Int J CARS (2013) 8:575–592
DOI 10.1007/s11548-013-0814-x

ORIGINAL ARTICLE

A robust method to track colonoscopy videos with non-informative images

Jianfei Liu · Kalpathi R. Subramanian · Terry S. Yoo

Received: 15 August 2012 / Accepted: 11 January 2013 / Published online: 3 February 2013

Abstract
Purpose Continuously, optical and virtual image alignment can significantly supplement the clinical value of colonoscopy. However, the co-alignment process is frequently interrupted by non-informative images. A video tracking framework to continuously track optical colonoscopy images was developed and tested.
Methods A video tracking framework with immunity to non-informative images was developed with three essential components: temporal volume flow, region flow, and incremental egomotion estimation. Temporal volume flow selects two similar images interrupted by non-informative images; region flow measures large visual motion between selected images; and incremental egomotion processing estimates significant camera motion by decomposing each large visual motion vector into a sequence of small optical flow vectors. The framework was extensively evaluated via phantom and colonoscopy image sequences. We constructed two colon-like phantoms, a straight phantom and a curved phantom, to measure actual colonoscopy motion.
Results In the straight phantom, after 48 frames were excluded, the tracking error was <3 mm of 16 mm traveled. In the curved phantom, the error was <4 mm of 23.88 mm traveled after 72 frames were excluded. Through evaluations with clinical sequences, the robustness of the tracking framework was demonstrated on 30 colonoscopy image sequences from 22 different patients. Four specific sequences among these were chosen to illustrate the algorithm's decreased sensitivity to (1) fluid immersion, (2) wall contact, (3) surgery-induced colon deformation, and (4) multiple non-informative image sequences.
Conclusion A robust tracking framework for real-time colonoscopy was developed that facilitates continuous alignment of optical and virtual images, immune to non-informative images that enter the video stream. The system was validated in phantom testing and achieved success with clinical image sequences.

Keywords Colonoscopy · Tracking · Region flow · Temporal volume flow · Egomotion

Electronic supplementary material The online version of this article (doi:10.1007/s11548-013-0814-x) contains supplementary material, which is available to authorized users.

J. Liu (✉)
Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, MD 20892, USA
e-mail: jianfei.liu@nih.gov

K. R. Subramanian
Department of Computer Science, The University of North Carolina at Charlotte, Charlotte, NC 28223, USA

T. S. Yoo
Office of High Performance Computing and Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Introduction

The mortality of colorectal cancer is estimated to be about 51,690 in the United States in 2012 [1]. *Optical colonoscopy* (OC) is a primary screening procedure to detect and remove cancerous polyps (tumors), despite the fact that OC procedures can miss polyps [2]. Summers [3] and Duncan [4] showed that the ability to correlate polyps in virtual colonoscopy (VC) with optical colonoscopy is clinically important for screening colorectal cancer. The co-alignment of OC and VC images [5] thus has the potential to improve OC procedures, as anus-to-cecum distance measurements indicate that the location difference of the majority of polyps is within 10 cm.

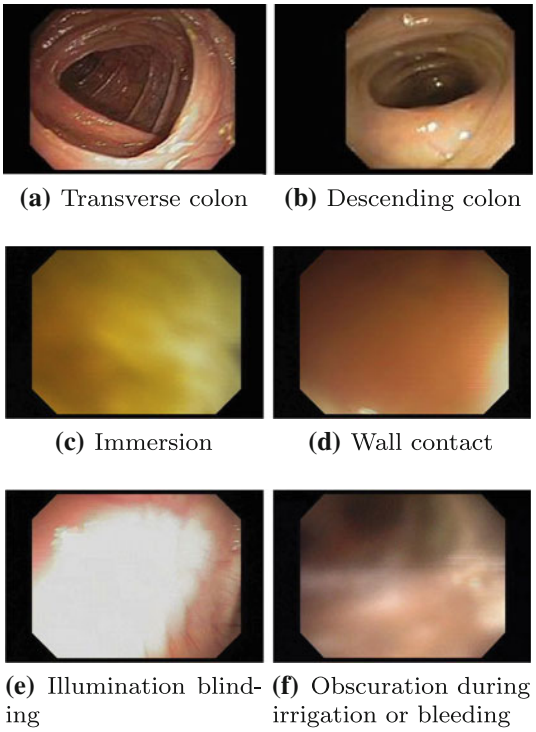

Fig. 1 Colonoscopy images can be classified as clear images (**a**, **b**) and non-informative images (**c**–**f**). Non-informative images may be due to fluid immersion (**c**), wall contact (**d**), extreme lighting (**e**), and irrigation (**f**)

