An Application Driven Comparison of Several Feature Extraction Algorithms in Bronchoscope Tracking During Navigated Bronchoscopy

Xióngbiāo Luó[1], Marco Feuerstein[1,2], Tobias Reichl[2], Takayuki Kitasaka[3], and Kensaku Mori[4,1]

[1]Graduate School of Information Science, Nagoya University, Japan
[2]Computer Aided Medical Procedures, Technische Universität München, Germany
[3]Faculty of Information Science, Aichi Institute of Technology, Japan
[4]Information and Communications Headquarters, Nagoya University, Japan

Abstract. This paper compares Kanade-Lucas-Tomasi (KLT), speeded up robust feature (SURF), and scale invariant feature transformation (SIFT) features applied to bronchoscope tracking. In our study, we first use KLT, SURF, or SIFT features and epipolar constraints to obtaininter-frame translation (up to scale) and orientation displacements and Kalman filtering to recover an estimate for the magnitude of the motion (scale factor determination), and then multiply inter-frame motion parameters onto the previous pose of the bronchoscope camera to achieve the predicted pose, which is used to initialize intensity-based image registration to refine the current pose of the bronchoscope camera. We evaluate the KLT-, SURF-, and SIFT-based bronchoscope camera motion tracking methods on patient datasets. According to experimental results, we may conclude that SIFT features are more robust than KLT and SURF features at predicting the bronchoscope motion, and all methods for predicting the bronchoscope camera motion show a significant performance boost compared to sole intensity-based image registration without an additional position sensor.

Keywords: Bronchoscope Tracking, Camera Motion Estimation, KLT, SURF, SIFT, Image Registration, Navigated Bronchoscopy

1 Introduction

In minimally invasive diagnosis and surgery of lung and bronchus cancer, a physician usually performs transbronchial needle aspiration (TBNA) to obtain tissue samples to assess suspicious tumors as well as to treat or remove precancerous tissue. However, it is difficult to properly navigate the biopsy needle to the region of interest (ROI) for sampling tissue inside the airway tree, because the TBNA procedure is usually guided by conventional bronchoscopy, which only provides 2-D information (bronchoscopic video images), and because of the complexity of the structure of the bronchial tree. Recently, bronchoscopic navigation systems have been developed to guide the TBNA procedure by fusing pre-operative and

intra-operative information such as 3-D multi-detector computed-tomography (CT) image data and real-time bronchoscopic video. This helps a physician to properly localize the biopsy needle during navigated bronchoscopy.

For navigated bronchoscopy the exact pose of the bronchoscope camera must be tracked inside the airway tree. Unfortunately it is really challenging to accurately track the position and orientation of the bronchoscope camera inside the patient's airway tree in real time during bronchoscopic navigation. So far, two main approaches (or their combination) for bronchoscope tracking have been proposed in the literature: (a) sensor-based and (b) vision-based tracking. The former uses an electromagnetic (EM) tracking system (e.g., the superDimension navigation system [12]) to locate an electromagnetic sensor that is usually fastened at the bronchoscope tip to directly measure the bronchoscope camera position and orientation. The latter analyzes the bronchoscopic video images obtained from the bronchoscope camera to continuously track the bronchoscope tip on the basis of image registration methods [10, 6]. This is a widely discussed topic in the field of bronchoscope tracking and also the topic of our paper.

Usually, vision-based methods use image registration techniques to align a real bronchoscope camera pose to a virtual camera pose generated by placing a virtual camera inside the 3-D CT data. However, a major drawback is that image registration techniques heavily depend on characteristic information of bronchial trees (e.g., bifurcations or folds), so they can fail easily to track the bronchoscope camera in the case of the shortage of such information [5]. Feature-based bronchoscope motion estimation is a promising means for dealing with this problem during bronchoscope tracking [10, 4]. Without any characteristic information, other texture feature information of real bronchoscopic video frames can be extracted and used to compensate the performance of image registration.

