A Method for Tracking the Camera Motion of Real Endoscope by Epipolar Geometry Analysis and Virtual Endoscopy System

Kensaku Mori[1,2], Daisuke Deguchi[2], Jun-ichi Hasegawa[3], Yasuhito Suenaga[2], Jun-ichiro Toriwaki[2], Hirotsugu Takabatake[4], and Hiroshi Natori[5]

[1] Image Guidance Laboratory, Department of Neurosurgery, Stanford University,
[2] Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
{ddeguchi,mori,toriwaki,suenaga}@nuie.nagoyap-u.ac.jp
[3] School of Computer and Cognitive Sciences, Chukyo University, Toyota, Japan
[4] Minami-ichijyo Hospital, Sapporo, Japan
[5] School of Medicine. Sapporo Medical University, Sapporo, Japan

Abstract. This paper describes a method for tracking the camera motion of a real endoscope by epipolar geometry analysis and image-based registration. In an endoscope navigation system, which provides navigation information to a medical doctor during an endoscopic examination, tracking the camera motion of the endoscopic camera is one of the fundamental functions. With a flexible endoscope, it is hard to directly sense the position of the camera, since we cannot attach a positional sensor at the tip of the endoscope. The proposed method consists of three parts: (1) calculation of corresponding point-pairs of two time-adjacent frames, (2) coarse estimation of the camera motion by solving the epipolar equation, and (3) fine estimation by executing image-based registration between real and virtual endoscopic views. In the method, virtual endoscopic views are obtained from X-ray CT images of real endoscopic images of the same patient. To evaluate the method, we applied it a real endoscopic video camera and X-ray CT images. The experimental results showed that the method could track the motion of the camera satisfactorily.

1 Introduction

An endoscope is a tool for observing the inside of a human body. A medical doctor inserts an endoscope inside a patient for diagnosis. The endoscope is controlled by only watching a TV monitor. Accordingly, it is hard to know the precise current location of the endoscopic camera. If it were possible to track the camera motion of an endoscope and display the current location to a medical doctor during an endoscopic examination in real time, this would be help the medical doctor greatly.

Virtual endoscope systems (VESs) are now widely used as systems that visualize the inside of the human body based on 3-D medical images in the medical

W. Niessen and M. Viergever (Eds.): MICCAI 2001, LNCS 2208, pp. 1–8, 2001.

field [1]. The user of a VES can fly through inside an organ or perform some quantitative measurements. The user can also create an examination path for a biopsy. The VES visualizes not only information about the target organ's wall but also information beyond the target organ by employing a translucent display technique. If we could fuse real and virtual endoscopes in the examination room, we could help medical doctors more easily operate endoscopes. Endoscope navigation systems, which provide navigation information to medical doctors during endoscopic examinations, should have two fundamental functions: (a) tracking of the camera motion of the endoscope and (b) presentation of the navigation information.

There are two types of endoscopes, i.e., rigid endoscopes and flexible endoscopes. It is difficult to acquire the tip position of a flexible endoscope by using an external positional sensor, since the body of the flexible endoscope can be bent into any form. Attaching a sensor at the tip is also hard due to space limitations.

Several research groups have already reported methods of registering real and virtual endoscopic images by using image-based registration techniques [2,3]. Bricault et al. [2] performed a frontier-work on the registration of real and virtual endoscopic images. They used branching structure information of the bronchi. This registration was statically performed. Unfortunately, they did not consider the continuous tracking of the camera motion. Our group also reported a method for the continuous tracking of the camera motion by using optical flow analysis and image-based registration [5]. We detected only the forward and backward motion of the endoscope from optical flow patterns. The image-based registration was performed as the precise estimation of the endoscope's position. The main problem of this previous methods was that it could not estimate the camera motion correctly around areas where there were no feature structures such as a branching structure. For example, the previous method could not detect the rotational motion inside a tube when we could not see branching structures.

This paper describes a method for tracking the camera motion of a flexible endoscope by using epipolar geometry analysis and image-based registration between real endoscopic (RE) and virtual endoscopic (VE) views enabling us to track the camera motion continuously. In Section 2, we briefly explain the epipolar geometry analysis, which is a fundamental theory of this paper. We describe the details of the proposed method in Section 3. Experimental results and discussions are shown in Section 4.

