Int J CARS (2013) 8:407–417
DOI 10.1007/s11548-012-0795-1

ORIGINAL ARTICLE

3D position estimation of flexible instruments: marker-less and marker-based methods

Rob Reilink · Stefano Stramigioli · Sarthak Misra

Received: 3 July 2012 / Accepted: 8 October 2012 / Published online: 31 October 2012

Abstract

Purpose Endoscopic images can be used to allow accurate flexible endoscopic instrument control. This can be implemented using a pose estimation algorithm, which estimates the actual instrument pose from the endoscopic images.

Methods In this paper, two pose estimation algorithms are compared: a marker-less and a marker-based method. The marker-based method uses the positions of three markers in the endoscopic image to update the state of a kinematic model of the endoscopic instrument. The marker-less method works similarly, but uses the positions of three feature points instead of the positions of markers. The algorithms are evaluated inside a colon model. The endoscopic instrument is manually operated, while an X-ray imager is used to obtain a ground-truth reference position.

Results The marker-less method achieves an RMS error of 1.5, 1.6, and 1.8 mm in the horizontal, vertical, and away-from-camera directions, respectively. The marker-based method achieves an RMS error of 1.1, 1.7, and 1.5 mm in the horizontal, vertical, and away-from-camera directions, respectively. The differences between the two methods are not found to be statistically significant.

Conclusions The proposed algorithms are suitable to realize accurate robotic control of flexible endoscopic instruments, enabling the physician to perform advanced procedures in an intuitive way.

R. Reilink (✉) · S. Stramigioli · S. Misra
MIRA—Institute for Biomedical Technology and Technical Medicine,
University of Twente, P.O. Box 217, 7500 AE, Enschede,
The Netherlands
e-mail: r.reilink@utwente.nl

S. Stramigioli
e-mail: s.stramigioli@ieee.org

S. Misra
e-mail: s.misra@utwente.nl

Keywords Computer-assisted surgery · Flexible endoscopy · Medical robotics · Image-guided surgery · Minimally invasive surgery

Introduction

Flexible endoscopy is a minimally invasive procedure that allows examination of the internal body cavities of the patient. The physician uses a flexible endoscope to perform this procedure. This endoscope consists of a flexible tube with a camera at the distal end. The camera can be moved in two degrees of freedom (DOFs) by turning two concentric wheels on the control handle. In addition to performing examinations, conventional endoscopes can also be used to perform small interventions, such as performing a biopsy, removing small sections of malignant mucosal tissue, or removing polyps. This is done using a long, flexible instrument that is inserted through the working channel of the endoscope. This instrument can be operated in two additional DOFs: insertion/retraction and rotation around the axis of the instrument.

Because of the limitations in the available motions, only simple interventions can be performed when using conventional endoscopes and their instruments. In order to broaden the range of possible interventions, advanced flexible endoscopes are currently being developed, such as the EndoSAMURAI (Olympus Corp., Tokyo, Japan) and the ANUBIS (Karl Storz GmbH & Co. KG, Tuttlingen, Germany). These endoscopes both allow multiple instruments to be used simultaneously, and their instruments can be operated in more DOFs. This gives the physician the dexterity that is required to perform more advanced interventions, such as the removal of larger sections of mucosal tissue, and Natural Orifice Transluminal Endoscopic Surgery (NOTES, [11]).

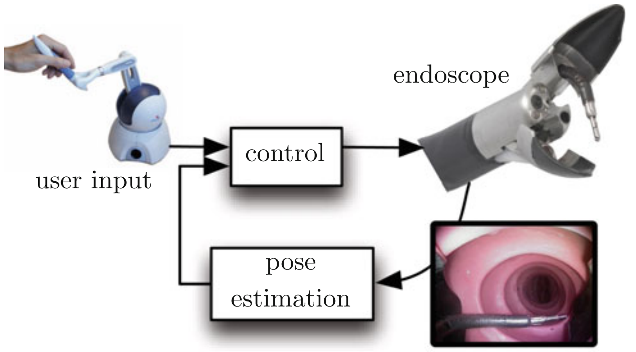


Fig. 1 Instrument control using visual feedback: The images that are captured by the endoscope are used by the pose estimation algorithm to find the actual instrument pose. The control actuates the endoscopic instrument such that it moves to the pose that is commanded by the user

However, the aforementioned flexible endoscopes are difficult to operate. Multiple physicians are required to operate all DOFs [14]. Since optimal coordination between the physicians is difficult, and because of the increased costs, this is undesirable. In addition, the control of the endoscope and the instruments is not intuitive, since there is no one-to-one mapping between the movement of the controls and the movement of the instrument. Intuitive control is also hindered by the presence of hysteresis due to friction and compliance in the mechanical control system of the instrument.

In order to overcome the aforementioned problems associated with current advanced flexible endoscopes, a robotic actuation system could be employed. If all DOFs of the endoscope and the instruments can be actuated, a telemanipulation setup (Fig. 1) can be constructed in which a single physician controls the complete system, like in the daVinci surgical system (Intuitive Surgical Inc, Sunnyvale, CA, USA). Because the coupling between the movement of the physician and the movement of the actuators is implemented in software, it can be designed to allow intuitive control.

There exists a significant amount of friction and compliance between the tip of the instrument and its control handle (where it is actuated), resulting in hysteresis. Abbott et al. [1], Bardou [2], and Bardou et al. [3] have proposed compensation of the hysteresis in the case that the amount of hysteresis is known in advance (i.e., determined preoperatively). However, because the friction and compliance vary with the (unknown) shape of the endoscope, feedback of the actual tip position is required in order to be able to control it accurately. Adding extra sensors to the instruments to measure this tip position will be expensive, because the space at the tip of the instrument is very limited. Therefore, it would be beneficial if the tip position can be measured without adding extra sensors. This can be accomplished by using the endoscopic images as a feedback.

Pose estimation of laparoscopic instruments has been studied by Doignon et al. [7], using both marker-based and marker-less techniques. They considered a general pose estimation problem, which has no model of the kinematics of the instrument. Moreover, for the marker-less estimation, the instrument was assumed to be straight, which is true for laparoscopy, but not for flexible endoscopy. In the case of flexible endoscopy, where the instrument has only three DOFs, the use of a kinematics model significantly reduces the solution space, improving the accuracy.

In this study, we compare two methods that use the endoscopic images to estimate the pose of a flexible endoscopic instrument. The first method uses feature points that are detected on the instrument tip (marker-less). The second method uses markers that are attached to the instrument (marker-based). The contributions of this study as compared to our previous work [17,18] are the following:

- In the current study, we perform a comparison of the marker-less and marker-based methods under equal experimental conditions.
- We have used an X-ray imager to reconstruct the ground-truth position of the instrument tip. This allows for an accurate evaluation of the estimation algorithm over the entire workspace.
- For the marker-based approach, we have developed a more robust method to match the marker regions that are found in the image to the markers in the model.

This paper is structured as follows: In section “Materials and methods,” the marker-less and marker-based estimation methods are presented, and the experimental setup for evaluation of these methods is described. The experimental results are presented in section “Results.” Finally, section “Discussion” concludes with the discussion.

Materials and methods

Our approach for the pose estimation is based on virtual visual servoing [13]. In this approach, the actual state of the estimator is used to find the estimated positions of certain feature points. This is done using a kinematics model of the instrument and a model of the camera. These estimated positions are compared to the positions of feature points that are observed in the endoscopic image. Based on the difference between the estimated and the actual positions, the state of the estimator is updated such that the estimated feature point positions move toward the actual feature point positions. From the state of the estimator, the pose (position and orientation) of the instrument tip can be derived using the kinematics model of the instrument.