Basically, a feature-based approach for motion estimation and recovery first needs to extract features from camera images, which can be utilized to compute the relative camera motion, for example by epipolar geometry (up to scale). Currently, two well-known methods for extracting features are the SURF and SIFT algorithm [2, 8]. Both return distinctive features from keypoints that are invariant to image scale and rotation. Also, the KLT tracker first detects good features by calculating the minimum eigenvalue of each 2×2 gradient matrix and selects features to be tracked using an optimization (e.g. Newton-Raphson) method for minimizing the difference between two feature windows from two consecutive images [11].

However, little work can be found that evaluates the effectiveness of these different feature extraction algorithms that are used for bronchoscope tracking during bronchoscopic navigation. This study utilizes these feature-based camera motion tracking methods to improve the performance of image registration-based bronchoscope tracking. We use the KTL, SURF, and SIFT features to estimate inter-frame pose displacements (up to scale) on the basis of epipolar constraints and Kalman filtering to get position estimates before performing image registration. We compare and evaluate the respective performances of KLT, SURF, and SIFT features used for bronchoscope tracking.

2 Method

Feature-based camera motion estimation algorithms are widely used in the field of structure from motion (SFM) or stereo vision. These approaches basically consist of two main steps: (1) feature extraction and (2) feature tracking. The first step usually characterizes some points or regions in each video image as interest features that carry motion information among video images. Sequentially, inter-frame motion parameters (up to scale) can be estimated in the second step by recognizing corresponding features between consecutive video frames. In our work, we detect interest features for each real bronchoscopic (RB) video image using a KLT-, SURF-, or SIFT-based method, respectively. We address the difficulty of determining the magnitude of motion (here referred to as scale factor) by Kalman filtering during feature-based motion estimation.

Our proposed bronchoscope tracking method has two major stages: rough camera motion estimation and intensity-based image registration. Figure 1 displays a flow-process diagram of our tracking method. First, KLT, SURF, or SIFT features are respectively detected from the current bronchoscopic video image and feature correspondences are identified in the previous frame. During epipolar geometry analysis, inter-frame camera motion up to scale is predicted on the basis of these feature correspondences. Kalman filtering is then applied to estimate the uncertain scale factor, or in other words, the magnitude of the relative motion. Finally, after combining the estimates of epipolar geometry analysis and Kalman filtering to a full Euclidean transformation matrix that moves the camera from the previous to the current pose, we can perform image registration initialized with this matrix.

Specifically, the feature-based bronchoscope camera motion tracking process is performed by the following five steps:

[Step 1] Feature detection. We extract 2-D feature points by using the KLT, SURF, or SIFT algorithm [11, 2, 8]. The KLT tracker is sometimes referred to as corner detector while the other two approaches, which are considered as scale invariant feature detectors, try to find characteristic blob-like structures in an image independent of its actual size. The SURF or SIFT detector can be constructed using a scale space representation of an original image at different resolutions. After detecting feature points, SIFT usually describes each feature point using a 128-dimensional vector while SURF does so with a 64-dimensional vector. All these vectors include the local gradient direction and magnitude information in a certain square neighborhood centered at the feature point. More details about these feature detection algorithms can be found in the original publications [11, 2, 8]. We note that the normal SURF algorithm is implemented by doubling the initial image resolution in our case, and hence we can obtain good performance, as shown in the work of Bauer et al. [1].

[Step 2] Feature correspondences. After feature point detection from bronchoscopic video sequences, we must determine feature correspondences that can be used to find the relative motion relation between two successive RB images.

