Combined Endoscopic Video Tracking and Virtual 3D CT Registration for Surgical Guidance

J. P. Helferty and W. E. Higgins

Abstract— **Bronchoscopic needle biopsy is a common step for early lung-cancer detection. This procedure uses two steps: (1) 3D computed-tomography (CT) chest image analysis, to choose a biopsy site; (2) live bronchoscopy, to perform the biopsy. CT-based virtual endoscopic analysis can improve results of biopsies, yet errors still can occur. We describe a procedure to combine the endoscopic video tracking (the "real" world) and CT-based virtual endoscopic registration (the "virtual" world). By bringing both sources of information together, a more robust surgical guidance system is realizable. Both the endoscope's video and thoracic CT scan are used as data sources in the tracking. An optical flow algorithm estimates the endoscope motion between successive video frames. The virtual CT rendering creates a range map for the optical flow equation. This simplifies the endoscope movement calculation into a straightforward linear system. We demonstrate this method for a phantom human airway-tree example.**

I. Introduction

Bronchoscopic needle biopsy is a common step in the early detection of lung cancer. Performing this procedure requires two steps: (1) initial 3D CT chest image analysis, to choose and plan a site; (2) follow-on live bronchoscopy to preform the planned biopsy [1]. This procedure is often unused since it requires the physician to deduce the best biopsy site by visually inspecting CT film slices. Virtual endoscopy can help this procedure by matching the endoscopic video to a virtual CT rendering that includes the target site of interest [1,2]. Unfortunately, this can lead to guidance errors, such as the physician being at a different site relative to the virtual world or being rotated incorrectly.

This paper presents a combined tracking/registration procedure for endoscopic guidance. An optical flow algorithm calculates the endoscope motion from analyzing a sequence of endoscope video frames [3]. This processing is then fused with corresponding virtual CT renderings to follow the 3D motion of the endoscope. The purpose of the system is to help prevent guidance and orientation errors during bronchoscopic procedures.

This tracking method is part of an analysis package that combines CT and video data sources to assist a physician during live brochoscopy [2]. The hidden tumors are drawn relative to the video scene, creating a target point for a needle biopsy. The tracking algorithm allows the virtual image to continuously follow the endoscope motion and functions as a component of the image-guided bronchoscopy system [4].

In other work, Bricault *et al.* introduced an algorithm to track the endoscope motion, but it is limited to bifurcation images [5]. Mori *et al.* presented a technique to track the motion of video frames [6,7]. Their system is demonstrated on a sequence of images, but runs at over 20 seconds per frame. Both methods are tested only on video-taped sequences and are not available during a live endoscopy procedure. Many investigators have studied the general problem of calculating a camera's 3D position and orientation change by analyzing 2D pixel movement [3,8–11]. These techniques involve solving for the 2D change in position of points in the image and translating that into a 3D motion vector [8,10]. This method is not well suited for endoscope video tracking since bronchoscopic images are smooth and corresponding points are difficult to calculate.

II. Tracking and Video/CT Registration

This section overviews the combined tracking and registration problem and describes the proposed method. The system is active during a live endoscopy. The physician has the endoscope inserted into the patient. The endoscope provides a live real-time video stream. Also, the patient's 3D CT chest scan is available.

The video tracking algorithm calculates the virtual CT viewpoints that correspond to each image in a sequence of endoscope frames. The frames start at time t_0 and end at time t_f with a time increment of δt. Let $E_V(x,y,t)$ be the endoscope video frame at time t. The collection of video frames are described as $\varepsilon_V(t_0,t_f) = \{E_V(x,y,t_0), E_V(x,y,t_0+\delta t), ..., E_V(x,y,t_f)\}$. $E_{\mathrm{CT}}^{\chi(t)}(x,y,t)$ is the virtual CT image that matches to the endoscope image $E_V(x,y,t)$. $\chi(t) = (X(t),Y(t),Z(t),\alpha(t),\beta(t),\gamma(t))$ is the virtual viewpoint in CT coordinates that corresponds to the endoscope location and orientation. $(X(t),Y(t),Z(t))$ is the virtual position and $(\alpha(t),\beta(t),\gamma(t))$ are the Euler angles rotated around x, y, and z respectively. For each endoscope frame, the goal is to find the virtual viewpoint $\chi(t)$ that causes $E_{\mathrm{CT}}^{\chi(t)}(x,y,t)$ to best match $E_V(x,y,t)$.