In our previous work [5], we treated the co-alignment problem in the presence of *clear* OC images, as illustrated in Fig. 1a, b. Optical flow can represent the actual image motion by measuring colon fold's displacements. A typical colonoscopy video stream, however, comprises many *non-informative images* that contain little or no image features; these images typically make up 30–40 % of the entire video stream [6,7]. Figure 1c–f illustrates four types of non-informative images caused by fluid immersion, wall contact, illumination blinding, and irrigation or bleeding. Optical flow cannot estimate image motion reliably from non-informative images due to lack of image features, leading to failure in tracking algorithms.

In this paper, we present our work on optical flow-based tracking algorithms to handle interruptions in the OC video stream due to non-informative images. We describe three novel techniques—*temporal volume flow, region flow, and incremental egomotion estimation*, resulting in a robust and accurate estimation of the camera motion across a non-informative image sequence. Figure 2 shows an example colonoscopy sequence (from our experiments) with non-informative images. It contains two clear OC image sequences bridged by a non-informative image sequence. Our optical flow-based approach [5] can track up to the end of the first clear image sequence, but fails when non-informative images are encountered. The exclusion of non-informative images generates a motion gap between frames 325 and 364. To estimate the camera motion, we need to find two images before and after the non-informative images, which contain similar visual features that carry camera motion information. Temporal volume flow is used to determine the best image pair by exploiting temporal coherence. As the camera motion is significant between the selected image pair, we propose region flow and incremental egomotion estimation to compute the final camera motion parameters. The combination of temporal volume flow, region flow, and optical flow was used to successfully track the entire image sequence shown in Fig. 2.

Endoscopy video tracking techniques can be broadly classified into three categories, based on image motion velocity and frame-to-frame coherence: (1) moderate velocity and contiguous frames, (2) rapid motion and contiguous frames, and (3) interruptions due to non-informative images.

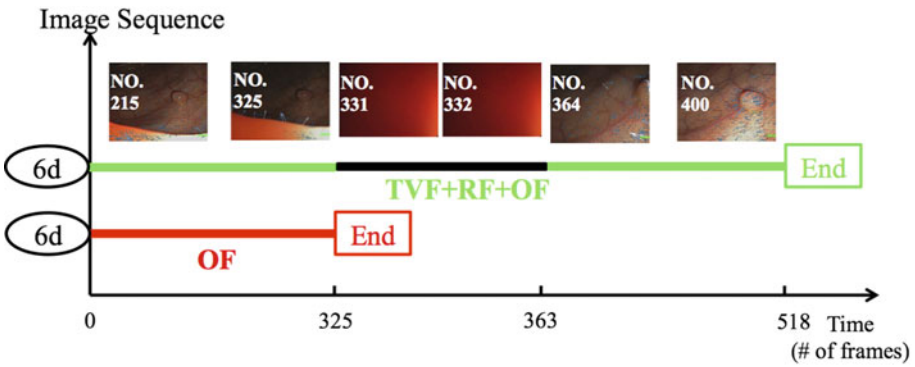

Fig. 2 Tracking optical colonoscopy (OC) images over non-informative image interruptions. The *horizontal* axis is over time (#frames) representing tracked images, and the *vertical* axis represents an image sequence tracked using two different schemes: optical flow (OF) approach [5] (in *red*) and the combination of temporal volume flow (TVF), region flow (RF), and optical flow (in *green*). The *black bar* represents a non-informative image sequence. The current colonoscopy image sequence contains 518 images (corresponding to Fig. 13) to illustrate the challenges of non-informative image interruption. The OF approach fails at frame 325 (onset of non-informative images), while our combined approach successfully tracks the entire sequence