2 Recovering the Camera Motion

2.1 Epipolar Geometry

The proposed method directly estimates the camera motion by using the epipolar geometry [6]. The epipolar geometry defines the relationship of cameras located at two different positions. Figure 1 is an illustration that explains the relationship of two position-varied cameras. Let $\mathbf{X}$ be a point located in a 3-D space, and $\mathbf{C}$

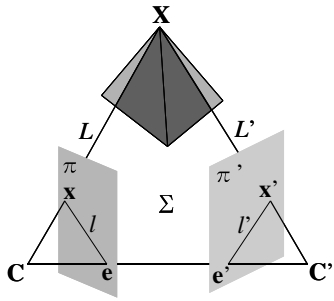

Fig.1 Epipolar geometry.

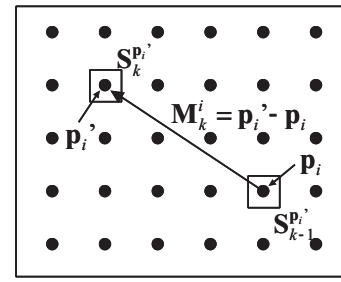

Fig.2 Finding a corresponding point-pair.

and $\mathbf{C}'$ the locations of the two cameras. The plane Σ is a plane defined by $\mathbf{C}$, $\mathbf{C}'$, and $\mathbf{X}$. $\mathbf{x}$ and $\mathbf{x}'$ are projected points of $\mathbf{X}$ on the image planes π and π'. Any point located on the line $\mathbf{L}$ should always be projected to the same point $\mathbf{x}$. The line $\mathbf{l}'$ is a projected line of $\mathbf{L}$ on the image plane π' defined by the camera $\mathbf{C}'$. $\mathbf{L}'$ and $\mathbf{l}$ are defined in the same way. The lines $\mathbf{l}$ and $\mathbf{l}'$ are called "epipolar lines". It is obvious that the epipolar lines pass the points $\mathbf{e}$ and $\mathbf{e}'$, called "epipoles", on each projection plane. The epipoles show the relationship between the two viewpoints $\mathbf{C}$ and $\mathbf{C}'$.

2.2 Epipolar Equation

Let us assume that the initial camera position $\tilde{\mathbf{x}}$ moves to $\tilde{\mathbf{x}'}$ by the rotational motion $\mathbf{R}$ and the translational motion $\mathbf{T}$. In this case, the relationship between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}'}$ is expressed by the following equation.

$$\lambda' \tilde{\mathbf{x}'} = \lambda R \tilde{\mathbf{x}} + \eta \mathbf{T}, \tag{1}$$

where λ and η are constants. Figure 2 shows the relation of each vector in Eq. (1). Since $\lambda' \tilde{\mathbf{x}'}$, $\lambda \mathbf{R} \tilde{\mathbf{x}}$, and $\eta \mathbf{T}$ exist in the same plane, the following equation should be satisfied.

$$\tilde{\mathbf{x}'} \cdot (\mathbf{T} \times \mathbf{R} \tilde{\mathbf{x}}) = 0, \tag{2}$$

where $\cdot$ and $\times$ mean the inner product and the cross product, respectively. Equation (2) can be rewritten as Eq.(3) by using $\mathbf{T} \times \mathbf{R} \tilde{x} = [\mathbf{T}]_\times \mathbf{R} \tilde{\mathbf{x}}$, $\mathbf{E} = [\mathbf{T}]_\times \mathbf{R}$, and $[\mathbf{T}]_\times = \begin{pmatrix} 0 & -T_3 & T_2 \\ T_3 & 0 & -T_1 \\ -T_2 & T_1 & 0 \end{pmatrix}$.

$$\tilde{\mathbf{x}'}^T \mathbf{E} \tilde{\mathbf{x}} = 0. \tag{3}$$

$\mathbf{E}$ is called the "Essential matrix". When we denote $\mathbf{A}$ as the camera parameter matrix defined by the focal length, the unit length of each axis, the center position of an image, and the angle of each axis, Eq. (3) is rewritten by using $\tilde{\mathbf{m}} = \mathbf{A} \tilde{\mathbf{x}}$ and $\tilde{\mathbf{m}'} = \mathbf{A} \tilde{\mathbf{x}'}$ as

$$\tilde{\mathbf{m}'}^T \mathbf{F} \tilde{\mathbf{m}} = 0. \tag{4}$$

Here

$$\mathbf{E} = \mathbf{A}^T \mathbf{F} \mathbf{A}. \tag{5}$$

If corresponding point-pairs of two views captured by two cameras are inputted, we can calculate **F** (fundamental matrix) from Eq (4). If the camera parameter **A** is known, we can calculate **E** by Eq. (5). Equations (3) and (4) are called "epipolar equations".