It is assumed that the initial viewpoint $\chi(t_0)$ is known. $\chi(t_0)$ corresponds to the endoscope frame $E_V(x,y,t_0)$. The problem is to calculate the remaining viewpoints $\chi(t_0+\delta t)$ to $\chi(t_f)$ that corresponds to the endoscope frames in the sequence $\varepsilon_V(t_0+\delta t, t_f)$.

In previous work, a registration system was designed that can calculate the virtual endoscope position relative to a video frame [1]. The CT/Video registration works well for a single frame but is too slow to be effective as a video tracking system on its own. An optical flow tracking algorithm can calculate the change in an endoscope's position between successive video frames. This measurement over time has a drift error since the viewpoint is

Penn State University, Dept. of Electrical Engineering, University Park, PA 16802,USA

This work was partially supported by NIH grants CA-74325 and CA-91534 from the National Cancer Institute, by the Olympus Corporation, and by the Whitaker Foundation.

the sum of incremental changes. The overall tracking system uses CT/Video registration to correct the drift errors caused by tracking with video frames alone.

An algorithm that combines registration and video-only tracking has benefits for both. The tracking system provides a better initial position for the registration algorithm. The registration system helps the tracking by giving a correct range map necessary for the next set of tracking calculations. The registration system also corrects for the accumulating errors in the tracking system. The combined tracking and registration algorithm starts with a sequence of video images and an initial registered viewpoint $\chi(t_0)$. The current registered viewpoint $\chi(t_f)$ results from applying the algorithm to all the images in the sequence. Figure 1 is a block diagram showing the steps in the combined registration and video-only algorithm. See below.

1. Get the initial viewpoint $\chi^*(t_0)$ that is close to the first video frame.
2. Get the first video frame from video source.
3. Render the viewpoint at $\chi{*}(t_0)$ to get a range map. The range map is used in the video frame tracking calculation.
4. Get N_{tr} video frames from the video source. N_{tr} is the number of video frames to track per a registration.
5. Start the tracking thread using N_{tr} video frames, and the range image at the first image in the N_{tr} frames. The thread applies (7) for each consecutive set of images.
6. In a seperate process, calculate a registration at $\chi^*(t_0{+}iN_{tr})$ The result is $\chi(t_0{+}(i{+}1)N_{tr})$. i counts the registration iterations and in this interval there were N_{tr} video tracking frames. This finds the best virtual viewpoint to match the first frame in the tracking iteration.
7. Wait for tracking thread to finish. The result is the change in endoscope position for N_{tr} frames. There are three translations changes and three rotation changes in each tracking result.
8. Increment the registered position $\chi(t_0 + (i+1)N_{tr})$ with the N_{tr} tracking movements. The result becomes $\chi^*(t_0 + (i+1)N_{tr})$. So at the end of an iteration, the best viewpoint comes from the registered viewpoint at the initial frame of the tracking set plus the endoscope differentials from video tracking.
9. Calculate the range image at $\chi * (t_0 + (i+1)N_{tr})$ for next tracking iteration. This is from rendering a image at this viewpoint and retrieving a range map from the OpenGL depth buffer.
10. Move the last video frame to first video frame for next tracking iteration. This maintains continuity in the video tracking calculations. Also, this sets up the next registration iteration.
11. If video source has frames remaining, go to Step 4. The process keeps repeating until the video source reaches the final frame.

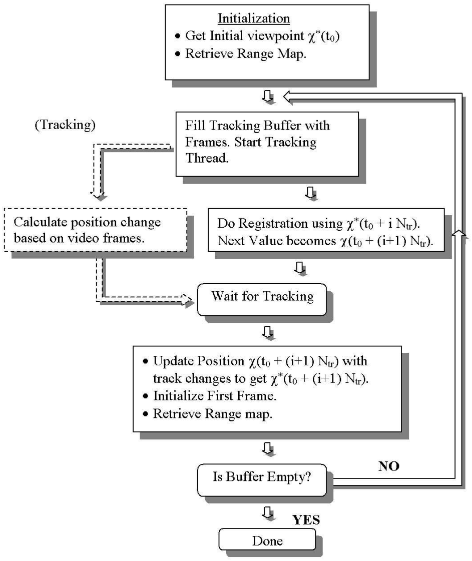

Initialization
• Get Initial viewpoint χ*(t₀)
• Retrieve Range Map.
(Tracking)
Fill Tracking Buffer with Frames. Start Tracking Thread.
Calculate position change based on video frames.
Do Registration using χ*(t₀ + i Nₜᵣ). Next Value becomes χ(t₀ + (i+1) Nₜᵣ).
Wait for Tracking
• Update Position χ(t₀ + (i+1) Nₜᵣ) with track changes to get χ*(t₀ + (i+1) Nₜᵣ).
• Initialize First Frame.
• Retrieve Range map.
Is Buffer Empty?
NO
YES
Done