Table 1 The three categories of endoscopy tracking problems and their corresponding solutions

Optical conditions	Existing approaches	Our approach
Moderate velocity and contiguous frames	[11–16] [17–21]	Optical flow
Rapid motion and contiguous frames	[8,9]	Region flow
Interruption from non-informative frames	[10]	Temporal volume flow

Most of the prior approaches were concerned with the first category, with few solutions for the second and third categories. Table 1 summarizes the status of endoscopy tracking methods. A detailed survey can be found in Liu [5]. Matching optical and virtual endoscopy images was a primary approach to track contiguous bronchoscopy frames with moderate velocity, given the minimal deformation and the ability to exploit features such as bifurcations in the bronchi. However, these methods become unstable when the motion is large or rapid, violating the small motion assumptions in their governing equations. External tracking devices are efficient tools to estimate rapid rigid motion because endoscopes equipped with such sensors are not subject to loss of position from non-informative video sequences. However, there are reports in the literature on bronchoscopy tracking [8,9] that sensor-based methods are sensitive to local image deformation and easily cause misalignment of optical and virtual images. Tracking errors can be moderated by estimating camera motion from local image motion [8,9]. Our approach complies with this strategy to compute rapid motion as the purpose of our work is to track endoscopy images without sensors.

The appearance of non-informative images further complicates the tracking problem because image motion cannot be reliably calculated when the video is obscured, blurred, or otherwise compromised. Neither computer vision techniques nor sensor-based approaches [8,9] can reliably track an endoscope at all times. Under these conditions, the solution in bronchoscopy tracking [10] was to let the user manually adjust bronchoscope. Manual adjustment is impractical in colonoscopy procedures, since more than 30 % of the images are non-informative [6,7], resulting in the user having to frequently reset the tracking system.

Thus, the work presented here aims to continuously and automatically track OC and VC images across interruptions caused by non-informative images. Our system comprises three stages:

- Temporal volume flow compares two temporal volumes before and after a non-informative image sequence, in order to search for an image pair by using temporal coherence.
- Region flow computes relative image displacements of all image regions between the selected image pair; significant image motion is accurately computed by employing region flow vectors that use reduced feature ranges.
- Incremental egomotion is used to estimate image motion across the non-informative image sequence by subdividing it into a sequence of small image displacements, and the camera motion is incrementally estimated from these displacements.

We evaluate our new methods using both phantom experiments and clinical colonoscopy data. Two colon-like phantoms, a straight phantom and a curved phantom (both in the shape of a tunnel), are constructed with a high degree of confidence of the accuracy necessary to generate ground truth for quantitatively validating the proposed methods. In the straight phantom, after 48 frames were excluded, the error was <3 mm of 16 mm traveled. In the curved phantom, after 72 frames were excluded, the error was <4 mm of 23.88 mm traveled. We also tested the colonoscopy tracking algorithm on 30 clinical colonoscopy image sequences from 22 patients, and spanning five different colon segments.[1] These experiments demonstrate the robustness of our method with respect to the following scenarios that result in non-informative images:

- Wall contact,
- Fluid immersion,
- Structural changes due to surgical removal of polyps, and
- Multiple non-informative image sequences.

Methods

Our strategy for tracking colonoscopy video in the presence of non-informative images is similar to real-world navigation; look for landmarks or features that might match those encountered earlier. Non-informative images in the video stream pose difficulties since their visual information is unreliable. Thus, our first step is to match visual information in images before and after the non-informative image sequence. However, rather than picking arbitrary images on either side of the sequence, we analyze the temporal volumes that bridge the non-informative image sequence and compute temporal volume flow (TVF), exploiting temporal coherence. This procedure results in selecting an image pair (guided by the TVF vectors) that has the best probability for success in determining the motion across the non-informative image sequence.