3 Methods

3.1 Preparation

Coordinate System in Tracking The goal of tracking the camera motion is to obtain a sequence of viewing parameters of the real endoscope that is represented in the CT coordinate system. Each viewing parameter is represented as $\mathbf{Q} = (\mathbf{P}, \mathbf{w})$ here. $\mathbf{P} = (x, y, z)$ is the camera position of the endoscope, and $\mathbf{w} = (\alpha, \beta, \gamma)$ represents the camera orientation. The final goal of the tracking is to find a **Q** able to generate a virtual endoscopic view that is similar to the current frame of the real endoscopic video for each frame.

Rendering of VE Images The proposed method uses VE images for tracking. Accordingly, VE views are obtained by rendering the shape of the target organ and setting the viewpoint inside the organ with the perspective projection. In the tracking, a lot of views should be rendered. Since the surface rendering can be accelerated with conventional PC-based graphics boards, we employ the surface rendering method to render VE views.

In a real endoscope, the light power decreases with the square of the distance from the light source. The VES can simulate this effect. Although all organs have complex reflection properties, we ignore these properties except for diffuse reflection properties when rendering for simplification purpose. All lighting parameters are manually adjusted so that the VE views are close to the real views. Lens system distortion is also ignored at present.

3.2 Processing Overview

The inputs of the system are a real endoscopic video and an X-ray CT image of the same patient. The method consists of two processing steps: (a) direct camera motion estimation by solving an epipolar equation and (b) image-based registration between RE and VE images. In the former step, we calculate corresponding point-pairs from two time-adjacent images. The rotation and the translation motion of the real endoscopic camera are found by substituting the coordinates of these point-pairs into the epipolar equation. Then we perform the image based registration [4]. The observation parameter of the VES, which generates the most similar VE view to the current frame of the real endoscopic video, is obtained here by generating VE views while changing the observation parameter. The search area of this process is limited to within the area around the motion computed by the direct estimation process.

3.3 Processing Procedure

Calculation of Corresponding Point-Pairs We calculate pairs of two corresponding points on two time-adjacent real endoscopic images, $\mathbf{R}_{k-1}$ and $\mathbf{R}_k$. $\mathbf{R}_{k-1}$ and $\mathbf{R}_k$ are the previous frame and the current frame of the real endoscopic video, respectively. The sampled points are defined on $\mathbf{R}_{k-1}$ as shown in Fig. 1. We define a small subimage $\mathbf{S}_{k-1}^{\mathbf{p}_i}$ for each sample point i on $\mathbf{R}_{k-1}$. The center point $\mathbf{p}_i$ of $\mathbf{S}_{k-1}^{\mathbf{p}_i}$ is located at the same position as sample point i. The size of $\mathbf{S}_{k-1}^{\mathbf{p}_i}$ is given beforehand. We search for a subimage $\mathbf{S}_k^{\mathbf{p}'_i}$ that is similar to $\mathbf{S}_k^{\mathbf{p}_i}$. The center point of $\mathbf{S}_k^{\mathbf{p}'_i}$ is noted as $\mathbf{p}'_i$, which can be found as the position that maximizes the correlation between subimages $\mathbf{S}_{k-1}^{\mathbf{p}_i}$ and $\mathbf{S}_k^{\mathbf{p}'_i}$. We consider the two points $\mathbf{p}_i$ and $\mathbf{p}'_i$ as a pair of corresponding points.

Calculation of the Fundamental Matrix This process calculates a fundamental matrix $\mathbf{F}$ from the point-pairs obtained in the previous step. Let $(u_i v_i)^T$ and $(u'_i v'_i)^T$ be $(u_i v_i) = \mathbf{A}\mathbf{u}_i$ and $(u'_i v'_i) = \mathbf{A}\mathbf{u}'_i$. We use $(u_i\ v_i\ 1)^T$ and $(u'_i\ v'_i\ 1)^T$ as $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{m}'}$ in Eq. (4) for the corresponding point-pair i. When we describe $\mathbf{F} = \begin{pmatrix} f_{11}\ f_{12}\ f_{13} \\ f_{21}\ f_{22}\ f_{23} \\ f_{31}\ f_{32}\ f_{33} \end{pmatrix}$, Eq. (4) can be rewritten as

$$\begin{pmatrix} uu'\ vu'\ u'\ uv'\ vv'\ v'\ u\ v\ 1 \end{pmatrix} \mathbf{f} = 0, \tag{6}$$