Fig. 1. Block diagram of the combined registration and tracking algorithm. The procedure starts with a video/CT registration. It then performs tracking on a small number of frames while running the registration algorithm. The registered positions gets updated with the calculations from tracking. The range map registration is used to help tracking. Tracking gives an improved initial starting point for the registration algorithm.

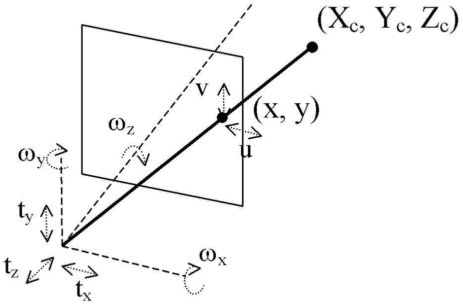


Fig. 2. A endoscope's movement in a rigid scene can be defined by three translational increments (t_x, t_y, t_z) and three Euler angle increments (w_x, w_y, w_z). These 3D movements result in the 2D pixel movement (u, v).

III. Endoscope Motion from Optical Flow

The mathematical problem of estimating 3D position change based on 2D pixel movement can be described as follows [3]. Figure 2 shows a rigid point being imaged by a moving endoscope. Let $\mathbf{X}_c = (X_c, Y_c, Z_c)$ be a point in endoscope coordinates, where x is to the right, y is up, and z is looking forward into the center of the screen. A 3D point in the endoscope view creates a pixel at the 2D endoscope screen point (x_i, y_i) from the perspective equations

$$x_i = \frac{fX_c}{Z_c} \qquad y_i = \frac{fY_c}{Z_c}. \tag{1}$$

with f as the focal length. Assume an endoscope can move in three small translations $\mathbf{t} = (t_x, t_y, t_z)$ or three small Euler rotations $\mathbf{w} = (w_x, w_y, w_z)$. The velocity of the 3D point relative to the endoscope center is

$$\dot{\mathbf{X}}_c = \mathbf{w} \times \mathbf{X}_c + \mathbf{t}. \tag{2}$$

Expanding the cross product gives the following equation:

$$\begin{bmatrix} \dot{X}_c \\ \dot{Y}_c \\ \dot{Z}_c \end{bmatrix} = \begin{bmatrix} -w_y Z_c + w_z Y_c - t_x \\ -w_z X_c + w_x Z_c - t_y \\ -w_x Y_c + w_y X_c - t_z \end{bmatrix}. \tag{3}$$

The point velocity $\dot{\mathbf{X}}_c$, relative to the endoscope, is a function of the point's position (X_c, Y_c, Z_c), the rotation (w_x, w_y, w_z), and the translation (t_x, t_y, t_z).

Let (u_i, v_i) be the 2D pixel velocity caused by the endoscope movement. This velocity is calculated from the time derivative of the pixel's position components, $u_i = \frac{dx_i}{dt}$ and $v_i = \frac{dy_i}{dt}$. Applying the derivatives to the motion equations, and substituting (3) into the projection equations (1) gives [3, 12],

$$u_i = \frac{-ft_x + x_c t_z}{Z_c} + \frac{x_c y_c}{f} w_x - \left(\frac{x_c^2}{f} + f\right) w_y + y_c w_z. \tag{4}$$

$$v_i = \frac{-ft_y + y_c t_z}{Z_c} + \left(\frac{y_c^2}{f} + f\right) w_x - \frac{x_c y_c}{f} w_y - x_c w_z. \tag{5}$$

(4) and (5) relates the endoscope translations (t_x, t_y, t_z) and rotations (w_x, w_y, w_z) to video pixel velocity (u_i, v_i). Separate equations of this form exist for every point in the image.