[1] Part of the Walter Reed Army Medical Center training dataset archive from the National Cancer Institute.

Rapid camera motion causes large visual motion between the two selected images. Wide-baseline point matching techniques [22–25] are potential solutions to estimate large image displacements by matching corresponding feature points. However, feature descriptors used in these techniques are indistinctive due to the lack of visual cues in OC images. Instead of performing point-to-point matching, we compute region flow between the two images by comparing all image regions within a global energy minimization framework. As region flow carries global motion information, it can significantly improve the accuracy and stability of large image displacement. Also, since camera motion cannot be accurately estimated from large image displacements, each large displacement vector is subdivided into a set of small optical flow vectors through an incremental egomotion estimation framework. Large camera motion is then computed by adding camera motion parameters from all the individual optical flow vectors. Figure 3 illustrates the three stages of our tracking algorithm for handling non-informative image sequences.

Finally, there is an initialization step. We manually adjust the VC images to match with the first OC image, using locations and orientations of polyps and/or colon folds. An egomotion estimator based on optical flow [5] is used to automatically track informative OC images thereafter. A blurry image detection algorithm [26] based on saturation values, intensity distribution, and edge information in the current OC image is employed to suspend the frame-by-frame tracking when non-informative images are encountered. The detection algorithm makes this decision on the current image based on a blurriness threshold. If the current image score is above this threshold, it is considered to be non-informative; otherwise, it is informative, corresponding to image t_2 in Fig. 3. As image t_1 (prior to the blurry image sequence) is recorded by our system, we can determine the informative image pair before and after the non-informative image sequence. Control then goes to the three stages of the algorithms described above by using the image pair t_1 and t_2.

Thus, we treat the overall problem from matching temporal volumes to image regions, and finally onto points, corresponding to our three main contributions: temporal volume flow, region flow, and incremental egomotion estimation. Their computations are all based on the dense comparison of visual information and represented by the following variational function:

$$E(\overrightarrow{u}(\boldsymbol{p})) = \int_{\Omega} (\underbrace{M(D^k F, \overrightarrow{u})}_{\text{Data Term}} + \alpha \underbrace{S(\nabla F, \nabla \overrightarrow{u})}_{\text{Smoothness Term}}) \mathrm{d}\boldsymbol{p} \qquad (1)$$

where $\boldsymbol{p} = (x_1, x_2, \ldots, x_n)$ denotes an n-coordinate point and $\overrightarrow{u} = (u_1, u_2, \ldots, u_m)$ is a m-tuple visual motion vector to be estimated. $D^k F$ is the set of all partial derivatives of I of order k. $M(D^k F, \overrightarrow{u})$ is a data term, and $S(\nabla F, \nabla \overrightarrow{u})$ is a smoothness constraint. α is a parameter to balance data and smoothness terms. The data term measures the similarity between two corresponding elements, such as intensity and gradient, and the smoothness term enforces the visual motion between two corresponding elements to vary smoothly except at data discontinuities.