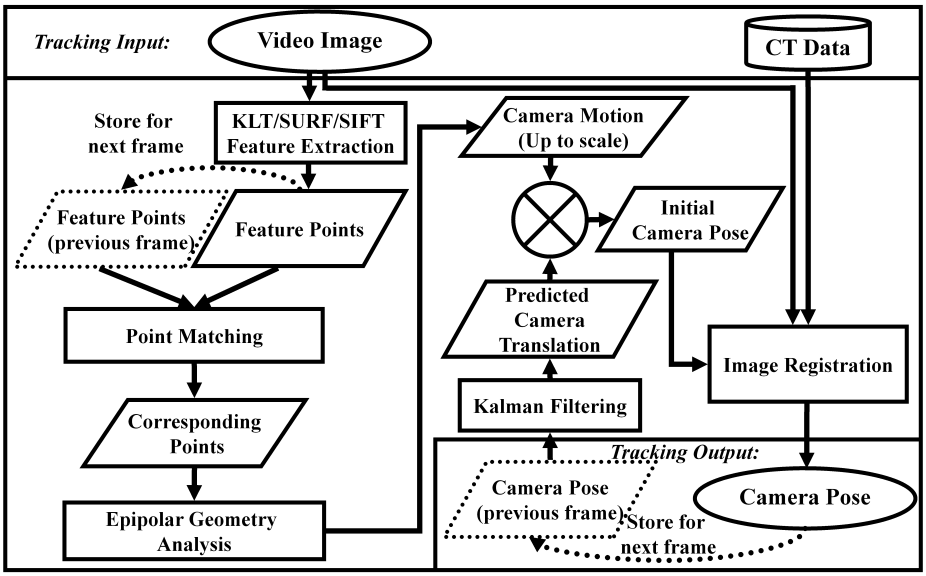


Fig. 1: Processing flowchart of our motion tracking method.

The KLT method extracts adequate feature points of an RB image and uses normalized cross correlation (NCC) to track (or match) them. However, for SURF or SIFT feature points, we recognize corresponding 2-D point pairs using the third matching strategy from the work of Mikolajczyk and Schmid [9]. Additionally, a simple outlier detection mechanism was performed on the basis of the standard deviation of the distances between corresponding points to remove unsuitable point pairs.

[Step 3] Epipolar geometry analysis. Inter-frame motion parameters $\Delta\tilde{\mathbf{Q}}^{(i)}$ between the (i-1)-th and (i)-th RB image contain a translation unit vector $\Delta\tilde{\boldsymbol{t}}^{(i)}$ and rotation matrix $\Delta\tilde{\mathbf{R}}^{(i)}$ that can be predicted with epipolar geometry analysis by solving the following equations sequentially:

$$\mathbf{E}^T \Delta\tilde{\boldsymbol{t}}^{(i)} = \mathbf{0}, \tag{1}$$

$$\Delta\tilde{\mathbf{R}}^{(i)}\mathbf{E}^T = \left[\Delta\tilde{\boldsymbol{t}}^{(i)}\right]_{\times}^{T} \tag{2}$$

where $\mathbf{E}$ is the essential matrix described epipolar constraints [7] that our corresponding points must satisfy. It needs to be clarified that the essential matrix $\mathbf{E}$ involves an arbitrary scale factor. Hence the absolute translation vector $\Delta\hat{\boldsymbol{t}}^{(i)}$ depends on an arbitrary scale factor $\tilde{\alpha}^{(i)} = |\Delta\tilde{\boldsymbol{t}}^{(i)}|$ that depicts the real magnitude of the translational motion. An effective method to predict this scale that is based on Kalman filtering is proposed in the next step.

[Step 4] Kalman filtering-based scale factor estimation. Kalman filtering is widely developed in the community for target position tracking on the basis of a state-space model [3]. In our work, Kalman-based motion filtering is employed to determine the magnitude of the bronchoscope translational motion. Basically, the scale factor $\hat{\alpha}^{(i)}$ can be determined by

$$\hat{\alpha}^{(i)} = |\Delta\hat{\boldsymbol{t}}^{(i)}| \;\; = \;\; |\hat{\boldsymbol{t}}^{(i)} - \hat{\boldsymbol{t}}^{(i-1)}|, \tag{3}$$

where the camera absolute translation vector $\hat{\boldsymbol{t}}^{(i-1)}$ and $\hat{\boldsymbol{t}}^{(i)}$ are calculated by Kalman filtering.