where $\mathbf{f} = \begin{pmatrix} f_{11}\ f_{12}\ f_{13}\ f_{21}\ f_{22}\ f_{23}\ f_{31}\ f_{32}\ f_{33} \end{pmatrix}^T$. To find $\mathbf{F}$, eight point-pairs are required at least. For n point-pairs, the above equation can be expressed as

$$\mathbf{U}\mathbf{f} = \begin{pmatrix} u_1u'_1\ v_1u'_1\ u'_1\ u_1v'_1\ v_1v'_1\ v'_1\ u_1\ v_1\ 1 \\ \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ u_nu'_n\ v_nu'_n\ u'_n\ u_nv'_n\ v_nv'_n\ v'_n\ u_n\ v_n\ 1 \end{pmatrix} \mathbf{f} = \mathbf{0}. \tag{7}$$

The fundamental matrix $\mathbf{F}$ is calculated by solving Eq. (7).

Calculation of the Translation and the Rotation Motion This process computes the translation and the rotation motions from the obtained $\mathbf{F}$. We can obtain $\mathbf{T}$ and $\mathbf{R}$ by solving the following equations sequentially.

$$\mathbf{E}^T\mathbf{T} = \mathbf{0} \tag{8}$$

$$\mathbf{R}\mathbf{E}^T = [\mathbf{T}]_\times^T \tag{9}$$

Image Based Registration This process finds the observation parameter that generates the VE view that is the most similar to $\mathbf{R}_k$. Let $\mathbf{V}(\mathbf{Q})$ be the VE view rendered by using the observation parameter $\mathbf{Q}$. The search process is performed by the following maximization process,

$$\max_{\mathbf{Q}} E\left(\mathbf{R}_k, \mathbf{V}(\mathbf{Q})\right). \tag{10}$$

Table 1. Tracking results

		Number of successful frames	
Video Clip	Number of frames	Previous method	Proposed method
(a)	543	375	487
(b)	100	8	73
(c)	299	173	247
(d)	320	51	190

$E(\mathbf{A}, \mathbf{B})$ is a function for measuring the similarity between two images $\mathbf{A}$ and $\mathbf{B}$. We use the correlation of the two images as E. The operation (10) is executed by employing Powell's method. We set the rotational motion of the initial parameter of the search as $\mathbf{R}$ acquired by Eq. (9). The direction specified by $\mathbf{T}$ is used for eliminating the search area of the translation motion.

4 Experimental Results and Discussion

We implemented the proposed method on a conventional Microsoft Windows-based PC machine (CPU: Pentium III 1GHz, Main memory: 1Gbyte, Graphics board: nVidia GeForce2 GTS). A bronchoscopic video and a chest X-ray CT image of the same patient were used for evaluating the performance of the proposed method. All of the processes were executed as off-line jobs due to the processing speed. The endoscopic video was recorded onto a digital video tape in an examination room and the images were captured frame by frame. Each frame of the video was converted from a color image into a gray-scale image. The real endoscope was performed under a conventional protocol. The specifications of the CT image were: 512x512x180 image size, 5mm thickness, and 1mm reconstruction pitch. A set of triangle patches was generated to render VE views from the bronchus region extracted from the CT image by the method in [1].

We evaluated the performance of the proposed method by counting the number of frames tracked correctly. The authors (including engineering and medical doctors) evaluated the results by observing real and virtual endoscopic images, since it is hard to obtain the real 3-D position of a bronchoscope camera. Table 1 shows the results of the evaluation. This table also includes results obtained by a previous method [5]. About 500 frames were successfully tracked. We can see improvements in the tracking, especially in frames where some specific structures cannot be seen.

Figure 3 shows results of camera motion tracking. The left side figures of each column are sequences of real endoscopic images. The corresponding virtual bronchoscopic images are displayed on the right side. These virtual images were generated by rendering VE views at estimated viewpoints. The results show that the method could track the real endoscope camera motion satisfactorily. In the case of video clip (a) where tracheal tumor can be observed, both the previous method and the proposed method can be seen to have worked well.