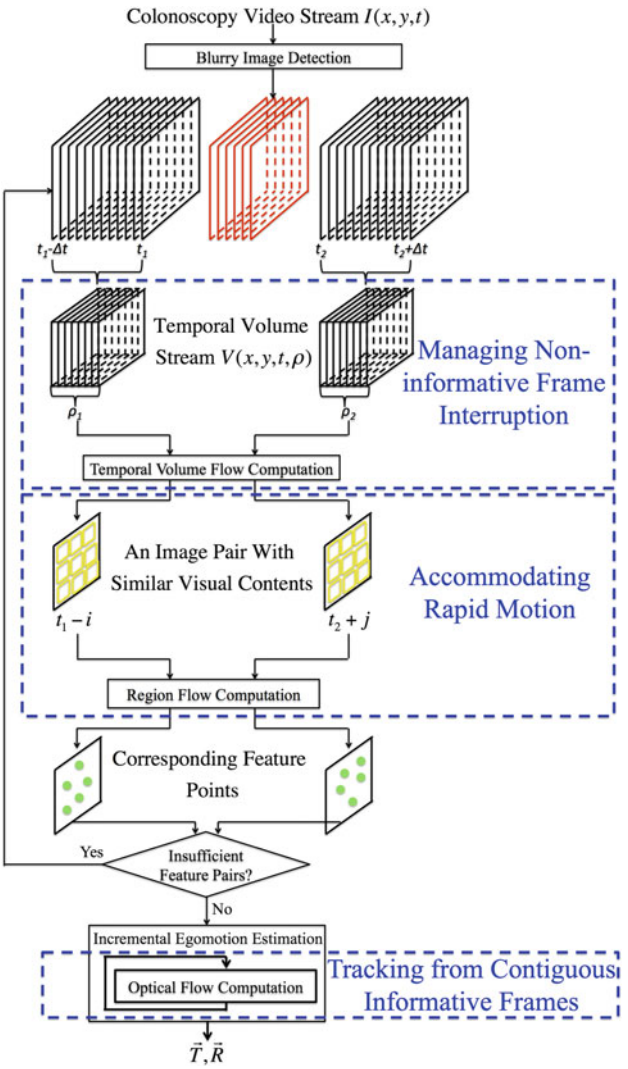

Fig. 3 The flowchart for managing non-informative image interruptions in colonoscopy video. The central idea behind our approach is to continuously reduce the tracking problem from the difficult issue of non-informative frame interruption to the easier problem of contiguous frame tracking with moderate velocities. Non-informative images (*red frames*) are first identified by a blurry image detection algorithm, and temporal volumes before and after the non-informative image sequence are then constructed. Temporal volume flow computation compares them to find a visually similar image pair. Region flow computation identifies corresponding feature points between the two selected images (shown as *green dots*). If there are an insufficient number of corresponding features, video times t_1 and t_2 are shifted and the process repeated with new temporal volumes. The results are input to the incremental egomotion estimation algorithm, an iterative optical flow computation to determine camera translational and rotational parameters

Temporal volume flow

This involves two steps, computation of the temporal volume flow field, followed by the search for the optimal image pair containing similar features.

Temporal volume flow computation

Assume there is a colonoscopy video stream $I(x, y, t)$ containing a non-informative image sequence at (t_1, t_2) (red frames, top row of Fig. 3). In the second row, two temporal volumes are constructed by collecting two stacks of colonoscopy images at $(t_1 - \Delta t, t_1)$ and $(t_2, t_2 + \Delta t)$. Without loss of generality, define ρ to represent the artificial time of a temporal volume stream. All temporal volumes form a continuous 4-d temporal volume stream, $V(x, y, t, \rho)$. Suppose ρ_1 and ρ_2 correspond to image sequences at $(t_1 - \Delta t, t_1)$ and $(t_2, t_2 + \Delta t)$, temporal volume flow performs a global comparison between $V(x, y, t, \rho_1)$ and $V(x, y, t, \rho_2)$ to find two similar images.

TVF densely matches $V(x, y, t, \rho)$ at time ρ_1 and ρ_2 as described in Eq. 1, and it is mathematically formulated as follows:

$$E(\vec{w}) = \iiint (\Psi(\underbrace{(V(x+w_x, y+w_y, t+w_t, \rho_2) - V(x, y, t, \rho_1))^2}_{\text{Intensity Constancy Term}} + \gamma(\underbrace{\nabla V(x+w_x, y+w_y, t+w_t, \rho_2) - \nabla V(x, y, t, \rho_1))^2}_{\text{Gradient Constancy Term}}) + \alpha\Psi(\underbrace{|\nabla w_x|^2 + |\nabla w_y|^2 + |\nabla w_t|^2}_{\text{Smoothness Term}}))\mathrm{d}x\mathrm{d}y\mathrm{d}t \quad (2)$$

where $\Psi(x^2) = \sqrt{x^2 + \epsilon^2}, \epsilon = 0.001$ is a modified $L1$ norm and allows the computation to handle occlusions and other non-Gaussian deviations of the matching criterion [27–29]. $\vec{w} = (w_x, w_y, w_t)$ is a TVF vector at a point $\boldsymbol{p} = (x, y, t, \rho_1)$. α and γ are two constants to balance different components in Eq. 2. They are experimentally set to 80 and 5, respectively. These values were also empirically validated by Brox [29]. Minimizing Eq. 2 with respect to $\vec{w}$ generates TVF.