We can now retrieve the absolute translation vector $\Delta\tilde{\boldsymbol{t}}_*^{(i)}$ between frames $(i-1)$ and i from the unit transaltion vector $\Delta\tilde{\boldsymbol{t}}^{(i)}$ (determined in the rough camera motion estimation stage) with respect to $\hat{\alpha}^{(i)}$

$$\Delta\tilde{\boldsymbol{t}}_*^{(i)} = \hat{\alpha}^{(i)} \frac{\Delta\tilde{\boldsymbol{t}}^{(i)}}{|\Delta\tilde{\boldsymbol{t}}^{(i)}|}. \tag{4}$$

Next, the estimated motion $\Delta\tilde{\mathbf{Q}}_*^{(i)}$ of the bronchoscope camera between frames $(i-1)$ and i can be computed by

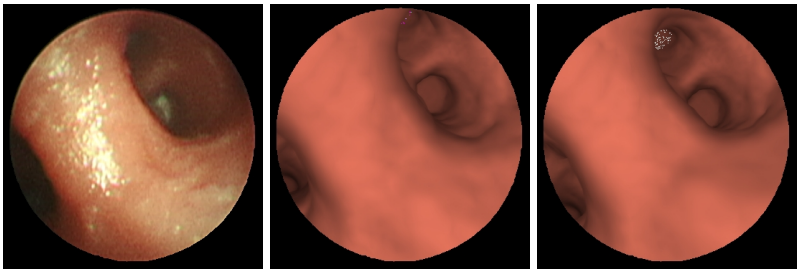
$$\Delta\tilde{\mathbf{Q}}_*^{(i)} = \begin{pmatrix} \Delta\tilde{\mathbf{R}}^{(i)} & \Delta\tilde{\boldsymbol{t}}_*^{(i)} \\ \mathbf{0}^T & 1 \end{pmatrix}, \tag{5}$$

where $\Delta\tilde{\mathbf{R}}^{(i)}$ is calculated by Eq. 2. Finally, the estimate $\Delta\tilde{\mathbf{Q}}_*^{(i)}$ is utilized as initialization of image registration, as described in the next step.

[Step 5] Intensity-based image registration. Intensity-based registration commonly defines a similarity measure and maximizes the similarities or minimizes the dissimilarities between an RB image $\mathbf{I}_R^{(i)}$ and a virtual bronchoscopic (VB) image $\mathbf{I}_V$. We here use a modified mean squared error (*MoMSE*) [5] similarity measure. Let $\mathbf{I}_V(\mathbf{Q}^{(i)})$ be a VB image generated from the predicted pose $\mathbf{Q}^{(i)} = \mathbf{Q}^{(i-1)}\Delta\mathbf{Q}^{(i)}$ of the current frame using volume rendering techniques, where $\mathbf{Q}^{(i-1)}$ denotes the previous camera pose and $\Delta\mathbf{Q}^{(i)}$ the inter-frame motion information between successive frames. By updating $\Delta\mathbf{Q}^{(i)}$, a series of VB images $\mathbf{I}_V(\mathbf{Q}^{(i-1)}\Delta\mathbf{Q}^{(i)})$ is generated and the most similar one corresponding to the RB image $\mathbf{I}_R^{(i)}$ is searched for. In summary, the intensity-based registration process optimizing $\Delta\mathbf{Q}^{(i)}$ can be formulated as

$$\Delta\mathbf{Q}^{(i)} = \arg\min_{\Delta\mathbf{Q}} MoMSE(\mathbf{I}_R^{(i)}, \mathbf{I}_V(\mathbf{Q}^{(i-1)}\Delta\mathbf{Q})). \tag{6}$$

For this optimization, the initialization of $\Delta\mathbf{Q}$ in Eq. 6 is one of the key components affecting tracking robustness and accuracy. $\Delta\mathbf{Q}$ is initialized as an identity matrix in previous work [5]. However, in our new method, we use our estimate $\Delta\tilde{\mathbf{Q}}_*^{(i)}$ (see Eq. 5) instead. Since we got this estimate by matching stable image features, it can overcome certain limitations of sole image registration such as dependencies on airway folds or bifurcations and hence enhances the tracking performance.



(a) (b) (c)

Fig. 2: Example of the tracking results from the two stages. (a) shows the real pose of the bronchoscope camera. (b) displays the predicted pose from rough camera motion estimation by using feature-based tracking. (c) shows the refined pose by performing image registration initialized by (b).