This section describes the kinematics model of the instrument, the model of the camera, the detection of the features from the endoscopic images, and the state estimation

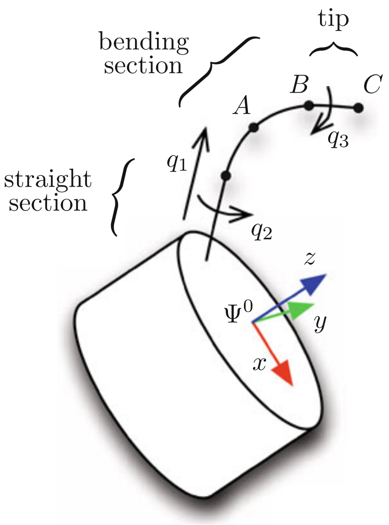


Fig. 2 The endoscopic instrument has three degrees of freedom: translation q_1, rotation q_2, and bending q_3. Points A and B are located midway and at the end of the bendable section, respectively. Point C is located at the tip. Frame Ψ^0 denotes the camera frame of the endoscopic camera

algorithms. Finally, the experimental setup that was used to evaluate the performance is presented.

Kinematics model of the instrument

The kinematics model of the instrument describes the positions of points on the instrument in the three-dimensional (3D) Euclidian space. The model consists of a straight section, a bendable section, and the tip (Fig. 2). This model is similar to that of Bardou et al. [4]. The model assumes that there are no significant forces acting on the instrument, resulting in a constant curvature along the bending section. This assumption is valid in our experiments. However, in clinical practice, external forces are present, which may have to be accounted for. These can be modeled as external disturbances to the model.

The state of our model (denoted $\mathbf{q}$) has three components: translation (q_1), rotation (q_2), and bending (q_3). We define three reference points, denoted A, B, and C, on the centerline of the instrument. A and B are located midway and at the end of the bendable section, respectively, while C is located at the tip. The model allows us to compute the positions of $A \ldots C$, denoted $\mathbf{p}_A \ldots \mathbf{p}_C$, using the forward kinematics function, denoted $f(\mathbf{q})$:

$$\begin{bmatrix} \mathbf{p}_A \\ \mathbf{p}_B \\ \mathbf{p}_C \end{bmatrix} = f(\mathbf{q}). \tag{1}$$

Additionally, we can compute the relation between the change of the state $\dot{\mathbf{q}}$ and the changes of the positions of the points $\dot{\mathbf{p}}_A \ldots \dot{\mathbf{p}}_C$:

$$\begin{bmatrix} \dot{\mathbf{p}}_A \\ \dot{\mathbf{p}}_B \\ \dot{\mathbf{p}}_C \end{bmatrix} = \mathbf{J_f}(\mathbf{q})\dot{\mathbf{q}}, \quad \text{where } \mathbf{J_f}(\mathbf{q}) := \begin{bmatrix} \frac{\partial \mathbf{p}_A}{\partial \mathbf{q}} \\ \frac{\partial \mathbf{p}_B}{\partial \mathbf{q}} \\ \frac{\partial \mathbf{p}_C}{\partial \mathbf{q}} \end{bmatrix}. \tag{2}$$

In (2), $\mathbf{J_f}$ denotes the analytical Jacobian of f. The detailed calculation of $\mathbf{J_f}$ is in the Appendix.

Endoscopic camera model

We have modeled the endoscopic camera using the pinhole camera model, with additional radial distortion. Since endoscopes have a wide-angle lens, the radial distortion is quite significant. The camera model $g(\mathbf{p})$ maps each point $\mathbf{p}$ in the 3D space to a point $\mathbf{x}$ in the 2D image space:

$$\mathbf{x} = g(\mathbf{p}). \tag{3}$$

For the marker-based method, the 2D image space positions of marker positions $A \ldots C$ are combined into the measurement vector $\mathbf{s}$:

$$\mathbf{s} = \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_B \\ \mathbf{x}_C \end{bmatrix} = \begin{bmatrix} g(\mathbf{p}_A) \\ g(\mathbf{p}_B) \\ g(\mathbf{p}_C) \end{bmatrix}. \tag{4}$$

Similar to (2), the derivative relation of (4) can be computed, showing the relation between the change of the feature point positions in 3D space $\dot{\mathbf{p}}$ and the change of the feature point positions in the 2D image space $\dot{\mathbf{x}}$:

$$\dot{\mathbf{s}} = \begin{bmatrix} \dot{\mathbf{x}}_A \\ \dot{\mathbf{x}}_B \\ \dot{\mathbf{x}}_C \end{bmatrix} = \begin{bmatrix} \mathbf{J_g}(\mathbf{p}_A)\,\dot{\mathbf{p}}_A \\ \mathbf{J_g}(\mathbf{p}_B)\,\dot{\mathbf{p}}_B \\ \mathbf{J_g}(\mathbf{p}_C)\,\dot{\mathbf{p}}_C \end{bmatrix}, \quad \text{where } \mathbf{J_g}(\mathbf{p}) := \frac{\partial g(\mathbf{p})}{\partial \mathbf{p}} \tag{5}$$

Equations (2) and (5) can be combined so as to obtain the relation between the change of the state $\dot{\mathbf{q}}$ and the change of the measurement vector $\dot{\mathbf{s}}$:

$$\dot{\mathbf{s}} = \mathbf{L}\dot{\mathbf{q}}, \tag{6}$$

where $\mathbf{L}$ is the (state-dependent) interaction matrix [6]. $\mathbf{L}$ is used by the state estimation algorithm as will be described later.

For the marker-less method, the computation of the interaction matrix is similar to the marker-based method. For the marker-less method, the marker locations $A \ldots C$ are replaced by the locations of feature points $f_1 \ldots f_3$, as described in the next section.

Feature detection

For the estimation of the instrument state, features are extracted from the acquired endoscopic images. For the marker-less method, three points on the instrument tip are used as the features. For the marker-based method, the features are the positions of the centroids of the markers in the image.

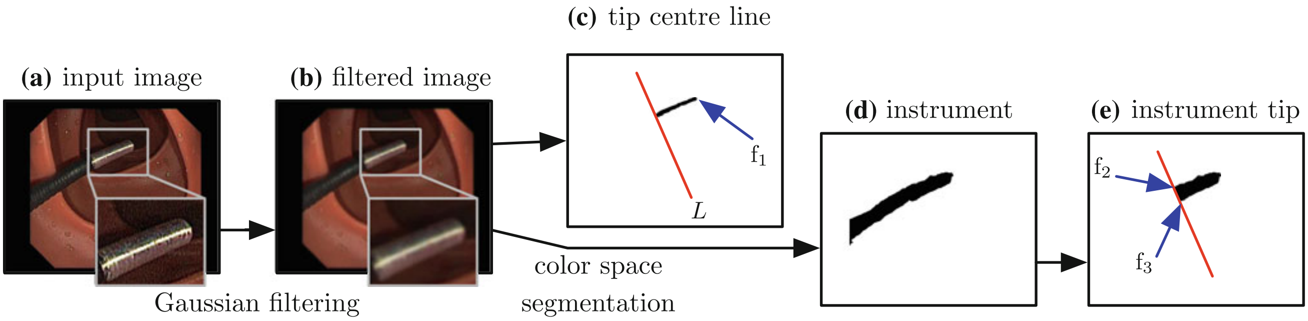


Fig. 3 Feature detection for the marker-less method: The input image (**a**) is filtered using a Gaussian filter kernel in order to remove noise. This results in image (**b**). This image is color-space-segmented twice using different parameters, resulting in the *tip centerline* (**c**) and the instrument (**d**) regions. From the *tip centerline*, the tip position is found, which is the first feature point (denoted f_1). Then, the *red line* L is determined, which is perpendicular to the *tip centerline*. This line is used to separate the instrument tip (**e**) from the instrument region (**d**). From the resulting instrument tip (**e**), two other feature points (f_2 and f_3) are detected