The Euler–Lagrange equation can be used to solve Eq. 2 and for the x component is given by

$$\begin{aligned}&\Psi'((\partial_\rho V)^2 + \gamma((\partial_{x\rho} V)^2 + (\partial_{y\rho} V)^2 + (\partial_{t\rho} V)^2))\\ &(\partial_x V \partial_\rho V + \gamma(\partial_{xx} V \partial_{x\rho} V + \partial_{xy} V \partial_{y\rho} V + \partial_{xt} V \partial_{t\rho} V))\\ &-\alpha \mathrm{div}\left(\Psi'(|\nabla w_x|^2 + |\nabla w_y|^2 + |\nabla w_t|^2)\nabla w_x\right) = 0 \end{aligned} \quad (3)$$

where

$$\begin{aligned}\partial_\rho V &= V(x+w_x, y+w_y, t+w_t, \rho_2) - V(x, y, t, \rho_1)\\ \partial_{x\rho} V &= \partial_x V(x+w_x, y+w_y, t+w_t, \rho_2) - \partial_x V(x, y, t, \rho_1)\end{aligned} \quad (4)$$

$\partial_{y\rho} V$ and $\partial_{t\rho} V$ are similarly defined.

However, Eq. 3 is a nonlinear equation with respect to $\vec{w}$ due to nonlinear intensity and gradient constancy terms. $\Psi(x^2)$ also generates non-convexity. Equation 2 is therefore difficult to minimize because Eq. 3 is non-convex and nonlinear. In order to remove nonlinearity and non-convexity, two numerical strategies are employed: multi-scale image representation and sequential linearization.

Multi-scale image representation. Non-convexity in the TVF computation arises mainly from the appearance of fine image details. Multi-scale image representations [30] effectively handle this issue because they suppress fine details at coarse scales, helping to better identify a global minimum. Temporal volume pyramids are used as the multi-scale image representations in TVF computation by down-sampling temporal volumes. They can smooth fine details as well as reduce computational cost. In our implementation, the down-sampling rate is chosen to be 0.75 between successive image resolution levels to ensure smooth transition across different scales.

Sequential linearization. This step aims to remove nonlinearity in Eq. 3. TVF vectors to be estimated are decoupled from other nonlinear functions, such as Ψ', through two nested iterations (as detailed in the appendix). Finally, each voxel has three linear equations corresponding to x, y, and z-directions, which leads to a large and sparse linear system to compute TVF. Successive over-relaxation (SOR) method [31] is applied to solve it. We summarize TVF computation in the online supplement.

In our implementation, a temporal volume consists of 20 consecutive OC images, which corresponds to 0.67-s interval between successive images, assuming the recording rate is 30 frames/s in OC videos. Across such short intervals, TVF will not miss polyps or other anatomical/pathological features. The computational cost is also significantly reduced. A thorough validation on the number of OC images to make up a temporal volume can be found in our earlier work [5]. An iterative search is performed in our recovery framework, as shown in Fig. 3. If there are insufficient feature matches between the selected images, we shift five frames ($t_1 - 5$ or $t_2 + 5$ in Fig. 3) to recreate new temporal volumes before and after non-informative images.

Figure 4 illustrates the TVF results on two temporal volumes separated by non-informative images due to wall contact. Here, temporal volumes are composited to two images through volume rendering techniques, and TVF vectors are represented as arrows. Note that the fold in the left image moves to the bottom left corner in the right image, and flow vectors accurately capture the movements between the twofolds.