3 Experimental Results and Discussion

We evaluated sole intensity-based registration (M1) and our proposed tracking methods (M2: KLT-based method, M3: SURF-based method, M4: SIFT-based method) on patient datasets, each consisting of an RB video sequence and a preinterventional 3-D chest CT. In-vivo patient data was acquired in accordance with a standard clinical protocol. The acquisition parameters of the CT images are 512×512 pixels, 72-209 slices, 2.0-5.0 mm slice thickness, and 1.0-2.0 mm reconstruction pitch. The image sizes of the bronchoscopic video frames are 362×370 and 256×263 pixels. We have done all implementations on a Microsoft Visual C++ platform and ran it on a conventional PC (CPU: Intel XEON 3.80 GHz×2 processors, 4-GByte memory).

A criterion for determining whether a method is more robust than another can be described by visual inspection and sum of the number of successfully tracked frames. If a VB image generated from the estimated camera parameters is greatly similar to the corresponding RB image, we consider it successfully tracked.

Table 1 gives quantitative results on the performance of all methods. Compared to M1, M2, and M3, in most cases the tracking performance has been improved significantly by using the proposed tracking algorithm M4. Figure 4 shows examples of RB images and the corresponding virtual images generated by volume rendering using the camera pose, calculated by the respective methods. The virtual images generated from the estimates of M4 are more similar than those of M1, M2, and M3, which means M4 more accurately predicts the real pose.

For the KLT method, we detect corner features and select 430 good features to be tracked from the previous frame [11]. The KLT tracker can usually track around 200 points per frame in our case. Because of the quality of KLT features, M2 has worse tracking results than M3 and M4, but is still better than M1. The

Table 1: Comparison of the tracking results for our patient studies, in terms of the number and percentage of successfully tracked frames and average processing time (seconds) per frame.

Cases ID	Num. of Frames	Number (Percentage) of frames successfully tracked			
		M1	M2 (KLT)	M3 (SIFT)	M4 (SURF)
Case 1	1200	450 (37.5%)	560 (46.7%)	683 (56.9%)	1120 (93.3%)
Case 2	200	116 (58.0%)	120 (60.0%)	70 (35.0%)	130 (65.0%)
Case 3	800	433 (54.1%)	618 (77.2%)	694 (86.8%)	774 (96.7%)
Case 4	800	437 (54.6%)	340 (42.5%)	605 (75.6%)	780 (97.5%)
Case 5	1000	431 (43.1%)	506 (50.6%)	575 (57.5%)	557 (55.7%)
Case 6	279	279 (100%)	279 (100%)	279 (100%)	279 (100%)
Case 7	400	240 (60.0%)	190 (32.5%)	210 (52.5%)	260 (65.0%)
Case 8	450	246 (54.7%)	217 (48.2%)	10 (2.22%)	10 (2.22%)
Total	**5120**	**2632 (51.4%)**	**2830 (55.3%)**	**3126 (61.1%)**	**3910 (76.4%)**
Average Times		**0.92 s**	**0.96 s**	**0.57 s**	**1.83 s**

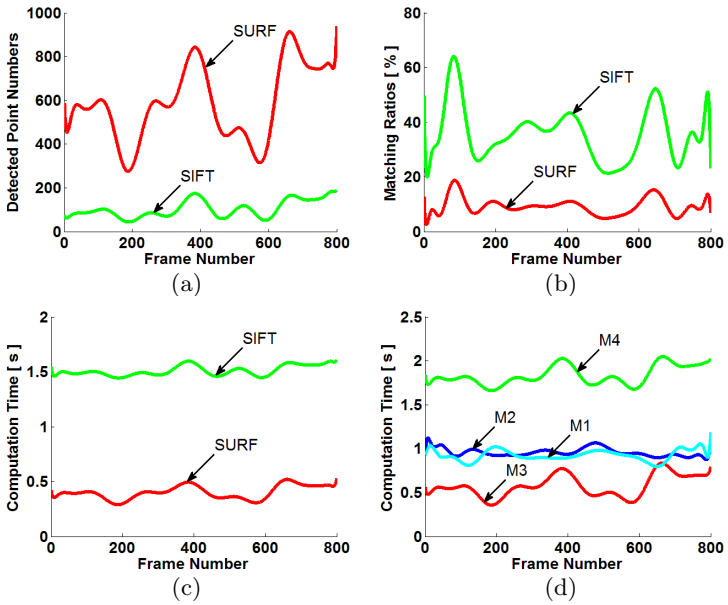


Fig. 3: Examples of detected feature numbers and computation times of Case 4. (a) shows the number of detected feature points, (b) gives the matching ratios calculated between the numbers of matching and detected points when using SURF and SIFT for each frame. (c) displays the time required for detecting SURF and SIFT features, and (d) illustrates the time needed to track the bronchoscope pose of each frame when using M1, M2, M3, and M4. It clearly shows that for M4 the average processing time with at least 1.5 seconds per frame is three times higher than that for M3, because it includes SIFT feature detection.