Marker-less feature detection

For the estimation without markers, three feature points are extracted from the endoscopic images. These are the tip of the instrument and two points on either side. It should be noted that the method could easily be expanded to take more feature points into account for increased accuracy and robustness. The extraction of the feature points is done as illustrated in Fig. 3. First, the endoscopic image is filtered using a Gaussian kernel with a scale of $\sigma = 3$ pixels (Fig. 3b). This reduces the effects of noise in the image. Then, the image is segmented using Fishers linear discriminant method [8], applied to the RGB color space. This results in a binary image of the centerline of the instrument tip (Fig. 3c). Using the same method, but with different parameters, a binary image of the complete instrument (Fig. 3d) is extracted.

The orientation of the instrument tip is computed using the singular value decomposition of the covariance matrix of the x- and y-coordinates of all points belonging to the instrument tip centerline region [17]. The largest singular value corresponds to the direction of the tip in the image. Using this principal direction, the point that is most toward the tip is selected as the first feature point f_1, as shown in Fig. 3c.

The tip direction is also used to define a line L, which is perpendicular to the tip direction, and intersects the instrument at the beginning of the tip region. L is shown in red in Fig. 3c, e. L is positioned such that it touches the binary image of the tip centerline. Using L, the instrument region (Fig. 3d) is separated, resulting in the instrument tip region (Fig. 3e). From this region, feature points f_2 and f_3 are derived.

Marker-based feature detection

The marker color was chosen to have a high contrast with the background of the image. As a result, the markers can be separated from the background relatively easily. As in the marker-less method, the endoscopic image is first filtered using a Gaussian filter. Then, color space segmentation is used to obtain a binary image of the markers [9]. The regions in this binary image are labeled using the `ndimage` module of the `scipy` package [19]. For every region, its centroid and its area are measured.

State estimation

The goal of the state estimation algorithm is to update the state of the instrument model, such that the feature points from the model match the actual features that were detected from the endoscopic images. The state estimation algorithm is similar for the marker-less and marker-based methods. However, since the features that are used are different, there are some differences in the state estimation algorithm between the two methods.

Marker-less state estimation

The algorithm of the estimator is illustrated in Fig. 4. The current state $\mathbf{q}$ of the estimator is used to compute the estimated positions of the feature points in the image space, denoted $\mathbf{s}$. This is done using the kinematic model f and the camera model g that were described earlier. Using the feature detection described in the previous section, the three feature points in the endoscopic image are found. These are denoted $\mathbf{s}^*$ in Fig. 4. The error $\mathbf{e}$ is defined as the difference between $\mathbf{s}$ and $\mathbf{s}^*$. $\mathbf{e}$ is the input to the controller, which is implemented as a multiplication by constant gain G and $\hat{\mathbf{L}}_{\mathbf{W}}^{\dagger}$, the pseudo-inverse of the interaction matrix $\mathbf{L}$. The computation of $\hat{\mathbf{L}}_{\mathbf{W}}^{\dagger}$ will be described later. The output of the controller is $\dot{\mathbf{q}}$, the desired change of state $\mathbf{q}$ that brings $\mathbf{s}$ closer to $\mathbf{s}^*$.

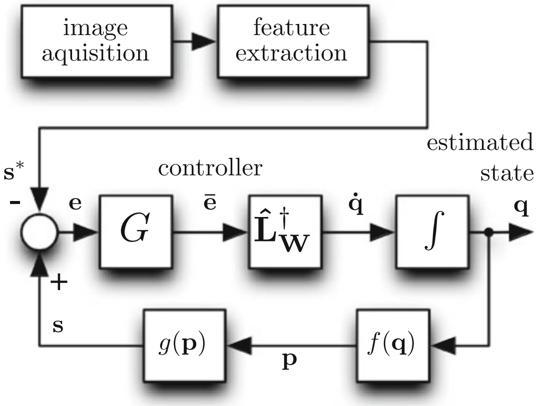


Fig. 4 Marker-less state estimation: For a given state **q**, the kinematic model f and camera model g are used to compute the expected positions of the feature points in the image space,denoteds. These are compared to $\mathbf{s}^*$, which are the positions of the feature points in the endoscopic image, as determined by the feature extraction algorithm. The difference $\mathbf{s} - \mathbf{s}^*$, denoted **e**, is input to the controller, which computes the state change $\dot{\mathbf{q}}$ to bring the model closer to the observed instrument

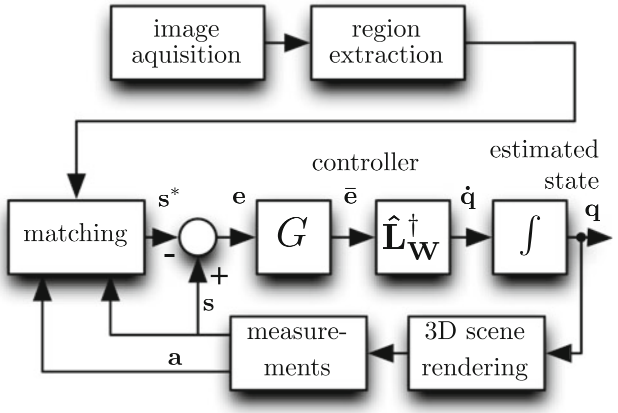


Fig. 5 Marker-based state estimation: The structure of the estimator is similar to the marker-less method, but the main difference is the extra matching step, in which the regions that are extracted from the endoscopic image are matched to the markers. For a given state **q**, a 3D rendering of the scene is created. From this scene, the centroids (denoted **s**) and areas (denoted **a**) of the rendered markers are computed. **s** is compared to $\mathbf{s}^*$, which are the centroids of the markers in the endoscopic image. The difference $\mathbf{s} - \mathbf{s}^*$, denoted **e**, is input to the controller, which computes the state change $\dot{\mathbf{q}}$ to bring the model closer to the observed instrument

Marker-based state estimation

For the marker-based method, the estimation algorithm is illustrated in Fig. 5. For this method, the features are the positions of the centroids of the markers in the image. Due to occlusion effects, these are in general not equal to the projection of the geometrical center of each marker. Therefore, in order to obtain accurate feature measurements from the model, a 3D rendering of the endoscopic instrument is created using OpenGL [16]. This ensures that occlusion effects that occur in the actual scene are also present in the model. The 3D rendering uses a camera model that replicates the severe lens distortion that is present in the endoscopic camera system. The camera parameters are obtained using the Camera Calibration Toolbox for Matlab [5]. The lens distortion and the movement of the instrument are computed on the Graphics Processing Unit (GPU) using vertex shaders. This improves the computational efficiency. The measurements **s**, which are the positions of the centroids of the markers in the rendered scene, are obtained from the rendered scene. Additionally, the areas of the markers, denoted **a** in Fig. 5, are measured. These are used by the matching algorithm as will be described below.

