Augmentation of human precision in computer-integrated surgery. *Innovation et Technologie en Biologie et Medicine*, 13(4):450–459, 1992.

[51] G. van HellenMondt, M. deKleuver, and P. Pavlov. Computer assisted pelvic osteotomies; clinical experience in 25 cases. In *First Annual Meeting of CAOS International*, page 123, 2001.

[52] C. VanderKolk, S. Zinreich, B. Carson, N. Bryan, and P. Manson. An interactive 3d-ct surgical localizer for craniofacial surgery. In A. Montoya, editor, *Craniofacial Surgery*. 1992.

[53] E. Watanabe, T. Watanabe, S. Manka, and et. al. Three-dimensional digitizer (neuronavigator): new equipment for computed tomography-guided stereotaxic surgery. *Surg Neurol*, 27:543–547, 1987.

[54] X. Wu and R. Taylor. A direction space interpolation technique for calibration of electromagnetic surgical navigation systems. In *Proceedings of the Sixth International Conference on Medical Image Computing and Computer Assisted Intervention*, volume II, pages 215–22. Springer Verlag, 2003.

[55] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[56] S. J. Zinreich, S. A. Tebo, D. M. Long, D. E. Mattox H. Brem, M. E. Loury, C. A. Vander Kolk, D. W. Kennedy W. M. Koch, , and R. N. Bryan. Frameless stereotaxic intergration of ct imaging data: Accuracy and initial applications. *Radiology*, pages 735–742, 1993.