tracking results of M3 are worse than those of M4, although SURF detected many more features and correspondences (around 587 detected points and 56 matching points per frame, matching ratios: 9.2%, as shown in Figure 3 (a) and (b)) than that of SIFT (around 103 detected points and 38 matching points per frame, matching ratios: 35.7%, as shown in Figure 3(a) and (b)). We believe that the SURF features-based method gives worse estimates to initialize the registration step than the SIFT features-based method, and hence fails to track the bronchoscope motion more often. This also demonstrates that the feature point quality from SURF is not as good as that of SIFT, as the authors already concluded in the work of Bay et al. [2]. Additionally, we note that all other approaches (M2-M4) show better tracking results than sole intensity-based image registration (M1). This can be explained by the usage of image texture features that depend less on airway folds or bifurcations.

Regarding computational efficiency, according to Table 1, M1 requires approximately 0.92 seconds to process a frame and the run-time of M2 is about 0.96 seconds per frame while that of M3 comes to 0.57 seconds per frame and M4 computes each frame in around 1.83 seconds. Compared to M1, M3 can improve the computational efficiency while M4 increases the processing time for each frame. From the work of Bay et al. [2] we know that SURF is faster than SIFT at detecting features, since the SURF method uses a fast-Hessian detector on the basis of an integral image. However, all methods cannot track the bronchoscope motion in real time (real time means 30 frames per second need to be processed in our case). This is because feature-based motion recovery methods are time-consuming in terms of detecting points and finding their correspondences, and so is the registration stage of bronchoscope tracking. However, we can utilize the GPU (graphics processing unit) to accelerate our implementations and make it (almost) real time.

Finally, in our patient study all methods failed to track the motion of the bronchoscope in some cases. This is because the estimation failed in the intensity-based registration process, which is usually caused by problematic bronchoscopic video frames such as RB images, on which bubbles appeared. Additionally, tracking failure also resulted from airways deformation, which was caused by patient movement, breathing, and coughing, and is also one particular challenge in navigated bronchoscopy. Currently, we do not explicitly address the problem of respiratory motion in our tracking method. Therefore, our future work will focus on improving intensity-based image registration for bronchoscope tracking during bronchoscopic navigation, as well as constructing a breathing motion model to compensate for respiratory motion.

4 Conclusion

This paper compared KLT, SURF, and SIFT features applied to bronchoscope tracking. We utilized the KLT-, SURF-, and SIFT-feature-based camera motion tracking method to improve the performance of image registration-based bronchoscope tracking without an additional position sensor. Furthermore, from