Due to shadows and specular reflections, the feature detection algorithm may sometimes fail to detect a marker, or detect regions which actually are not markers. Also, in clinical practice, markers may sometimes be invisible due to occlusions. In order to provide a robust matching between the regions that are found by the feature detection algorithm and the markers of the model, a maximum-likelihood approach is used [10].

We will use k to denote the number of regions found by the feature detection algorithm. We define the likelihood function $\mathcal{L}(i, j)$ as the likelihood that marker i $(i = 1 \ldots 3)$ corresponds to region j $(j = 0 \ldots k)$. $\mathcal{L}(i, 0)$ is defined as the likelihood that marker i is missing (i.e., not detected by the feature detection algorithm).

Given the individual likelihoods $\mathcal{L}(i, j)$, the total likelihood $\mathcal{L}_T$ that the three markers in the model are represented by, respectively, regions r, s, and t is given by:

$$\mathcal{L}_T(r, s, t) = \mathcal{L}(1, r)\mathcal{L}(2, s)\mathcal{L}(3, t), \tag{7}$$

where r, s, and t $(0 \ldots k)$ denote the regions that are selected as representing the first, the second, and the third marker, respectively.

The state estimator matches the regions to the markers by finding the maximum $\mathcal{L}_T(r, s, t)$ under the condition

$$r \neq s, \quad s \neq t, \quad r \neq t. \tag{8}$$

The positions of the centroids of the resulting regions r, s, and t are combined into measurement vector $\mathbf{s}^*$. $\mathbf{s}^*$ is used as an input to the virtual visual servoing loop just as in the marker-less approach.

The likelihood function $\mathcal{L}(i, j)$ was chosen to be a function of the Euclidian distance between the position of the region and the position of the marker, and the ratio between the area of the region and the area of the marker:

$$\mathcal{L}(i, j) := \begin{cases} \mathcal{L}_M(i), & j = 0 \\ \mathcal{L}_D(i, j)\, \mathcal{L}_A(i, j), & j \neq 0 \end{cases}. \tag{9}$$

In (9), $\mathcal{L}_D(i, j)$ denotes a likelihood function that is dependent on the Euclidian distance between the position of marker i and region j $(j = 1 \ldots k)$. $\mathcal{L}_A(i, j)$ denotes a likelihood function that depends on the ratio between the area of marker i and region j $(j = 1 \ldots k)$. $\mathcal{L}_M(i)$ denotes the constant likelihood that marker i is missing (i.e., it was not detected by the feature detection algorithm). $\mathcal{L}_D$ and $\mathcal{L}_A$ were chosen as exponential functions, since this matched the distributions that were observed during the actual experiment.

The distance-dependent likelihood function $\mathcal{L}_D$ is

$$\mathcal{L}_D(i, j) := \exp\left(-\frac{||\mathbf{x}_{\mathrm{m}}(i) - \mathbf{x}_{\mathrm{r}}(j)||}{\sigma_D}\right), \tag{10}$$

where $\mathbf{x}_{\mathrm{m}}(i)$ and $\mathbf{x}_{\mathrm{r}}(j)$ denote the position of the centroid of marker i and region j in the image, respectively (subscript m for marker and r for region). $||\cdot||$ denotes the Euclidian distance. σ_D is a parameter that controls the decay of the exponential function.

The area-dependent likelihood function $\mathcal{L}_A$ is as follows:

$$\mathcal{L}_A(i, j) := \exp\left(-\frac{\left|\log\left(\frac{a_{\mathrm{m}}(i)}{a_{\mathrm{r}}(j)}\right)\right|}{\sigma_A}\right), \tag{11}$$

in which $a_{\mathrm{m}}(i)$ and $a_{\mathrm{r}}(j)$ denote the area of marker i and region j in the image, respectively. σ_A is a parameter that controls the decay of the exponential function. Note that $\mathcal{L}_A(i, j)$ can alternatively be written as:

$$\mathcal{L}_A(i, j) = \begin{cases} \left(\frac{a_{\mathrm{r}}}{a_{\mathrm{m}}}\right)^{\frac{1}{\sigma_A}}, & a_{\mathrm{r}} < a_{\mathrm{m}} \\ \left(\frac{a_{\mathrm{m}}}{a_{\mathrm{r}}}\right)^{\frac{1}{\sigma_A}}, & a_{\mathrm{r}} \geq a_{\mathrm{m}} \end{cases}. \tag{12}$$

This shows that $\mathcal{L}_A(i, j)$ is an exponential function with the ratio between a_{m} and a_{r} as its base.

The likelihood functions $\mathcal{L}_D$ and $\mathcal{L}_A$, and the parameters σ_D, σ_A, and $\mathcal{L}_M(i)$ were chosen so as to represent the distribution of the distances and area ratios that were observed during the actual experiment.

Controller

For both the marker-less and the marker-based method, the visual servo loop contains a controller consisting of a proportional gain G and the pseudo-inverse of the interaction matrix, denoted $\mathbf{L}_{\mathbf{W}}^{\dagger}$. For the marker-based approach, the matrix $\mathbf{L}$ in (6) is an approximation. $\mathbf{L}$ relates to the positions of the center of each of the markers, while actually the centroids of the projections of the markers are used as measurements. Note that this approximation is only used for the computation of the interaction matrix, not in the actual virtual visual servo control loop.

For a given error $\mathbf{e}$ between $\mathbf{s}$ and $\mathbf{s}^*$, the proportional gain G yields the desired change of error $\mathbf{e}$ (denoted $\bar{\mathbf{e}}$) that will cause $\mathbf{e}$ to decrease: $\bar{\mathbf{e}} = -G\mathbf{e}$, with G a positive scalar constant.

Since the dimension of $\bar{\mathbf{e}}$ (six) is higher than the dimension of $\mathbf{q}$ (three), it is in general not possible to find a $\dot{\mathbf{q}}$ that will result in the desired change of error (i.e., that results in $\dot{\mathbf{e}} = \bar{\mathbf{e}}$). Therefore, we use the weighted Moore–Penrose pseudo-inverse of the approximated interaction matrix to obtain the state change $\mathbf{q}$ that minimizes the weighted error $||\mathbf{W}(\dot{\mathbf{e}} - \bar{\mathbf{e}})||_2$, where $\mathbf{W}$ denotes a weighting matrix [15]:

$$\hat{\mathbf{L}}_{\mathbf{W}}^{\dagger} := \left(\mathbf{L}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{L}\right)^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}. \tag{13}$$

For the marker-less method, the identity matrix is used for $\mathbf{W}$. For the marker-based method, we take advantage of the likelihoods that were computed for the matching between the regions and the markers. We use a weighting matrix:

$$\mathbf{W} = \mathrm{diag}(\mathcal{W}(1, r), \mathcal{W}(1, r), \mathcal{W}(2, s), \mathcal{W}(2, s), \mathcal{W}(3, t), \mathcal{W}(3, t)), \tag{14}$$

with

$$\mathcal{W}(i, j) := \begin{cases} 0, & j = 0 \\ \mathcal{L}(i, j), & j \neq 0 \end{cases}. \tag{15}$$

The definition of $\mathcal{W}(i, j)$ ensures that if there was no match found for a given marker $(j = 0)$, a weight of 0 is used. If a match was found $(j \neq 0)$, markers with a higher likelihood are weighted more than those with a lower likelihood. Because of (8), only one of the markers can have a zero weight. This ensures the term $\mathbf{L}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{L}$ in (13) remains full rank and therefore invertible.