The optical flow method assumes that pixels in a scene can move, but the overall image brightness remains consistent. This leads to the classic optical flow constraint equation [3],

$$\frac{\partial E}{\partial x} u + \frac{\partial E}{\partial y} v + \frac{\partial E}{\partial t} = 0. \tag{6}$$

The partial $\frac{\partial E}{\partial t}$ is calculated with an image difference, and $\frac{\partial E}{\partial x}$ and $\frac{\partial E}{\partial y}$ are calculated from the image gradient. The motion parameters, which are three translation and three rotation variables, determine how the pixels move in the 2D scene. The pixel's movement corresponds to the variables, u and v. (6) will be used as the measurement portion of a least squares system to solve for the 3D motion parameters.

By substituting (4) and (5) into (6), it is possible to create a linear system to solve for the motion parameters. There is no need to calculate corresponding points. This leads to the optical flow equation in terms of the three-dimensional motion,

$$-E_t = \begin{bmatrix} -\frac{E_x f}{Z_c} \\ -\frac{E_y f}{Z_c} \\ \frac{E_x x_i}{Z_c} + \frac{E_y y_i}{Z_c} \\ E_x \frac{x_i y_i}{f} + E_y(\frac{y_i^2}{f} + f) \\ -E_x(\frac{x_i^2}{f} + f) - E_y \frac{x_i y_i}{f} \\ E_x y_i - E_y x_i \end{bmatrix}^T \begin{bmatrix} t_x \\ t_y \\ t_z \\ w_x \\ w_y \\ w_z \end{bmatrix}. \tag{7}$$

In (7), only the change motion parameters $(t_x, t_y, t_z, w_x, w_y, w_z)$ are unknown. E_t is the image difference, E_x and E_y are image gradients, Z_c is the image depth map, f is the perspective focal length, and x_i and y_i are the image pixel values.

There is a separate equation of the form (7) for every pixel in the image. The goal is to calculate the six unknown motion parameters $(t_x, t_y, t_z, w_x, w_y, w_z)$. If there are N pixels in the image, the six parameters result from solving a system of N linear equations. (7) can be used to calculate the change in endoscope motion by comparing two video frames. Summing the incremental changes over time will result in the motion change over a longer interval. This has a problem of drift error, since any error from the current increment is added to the accumulated total from previous increments. This compounding error increases with each increment. In order to limit this error, the system augments the video tracking system with position measures.

IV. Results

The endoscope sequence in Figure 3 shows an airway phantom example of the combined tracking and registration algorithm. There are a total of 412 frames in the sequence. Tracking starts at a known location as shown by the carina view in Frame 1. The endoscope sequence moves through the left mainstem bronchus towards the first subdivision. The tracking system calculates the virtual view to match the endoscope video. In this example, a registration took about 1.8 seconds so the average frame rate was 2.22 frames per second.

Figure 4 shows a screen display of the tracking system as a component of an image-guided bronchoscopy system. As the virtual position is tracked, the location is updated in all the separate displays. The physician can reference the endoscope position in a global surface view. The immediate feedback of the endoscope position should help prevent guidance and orientation errors.

References

[1] J. P. Helferty, A. J. Sherbondy, A. P. Kiraly, J. Z. Turlington, E. A. Hoffman, G. McLennan, and W. E. Higgins, "Image-guided endoscopy system for lung cancer assessment," *IEEE International Conference on Image Processing 2001*, vol. 2, pp. 307–310, Oct. 7-10 2001.

[2] J. P. Helferty and W. E. Higgins, "Technique for registering 3D virtual CT images to endoscopic video," *IEEE International Conference on Image Processing 2001*, vol. 2, pp. 893–896, Oct. 7-10 2001.

[3] B. K. P. Horn and R. J. Weldon, Jr., "Direct methods for recovering motion," *International Journal of Computer Vision*, vol. 2, pp. 51–76, 1988.

[4] J. P. Helferty, *Image-Guided Endoscopy and its Application to Pulmonary Medicine*, Ph.D. thesis, Penn State University, Dec 2001.

[5] I. Bricault, G. Ferretti, and P. Cinquin, "Registration of real and CT-derived virtual bronchoscopic images to assist transbronchial biopsy," *IEEE Transactions on Medical Imaging*, vol. 17, no. 5, pp. 703–714, Oct. 1998.

[6] H. Shoji, K. Mori, J. Sugiyama, Y. Suenaga, J. Toriwaki, H. Takabatake, and H. Natori, "Camera motion tracking of real endoscope by using virtual endoscopy system and texture information," *SPIE Medical Imaging 2001: Physiology and Function from Multidimensional Images, A. Clough and C.T. Chen, eds.*, vol. 4321, Feb 18-22 2001.