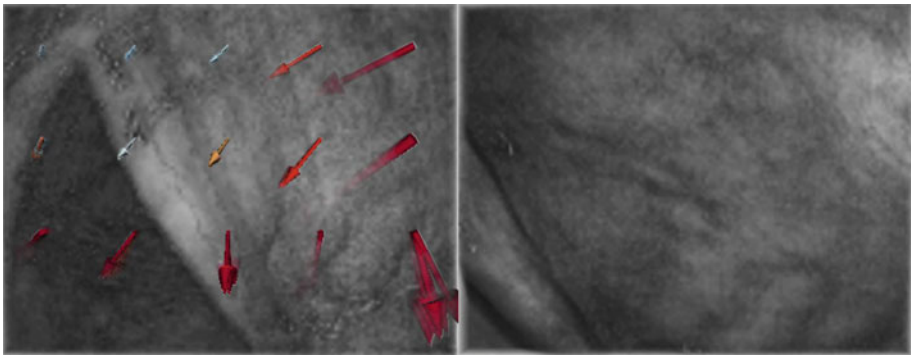

Fig. 4 Temporal volume flow (TVF) results on a non-informative image sequence. Here, volume rendering techniques are used to visualize temporal volumes and to composite *left* and *right* images. TVF pointing to *bottom left corner* accurately reflects relative displacements between colon folds in the *left* and *right* images

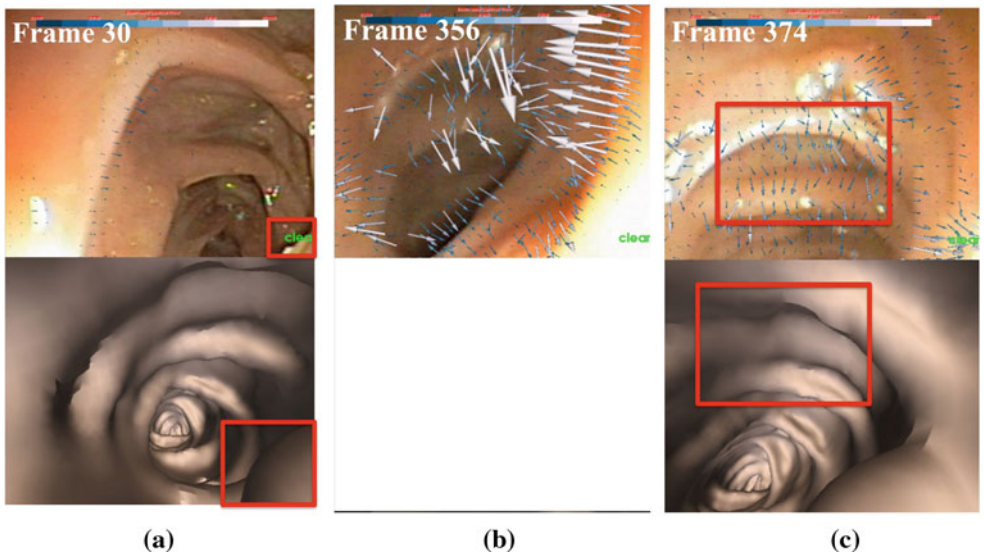

Fig. 5 Comparison of tracking results with and without temporal volume flow (TVF) in the descending colon after polyp removal. **a** OC and VC images at frame 30 before a non-informative sequence; *red rectangles* indicate polyp locations, **b** selecting frame 356 after the non-informative image sequence for matching causes a recovery failure, **c** TVF chooses frame 374 to successfully continue tracking, because the same folds (*red rectangles*) appear in both OC and VC images

Image pair search

After TVF is computed, we track all possible voxel displacements between the two temporal volumes. Then we count the number of all possible voxel correspondences connected by TVF vectors, between every image pair between the temporal volumes. Thus, if there are N images in both temporal volumes, there are $N \times N$ pairs of images that will be considered. We select the image pair that has the largest number of voxel correspondences.