Since the estimation methods are iterative, an initialization is required. Currently, this initialization is done by starting the experiment with the instrument in a known position. In our proposed application, where the instrument is robotically actuated, the (known) state of the actuators may be used to initialize the estimation to a state that is close to the actual state.

Experimental evaluation

In order to evaluate the pose estimation system that was described in the previous sections, experiments were conducted. A flexible endoscopic instrument was operated inside a colon model, and the tip position was estimated. This was compared to a reference tip position which was obtained using an X-ray imager. Although the proposed methods can also estimate the orientation of the tip, the orientation was not evaluated since an accurate ground-truth orientation was not available.

Figure 6 shows the experimental setup that was constructed to evaluate the performance of the marker-less and

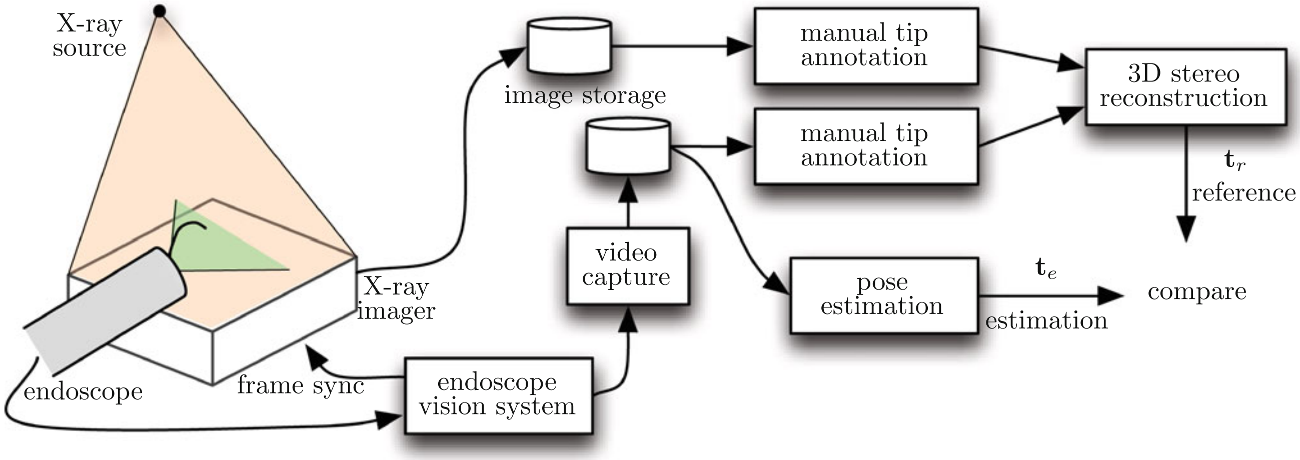


Fig. 6 The estimator was evaluated using an X-ray imaging setup. Images from the endoscopic camera and the X-ray imager were synchronously acquired and stored. In both the X-ray images and the endoscopic images, the tip position was manually annotated. From these positions, the 3D reference position $\mathbf{t}_r$ was constructed. This was compared to the estimated 3D position $\mathbf{t}_e$ as obtained from the pose estimation. During the experiment, the endoscope and the instrument were inside a colon model. The colon model is not shown in the figure for clarity

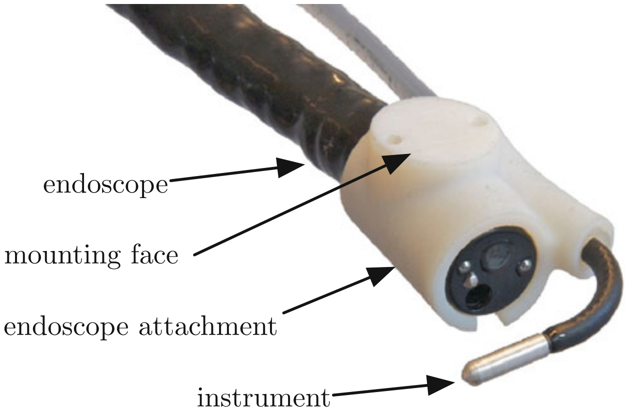


Fig. 7 An endoscope attachment was designed to let the endoscopic instrument emerge near the endoscopic camera, similar to the Anubis endoscope. The attachment also has a mounting face, which enables the endoscope to be fixed inside the X-ray imaging setup

marker-based methods. The endoscope was stationary during the experiment. An endoscope attachment was designed to locate the endoscopic instrument near the endoscope tip (Fig. 7). In order to obtain a reference measurement of the tip position, an X-ray imaging setup was used (Fig. 8). The X-ray imager was positioned such that a top view of the scene was obtained. The X-ray imager and the endoscopic camera were used as a stereo camera rig, enabling reconstruction of the tip position in 3D. The acquired images of the X-ray imager were 1,024×768 pixels, resulting in a resolution of 0.25 mm per pixel. The resolution of the endoscopic images was 720×576 pixels. The X-ray imager was synchronized to the endoscopic camera, using the synchronization information that is available in the composite video output of the endoscopic camera unit. Both image sequences were stored for the processing, which was performed off-line.

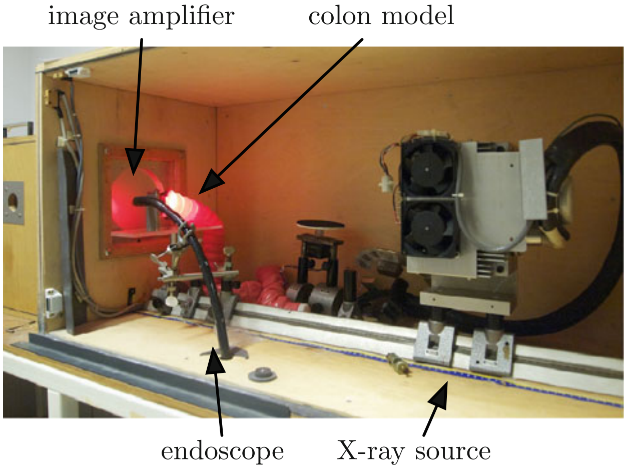
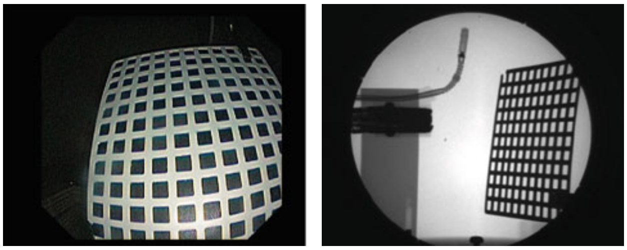


Fig. 8 A custom-built X-ray imaging setup was used for the experiment. The X-ray source generates the X-rays which are captured by the image amplifier. The images are digitized by a digital camera (not visible in the image). The endoscope is positioned inside a colon model during the experiment

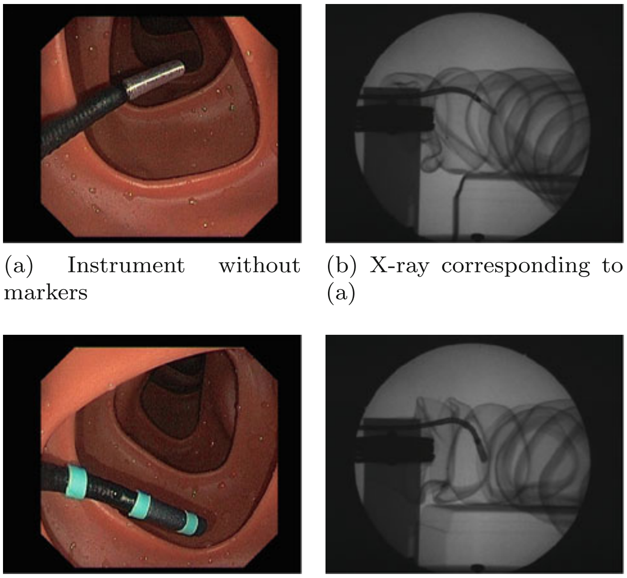
For both the endoscopic image and the X-ray images, the tip position was manually annotated in each frame. From the 2D tip position in the X-ray and the endoscopic images, the 3D tip position was reconstructed using the Camera Calibration Toolbox for Matlab [5]. The stereo rig had been calibrated before the experiment. A punched metal sheet was used for the calibration as a substitute for the more commonly used checkerboard pattern, because this sheet could be clearly imaged using both imaging modalities (Fig. 9).