[7] K. Mori, Y. Suenaga, J. Toriwaki, J. Hasegawa, K. Katada, H. Takabatake, and H. Natori, "A method for tracking camera motion of real endoscope by using virtual endoscopy system," *SPIE Medical Imaging 2000: Physiology and Function from Multidimensional Images, A. Clough and C.T. Chen, eds.*, vol. 3978, pp. 122–133, Feb. 12-17, 2000.

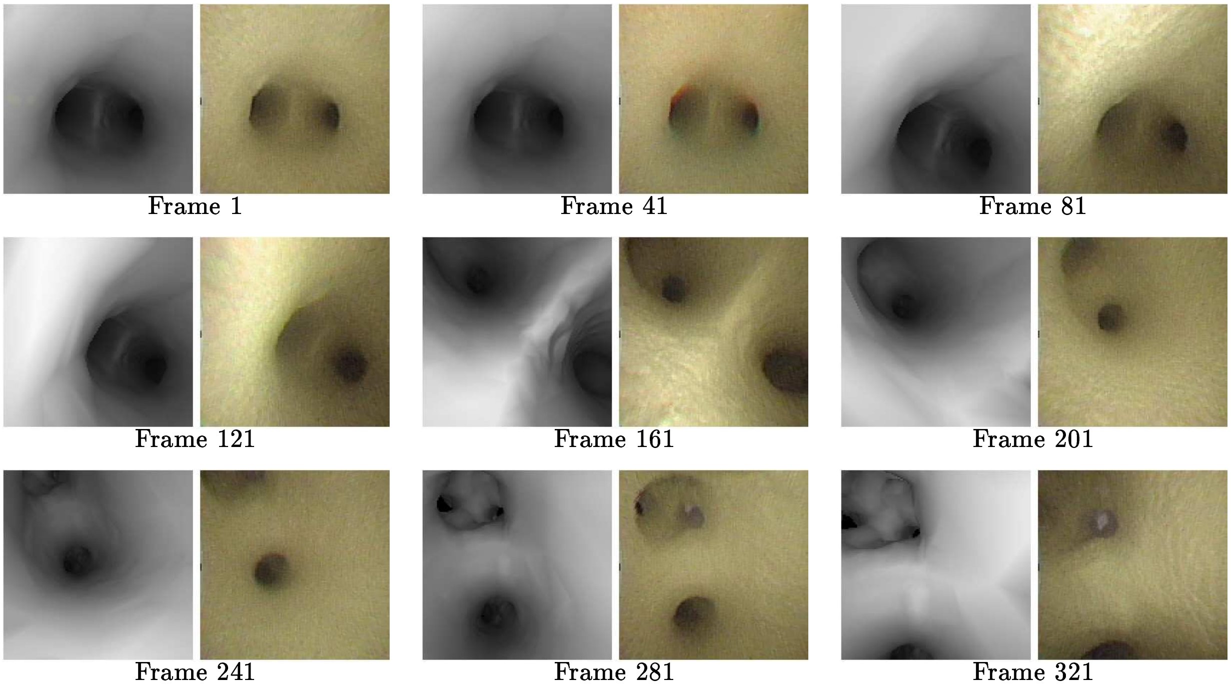
[8] T. Jebara, A. Azarbayejani, and A. Pentland, "3D structure from 2D motion," *IEEE Signal Processing Magazine*, pp. 66–84, May. 1999.

[9] C. Stiller and J. Konrad, "Estimating motion in image sequences," *IEEE Signal Processing Magazine*, pp. 70–90, Jul. 1999.

[10] K. I. Diamantaras, T. Papadimitriou, M. G. Strintzis, and M. Roumeliotis, "Total least squares 3-D motion estimation," *IEEE International Conference on Image Processing-98*, vol. 1, pp. 923–927, 1998.

[11] T. T. Kim and H. M. Kim, "Recursive total least squares algorithm for 3-D camera motion estimation from image sequences," *IEEE International Conference on Image Processing 98*, vol. 1, pp. 913–917, Oct. 1998.

[12] S. Soatto and P. Perona, "Reducing "structure from motion": A general framework for dyanamic vision part 1: Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 933–941, 1998.