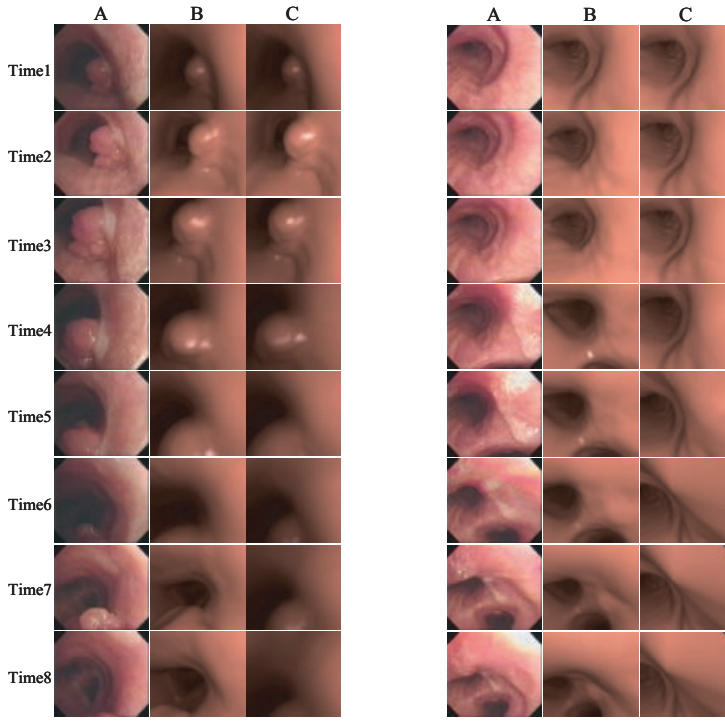

Fig.3 Results of camera motion tracking. The images in column A are real endoscopic images, those in B results obtained by the proposed method, and those in C were results obtained by a previous method [5].

From the scene where the tumor disappears from the views by due to endoscope movements, the proposed method can be seen to have continued its successful tracking, while the previous method failed. In video clip (b), the endoscope passes through an area where there are few feature structures. The proposed method can be seen to have performed the tracking correctly, while the previous method failed in tracking at Time3 and could no longer perform estimation, for the rest of the frames.

The processing time for one frame was also significantly reduced from fifteen second of the previous method to six seconds for the proposed one. This was because the result obtained by the direct estimation process eliminated the search area of the image-based registration process. Generally speaking, the image-based registration process is a time-consuming task, since we should generate a lot of virtual views to obtain an optimal parameter.

5 Conclusion

This paper described a method for tracking the camera motion of a real endoscope by using a camera motion recovery technique and an image-based registration technique. The translation and the rotational motion were directly estimated from corresponding point-pairs of two images by solving an epipolar equation. Then, precise estimation was performed by comparing VE and RE images. The experimental results showed that the method can track the camera motion in scenes where we cannot see feature structures in endoscopic views. Future work includes: (a) evaluation with a large number of cases, (b) development of a precise validation method, (c) improvement of the procedure for finding corresponding point-pairs by employing sub-pixel matching, and (d) reduction of the processing time.

Acknowledgement

The authors thank our colleagues for their useful suggestions and discussions. Parts of this research were supported by the Grant-In-Aid for Scientific Research from the Ministry of Education, the Grant-In-Aid for Scientific Research from Japan Society for Promotion of Science, and the Grant-In-Aid for Cancer Research from the Ministry of Health and Welfare of Japanese Government.

References

1. K. Mori, J. Hasegawa, J. Toriwaki, et al., Automated extraction and visualization of bronchus from 3-D CT images of Lung, In N.Ayache, (ed.) Computer Vision, Virtual reality and Robotics in Medicine, Lecture Notes in Computer Science, Vol. 905, pp.542-548, Springer-Verlag, Berlin Heidelberg New York, 1995
2. I. Bricault G. Ferretti, P. Cinquin, Registration Real and CT-Derived Virtual Bronchoscopic Images to Assist Transbronchial Biopsy, IEEE Trans. on Medical Imaging, 17, 5, pp.703-714, 1998
3. A.J. Sherbondy, A.P. Kiraly, A.L. Austin, et al., Virtual bronchoscopic approach for combining 3D CT and endoscopic video, Processing of SPIE Vol.3978, pp.104-115, 2000
4. A. Roche, G. Malandain, N. Ayache, S. Prima, Toward a Better Comprehension of Similarity Measures Used in Medical Image Registration, Lecture Notes Computer Science, 1679 (Proc. of MICCAI'99), pp.555-566, 1999
5. K.Mori, Y.Suenaga, J.Toriwaki, et al., Tracking of camera motion of real endoscope by using the Virtual Endoscope System, Proc. of CARS2000, pp.85-90, 2000
6. Gang Xu, Zhengyou Zhang, Epipolar Geometry in Stereo, Motion and Object Recognition, Kluwer Academic Publishers, Sept 1996