A conventional colonoscope (Exera, Olympus Imaging Corp, Tokyo, Japan) was used in the experiment. The images were captured using the FireWire output of the colonoscope imaging unit. The Anubis endoscopic instruments (Karl Storz GmbH & Co. KG, Tuttlingen, Germany) were manually



(a) Endoscopic image **(b)** X-ray image

Fig. 9 The stereo rig composed of **a** the endoscope camera and **b** the X-ray imager is calibrated by imaging a reference object using both image modalities. In (**a**), it can be observed that severe barrel distortion is present in the endoscopic images



(a) Instrument without markers (b) X-ray corresponding to (a)

(c) Instrument with markers (d) X-ray corresponding to (c)

Fig. 10 During the experiment, the instrument was manually operated while the endoscope was inside the colon model. A top view of the scene was simultaneously imaged using an X-ray imager. In (**a**) and (**b**), endoscopic and X-ray images of the experiment without markers are shown. In (**c**) and (**d**) images of the experiment with markers are shown

operated. The experiment was performed inside a colon model (KKM40, Kyoto Kagaku, Kyoto, Japan) that is commonly used for colonoscopy training. A viscous fluid was used to coat the inside of the model as per the manufacturers instructions, in order to replicate the lighting conditions of clinical images. Specifically, this fluid causes specular reflections which are also commonly present in clinical images.

Results

Figure 10a, b shows endoscopic and X-ray images of the instrument, respectively, while it was operated inside the colon model during the marker-less experiment. Figure 11 shows the results of the marker-less pose estimation. It shows the x-, y-, and z-components of the estimated tip position, and the reference as obtained by the 3D reconstruction from the X-ray and endoscopic images. The positions are expressed in the camera frame Ψ^0 (Fig. 2). The root-mean-square (RMS) differences between the estimated and the reference position were 1.5, 1.6, and 1.8 mm in the x-, y-, and z-directions, respectively.

Figure 10c, d shows endoscopic and X-ray images for the marker-based estimation experiment, respectively. Figure 12 shows the position estimation results for the marker-based estimation. For the marker-based method, the RMS differences between the estimated and the reference position were 1.1, 1.7, and 1.5 mm in the x-, y-, and z-directions, respectively.

The two methods were compared statistically using the Mann–Whitney–Wilcoxon test [12]. The experimental data were subsampled at 5-s intervals in order to prevent unacceptable dependence between the samples, resulting in 35 samples for each method. No significant differences between the methods were found ($p = 0.2$).

Discussion

Two methods were compared for estimating the pose of an endoscopic instrument, one with and one without markers on the instrument. The methods were tested inside a colon model, and the accuracy of the estimated tip position was evaluated using an X-ray imager to provide a ground-truth value. Both methods were able to track the motions of the endoscopic instrument and performed similarly in terms of tip position accuracy. No significant difference between the methods was found in terms of accuracy. The kinematics model can also be used to derive the tip orientation. However, this was not evaluated in this study.

For the marker-based method, a maximum-likelihood approach was used to match the regions in the endoscopic image to the markers of the model of the endoscopic instrument. This approach makes the state estimator robust against missing markers that may be caused by, for example, occlusions or shadows. This is a potential advantage over the marker-less method. However, we have not evaluated the robustness of the two methods in the current study. Also, the computed likelihood value gives a measure of how reliable the estimated position is. An alternative control method for the instrument could be used as a backup if the likelihood is too low. This would create the robustness that is required for the system to be implemented in clinical practice. Another advantage of the marker-based method is that the apparent size of the markers could be used as a cue for the z-position of each marker. In this case, the area of

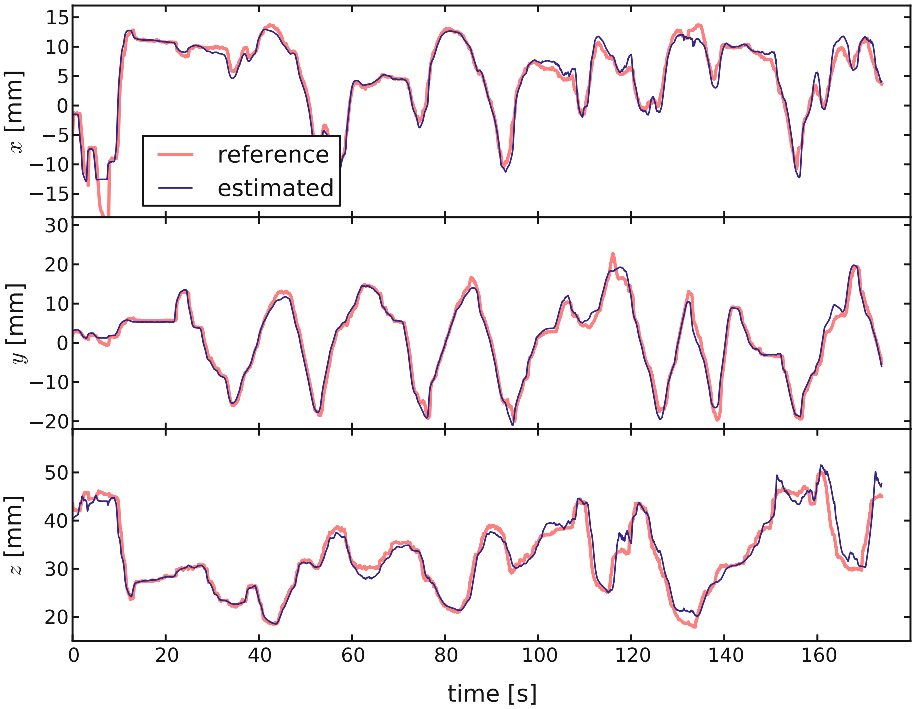


Fig. 11 Marker-less estimation results: The graphs show the x-, y-, and z-coordinates of the estimated tip position, and the reference that was obtained using the X-ray imager. The RMS errors were 1.5, 1.6, and 1.8 mm in the x-, y-, and z-directions, respectively