Frame 1

Frame 41

Frame 81

Frame 121

Frame 161

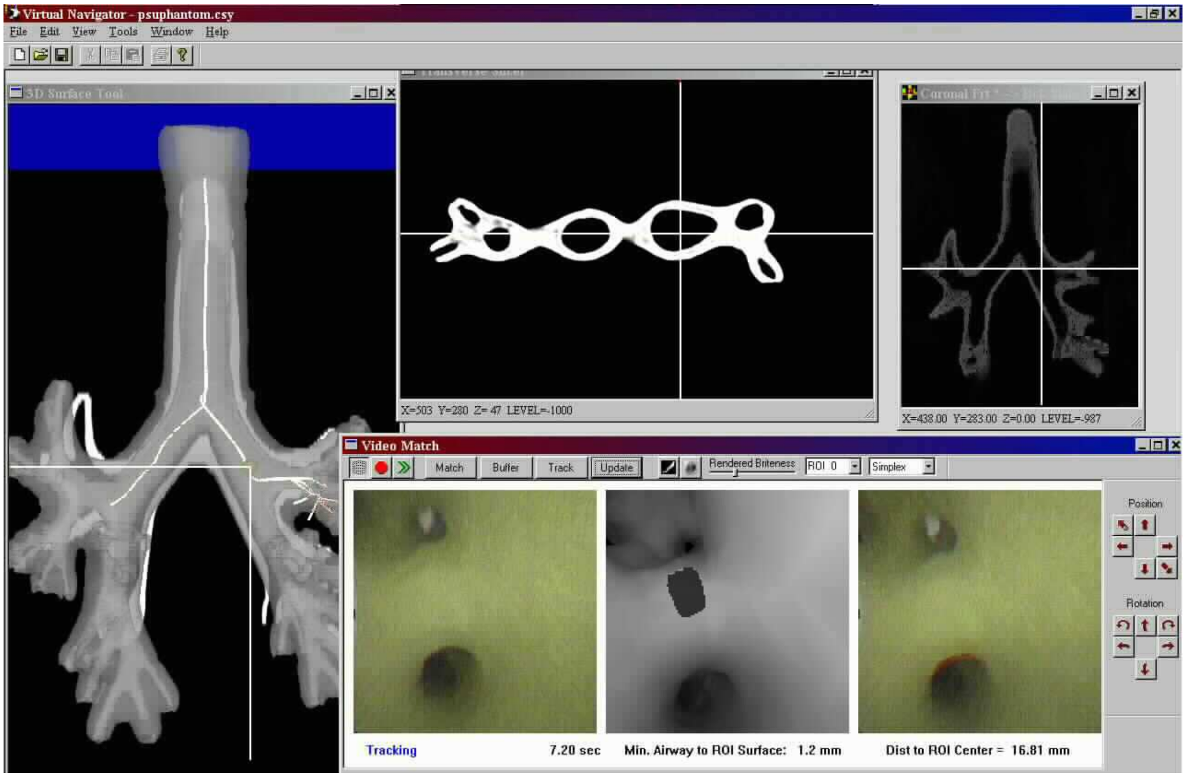
Frame 201

Frame 241

Frame 281

Frame 321

Fig. 3. Results of the combined registration and tracking algorithm for an endoscope video image. In each frame, the view on the left is the virtual CT rendering and on the right is the endoscope image. The endoscope starts at a well-known location like the carina. The physician moves the endoscope to the desired spot and the endoscope follows. This maintains a consistent tie between the endoscope position and the video world. Nine frames are selected from a sequence of endoscopic images with each separated by 40 frames. Overall, the total number of frames tracked was 412. The average frame rate in this example was 2.22 frames per second. This was a Dell Precision 620 workstation with dual Pentium III Xeon processors at 933Mhz. The graphics card was a Quadro II pro with 64 MB.



X=503 Y=280 Z= 47 LEVEL=-1000

X=438.00 Y=283.00 Z=0.00 LEVEL=-987

Tracking 7.20 sec Min. Airway to ROI Surface: 1.2 mm Dist to ROI Center = 16.81 mm

Fig. 4. A composite view of Virtual Navigator during bronchoscopy. The *Video Match* tool in the lower right contains the combined tracking and registration algorithm that maintains a tie between the virtual world and the video endoscope. The *3D Surface* tool shows a global view of the airway tree with the endoscope location and centerline paths. There is a transverse slice view in the upper middle. The upper right shows a coronal slab view. In the *Video Match* tool, the image on the left is the live video, the view in the center is the matching virtual CT image, and the image on the right is the current endoscope frame being processed. The calculated endoscope postion in shown in the *3D Surface* tool by the crosshair location. The coronal and transverse displays reflect the endoscope position by their own crosshairs.