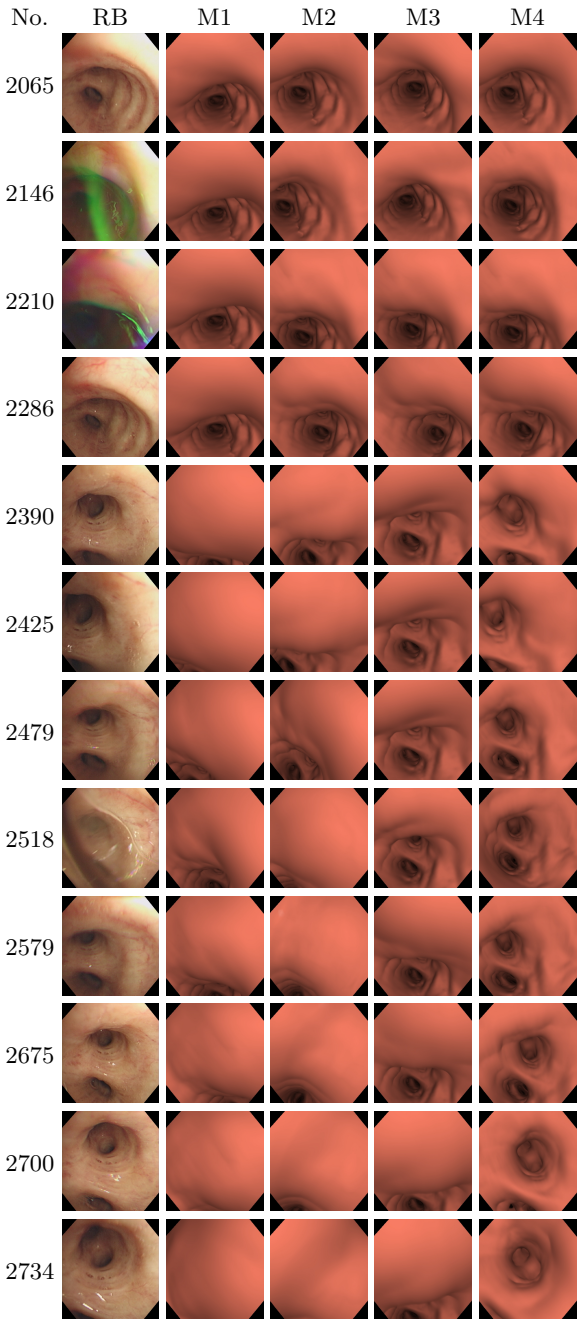


Fig. 4: Results of camera motion tracking for the patient assessment. The second column shows selected frames from a sequence of patient RB images and the first column their corresponding frame numbers. The other columns show tracking results for methods M1∼M4, all generated by volume rendering of the airways from the estimated viewpoints.

experimental results, we may conclude that SIFT features are more robust than the other two features when applied to predict bronchoscope motion, since the SIFT-based method successfully tracked 76.4% frames, compared to the KLT-based and the SURF-based methods with 55.3% and 61.1%, respectively. However, with about half to a third the processing time of the other methods, the SURF-based method seems to be a good compromise between tracking accuracy and computational efficiency.

References

1. Bauer, J., Süunderhauf, N., Protzel, P.: Comparing several implementations of two recently published feature detectors. In: Proc. of the International Conference on Intelligent and Autonomous Systems (2007)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), 346–359 (2008)
3. Cuevas, E., Zaldivar, D., Rojas, R.: Kalman filter for vision tracking. Tech. Rep. B 05-12, Freie University Berlin (October 2005)
4. Darius Burschka, Ming Li, M.I.R.H.T., Hager, G.D.: Scale-invariant registration of monocular endoscopic images to CT-scans for sinus surgery. Medical Image Analysis 9(5), 413–426 (2005)
5. Deguchi, D., Mori, K., Feuerstein, M., Kitasaka, T., Maurer Jr., C.R., Suenaga, Y., Takabatake, H., Mori, M., Natori, H.: Selective image similarity measure for bronchoscope tracking based on image registration. Medical Image Analysis 13(4), 621–633 (2009)
6. Deligianni, F., Chung, A.J., Yang, G.Z.: Nonrigid 2-D/3-D registration for patient specific bronchoscopy simulation with statistical shape modeling: Phantom validation. IEEE Transactions on Medical Imaging 25(11), 1462–1471 (2006)
7. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press (2004)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
10. Mori, K., Deguchi, D., Sugiyama, J., Suenaga, Y., Toriwaki, J., Maurer Jr., C.R., Takabatake, H., Natori, H.: Tracking of a bronchoscope using epipolar geometry analysis and intensity based image registration of real and virtual endoscopic images. Medical Image Analysis 6, 321–336 (2002)
11. Shi, J., Tomasi, C.: Good features to track. In: CVPR. pp. 593–600 (1994)
12. Solomon, S.B., P. White, J., Wiener, C.M., Orens, J.B., Wang, K.P.: Three-dimensionsal CT-guided bronchoscopy with a real-time electromagnetic position sensor: a comparison of two image registration methods. Chest 118(6), 1783–1787 (2000)