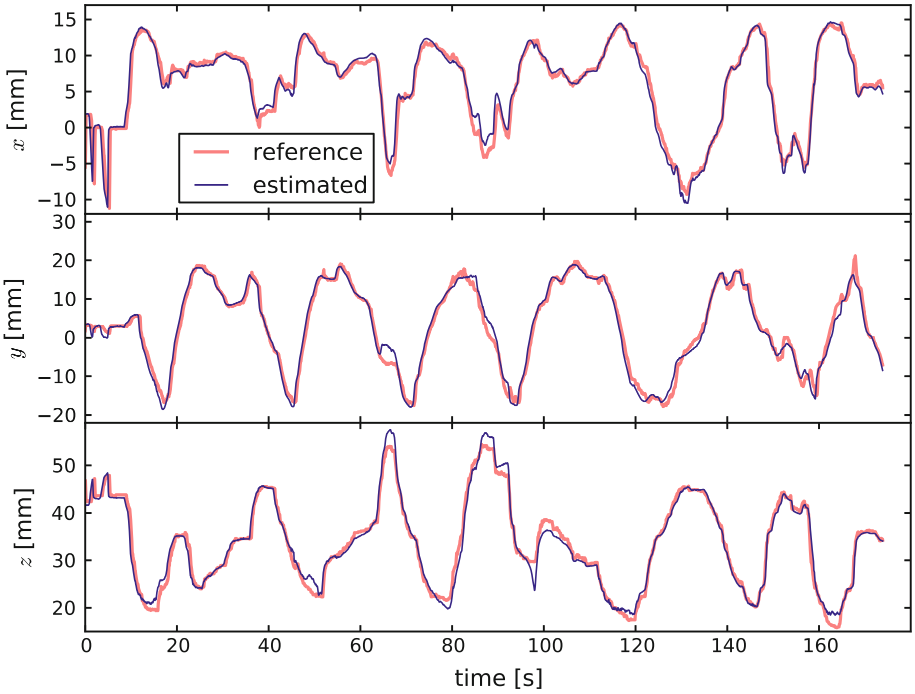


Fig. 12 Marker-based estimation results: The accuracy for the marker-based estimation is similar to the accuracy for the marker-less method. The RMS errors were 1.1, 1.7, and 1.5 mm in the x-, y-, and z-directions, respectively

each marker would be included in the vector **s** in (4). This could improve the estimation accuracy, especially in the z-direction.

An advantage of the marker-less method is that current instruments can be used without adding any markers. However, it might be necessary to adapt the feature detection algorithm depending on the type of instrument that is used.

For the future work, our goals are twofold. Firstly, we want to test the performance of the algorithms under various lighting conditions and in the presence of occlusions. Secondly, we plan to add actuators to the endoscopic instrument. The estimation methods that were developed will be used as a feedback in order to be able to obtain accurate and intuitive control of the instrument. This will be incorporated in an endoscope system in which a single physician is able to control all the DOFs of the endoscope and the instrument. Such a system will enable advanced endoscopic procedures to be performed accurately and efficiently.

Acknowledgments This research is conducted within the TeleFLEX project, which is funded by the Dutch Ministry of Economic Affairs and the Province of Overijssel, within the Pieken in de Delta (PIDON) initiative. The ANUBIS endoscopic instrument was provided by Karl Storz GmbH & Co. KG, Tuttlingen, Germany. The endoscope was provided by Olympus Corp., Tokyo, Japan.

Appendix

Here we show the derivation of the analytical Jacobian $\mathbf{J_f}(\mathbf{q})$ of the forward kinematics function $f(\mathbf{q})$ in (2). We define five frames on the instrument (Fig. 13). Frame Ψ^0 is the camera frame, with the z-axis in the direction of the camera optical axis. Frame Ψ^1 is located at the point where the instrument emerges from the endoscope, with the z-axis aligned with the instrument direction. Frame Ψ^2 is at the end of the straight section, rotating with the instrument rotation q_2. Frame Ψ^3 is midway the bending section, and frame Ψ^4 is at the end of the bending section.

We first derive the unit twists of frames Ψ^2, Ψ^3, and Ψ^4 associated with each of the three DOFs. We denote the motion of frame Ψ^l with respect to frame Ψ^m, expressed in frame Ψ^k as the infinitesimal twist $\mathbf{T}_l^{k,m}$. We denote the unit twist of frame Ψ^l associated with q_j, with respect to frame Ψ^0, expressed in frame Ψ^0 as $\hat{\mathbf{T}}_{l,j}$. From the unit twists, the Jacobian $\mathbf{J_f}(\mathbf{q})$ is derived.

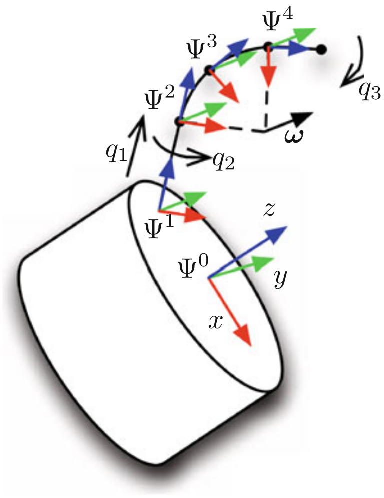


Fig. 13 Five frames are defined: frame Ψ^0 and Ψ^1 are fixed to the endoscope, while frame Ψ^2, Ψ^3 and Ψ^4 are fixed along the instrument. q_1, q_2, and q_3 denote the three DOFs: insertion, rotation, and bending, respectively

Straight section

The pose of frame Ψ^2, located at the end of the straight section, is defined by q_1 and q_2, which are a translation along the z-axis of frame Ψ^1 and a rotation around the same axis, respectively. Thus, the pose of frame Ψ^2 with respect to frame Ψ^1 is given by:

$$ {}_2^1\mathbf{H} = \begin{bmatrix} \mathbf{R_z}(q_2) & \begin{matrix}0\\0\\q_1\end{matrix} \\ 0\;0\;0 & 1 \end{bmatrix}, \tag{16}$$

where $\mathbf{R_z}(\cdot)$ denotes the 3-by-3 rotation matrix around the z-axis. The pose of frame Ψ^1 with respect to frame Ψ^0 is determined by the geometry of the endoscope and is thus fixed.

The motion of frame Ψ^2 with respect to frame Ψ^0 is described by the infinitesimal twist:

$$\mathbf{T}_2^{0,0} = \hat{\mathbf{T}}_{2,1}\dot{q}_1 + \hat{\mathbf{T}}_{2,2}\dot{q}_2\,, \tag{17}$$

where $\hat{\mathbf{T}}_{2,1}$ and $\hat{\mathbf{T}}_{2,2}$ represent a translation along the z-axis of frame Ψ^1 and a rotation around that z-axis, respectively. They are:

$$\hat{\mathbf{T}}_{2,1} = \mathrm{Ad}_{{}_1^0\mathbf{H}}\left[0\;0\;0\;0\;0\;1\right]^{\mathrm{T}} \tag{18}$$

$$\hat{\mathbf{T}}_{2,2} = \mathrm{Ad}_{{}_1^0\mathbf{H}}\left[0\;0\;1\;0\;0\;0\right]^{\mathrm{T}}, \tag{19}$$

where $\mathrm{Ad}_{{}_1^0\mathbf{H}}$ denotes the Adjoint operator that changes the coordinates of the twist from frame Ψ^1 to frame Ψ^0.

Bending section

The bending section is modeled as a constant curvature. It can be defined by a finite twist around axis $\boldsymbol{\omega} = \left[0\;\omega\;0\right]^{\mathrm{T}}$ (Fig. 13), where ω is the angle of the arc. The axis $\boldsymbol{\omega}$ is in the y-direction of frame Ψ^2, located at $\left[\rho\;0\;0\right]^{\mathrm{T}}$ in frame Ψ^2, where ρ denotes the curve radius. The chord length, denoted ℓ, is given by $\ell = \omega\rho$. q_3 is defined as $q_3 := \omega$. This results in the finite twist describing the bending section:

$$\mathbf{S}_4^{2,2} = \begin{bmatrix} \boldsymbol{\omega} \\ \begin{bmatrix}\rho\\0\\0\end{bmatrix} \wedge \boldsymbol{\omega} \end{bmatrix} = \begin{bmatrix}0\\q_3\\0\\0\\0\\\ell\end{bmatrix}, \tag{20}$$

where $\mathbf{S}_4^{2,2}$ denotes the finite twist of frame Ψ^4 with respect to frame Ψ^2 expressed in frame Ψ^2. The infinitesimal twist $\mathbf{T}_4^{2,2}$ can be derived from the finite twist $\mathbf{S}_4^{2,2}$ using the definition of the twist in matrix form (denoted by the tilde: $\tilde{\mathbf{T}}_l^{k,m}$):

$$\tilde{\mathbf{T}}_4^{2,2} := {}^2_4\dot{\mathbf{H}}\,{}^4_2\mathbf{H} \tag{21}$$

$$= \frac{\partial\,{}^2_4\mathbf{H}}{\partial q_3}\dot{q}_3 \exp\left(\tilde{\mathbf{S}}_4^{2,2}\right) \tag{22}$$

$$= \begin{bmatrix} 0 & 0 & 1 & \frac{\ell}{{q_3}^2}(-1+\cos q_3) \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & \frac{\ell}{{q_3}^2}(q_3 - \sin q_3) \\ 0 & 0 & 0 & 0 \end{bmatrix} \dot{q}_3. \tag{23}$$

The unit twist $\hat{\mathbf{T}}_{4,3}$ is found by writing (23) in vector form, and transforming it to frame Ψ^0:

$$\hat{\mathbf{T}}_{4,3} = \mathrm{Ad}_{{}^0_2\mathbf{H}} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \frac{\ell}{{q_3}^2}(-1+\cos q_3) \\ 0 \\ \frac{\ell}{{q_3}^2}(q_3 - \sin q_3) \end{bmatrix} \tag{24}$$

Since frame Ψ^3 is located midway the bending section, unit twist $\hat{\mathbf{T}}_{3,3}$ is found by substituting ℓ by $\frac{\ell}{2}$ in (24).

The velocity of a point $\mathbf{p}_i$, that is fixed to frame Ψ^l, is [20] as follows:

$$\dot{\mathbf{p}}_i = \tilde{\mathbf{T}}_l^{0,0}\mathbf{p}_i\ , \tag{25}$$

with respect to frame Ψ^0 and expressed in frame Ψ^0. Since point A (Fig. 2) is fixed to frame Ψ^3, and point B and C are fixed to frame Ψ^4, the Jacobian $\mathbf{J_f}$ is as follows:

$$\mathbf{J_f} = \begin{bmatrix} \tilde{\hat{\mathbf{T}}}_{3,1}\mathbf{p}_A & \tilde{\hat{\mathbf{T}}}_{3,2}\mathbf{p}_A & \tilde{\hat{\mathbf{T}}}_{3,3}\mathbf{p}_A \\ \tilde{\hat{\mathbf{T}}}_{4,1}\mathbf{p}_B & \tilde{\hat{\mathbf{T}}}_{4,2}\mathbf{p}_B & \tilde{\hat{\mathbf{T}}}_{4,3}\mathbf{p}_B \\ \tilde{\hat{\mathbf{T}}}_{4,1}\mathbf{p}_C & \tilde{\hat{\mathbf{T}}}_{4,2}\mathbf{p}_C & \tilde{\hat{\mathbf{T}}}_{4,3}\mathbf{p}_C \end{bmatrix}. \tag{26}$$

Note that $\tilde{\hat{\mathbf{T}}}_{3,1} = \tilde{\hat{\mathbf{T}}}_{4,1} = \tilde{\hat{\mathbf{T}}}_{2,1}$ and $\tilde{\hat{\mathbf{T}}}_{3,2} = \tilde{\hat{\mathbf{T}}}_{4,2} = \tilde{\hat{\mathbf{T}}}_{2,2}$ since the poses of frame Ψ^3 and Ψ^4 with respect to frame Ψ^2 are independent of q_1 and q_2.

Conflict of interest This is to certify that the authors have no financial or personal relationships with other people or organizations that would inappropriately influence our work.

References

1. Abbott D, Becke C, Rothstein R, Peine W (2007) Design of an endoluminal NOTES robotic system. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS). San Diego, pp 410–416 doi:10.1109/IROS.2007.4399536
2. Bardou B (2011) Développement et étude d'un système robotisé pour l'assistance à la chirurgie transluminale. Ph.D. thesis, Université de Strasbourg
3. Bardou B, Nageotte F, Zanne P, De Mathelin M (2012) Improvements in the control of a flexible endoscopic system. In: Proceedings of the IEEE international conference on robotics and automation (ICRA). St. Paul, pp 3725–3732
4. Bardou B, Nageotte F, Zanne P, de Mathelin M (2009) Design of a telemanipulated system for transluminal surgery. In: 31st annual international conference of the IEEE EMBS
5. Bouguet JY Camera calibration toolbox for Matlab. http://www.vision.caltech.edu/bouguetj/calib_doc
6. Chaumette F, Hutchinson S (2006) Visual servo control. I. basic approaches. IEEE Robot Autom Mag 13(4):82–90. doi:10.1109/MRA.2006.250573
7. Doignon C, Nageotte F, Maurin B, Krupa A (2008) Pose estimation and feature tracking for robot assisted surgery with medical imaging. In: Kragic D, Kyrki V (eds) Unifying perspectives in computational and robot vision. Springer, Berlin, pp 79–101
8. Fisher R (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188
9. Gonzalez RC, Woods RE (2002) Digital image processing. Prentice Hall, Englewood Cliffs
10. Harris J, Stocker H (1998) Handbook of mathematics and computational science. Springer, Berlin
11. Kalloo AN et al (2004) Flexible transgastric peritoneoscopy: a novel approach to diagnostic and therapeutic interventions in the peritoneal cavity. Gastrointest Endosc 60(1):114–117. doi:10.1016/S0016-5107(04)01309-4
12. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50–60
13. Marchand É, Chaumette F (2002) Virtual visual servoing: a framework for real-time augmented reality. Eurographics 21(3):289–298
14. Marescaux J, Dallemagne B, Perretta S, Wattiez A, Mutter D, Coumaros D (2007) Surgery without scars: report of transluminal cholecystectomy in a human being. Arch Surg 142(9):823–826
15. Nakamura Y (1991) Advanced robotics, redundancy and optimization. Addison-Wesley, Reading
16. OpenGL: The industry standard for high performance graphics. http://www.opengl.org
17. Reilink R, Stramigioli S, Misra S (2011) Three-dimensional pose reconstruction of flexible instruments from endoscopic images. In: IEEE/RSJ international conference intelligent robots and systems (IROS). San Francisco, pp 2076–2082
18. Reilink R, Stramigioli S, Misra S (2012) Pose reconstruction of flexible instruments from endoscopic images using markers. In: Proceedings of the IEEE international conference on robotics and automation (ICRA). St. Paul, pp 2939–2943
19. SciPy: scientific tools for Python. http://www.scipy.org
20. Stramigioli S, Bruyninckx H (2001) Geometry and screw theory for robotics. Seoul