Figure 5 compares the tracking results on an OC sequence with and without TVF-assisted image pair search. Figure 5a shows the co-aligned OC image (top) and VC image (bottom) at frame 30 prior to non-informative images. Frame 356 is the image just after the non-informative sequence. Note that the colon folds in Fig. 5a are scaled down and lifted, which causes substantial dissimilarity between frame 30 and 356. Thus, the VC image fails to co-align with the OC counterpart, represented as a blank image in Fig. 5b. Instead, TVF selects frames 30 and 374 to compute the motion. The corresponding folds remain at approximately the same regions although the folds are lowered. Figure 5c shows that OC and VC images are successfully co-aligned due to a more similar image pair chosen by the TVF.

Region flow

Region flow is a global region-to-region matching method to measure large image motion between two images (second blue frame of Fig. 3). After TVF computation, we obtain an image pair with large image displacements, such as Fig. 5a, c. Large image motion displacements have been defined by wide-baseline image matching methods [22–25,32,33], to be larger than 1 or 2 pixels in successive video images. The difficulty in estimating large image motion increases quadratically with the magnitude of image displacement, making wide-baseline image matching impractical for this purpose. Region flow explores all image regions to globally

understand large displacements of all image points. Local SIFT feature matching [22] constrained by global region flow can significantly enhance the accuracy of large image displacement estimation. This section describes region flow–based image motion estimation.

Region flow computation

To simplify the description, let $I_1(x, y)$ and $I_2(x, y)$ be a pair of selected images at $t_1 - i$ and $t_2 + j$ shown in Fig. 3. The essential feature of region flow computation is to use region comparison to replace point matching in optical flow methods [28,34–36], to reduce its sensitivity to large image displacements. The similarity between two regions of $I_1(x, y)$ and $I_2(x, y)$ can be measured by normalized cross-correlation(NCC) [37]:

$$NCC(x, y, \overrightarrow{r}) = \iint \hat{I}_2(x + r_x, y + r_y)\hat{I}_1(x, y)\mathrm{d}x\mathrm{d}y$$
$$\hat{I}_1(x, y) = \frac{I_1(x, y) - \bar{I}_1}{\sigma_{I_1}} \quad \hat{I}_2(x, y) = \frac{I_2(x, y) - \bar{I}_2}{\sigma_{I_2}} \tag{5}$$

where $\bar{I}_1$ and $\bar{I}_2$ are mean values, and σ_{I_1} and σ_{I_2} are standard deviations. $\overrightarrow{r} = (r_x, r_y)$ represents a region flow vector at point (x, y). NCC values are in the range $[-1, 1]$, and two regions are matched if the NCC value is maximized.

In order to fit NCC measurement into the minimization framework of region flow computation, Eq. 5 is rewritten to minimize $1.0 - NCC(x, y, \overrightarrow{r})$. Similar to TVF computation, a global energy function similar to Eq. 1 is applied to compute region flow, within a minimization framework:

$$E(r_x, r_y) = \iint \underbrace{\min(|1.0 - NCC(x, y, r_x, r_y)|, \alpha)}_{\text{NCC Term}} + \lambda \underbrace{\min((|\nabla r_x|^2 + |\nabla r_y|^2), \beta)}_{\text{Smoothness Term}} \mathrm{d}x\mathrm{d}y \tag{6}$$

where α and β are truncation values to prevent oversmoothing and λ is a parameter to balance data and smoothness constraints. In our implementation, we set $\alpha = 0.8$, $\beta = 50$, and $\lambda = 1$ to process both phantom and clinical image sequences in our experiments.

The original resolution of OC images is 720×480, and the selected two OC images are first down-sampled by a factor of 4 to reduce the computational cost. The accuracy of region flow might be decreased because of its computation on down-sampled images. However, highly accurate region flow is unnecessary at this step because the purpose of region flow is to provide search ranges for SIFT feature matching described in section "Feature matching." The final large image motion is determined by measuring relative displacements between matched SIFT features (SIFT features are extracted from the original OC images). Consequently, the estimation of large image motion will not be affected by down-sampling OC images for region flow computation.

The computation begins by calculating NCC measurement to match image regions in the down-sampled source image to corresponding regions in the down-sampled target image at every pixel. Its computational cost is $O(N^4)$ for $N \times N$-sized images. Equation 5 indicates that the NCC value not only depends on (x, y) and $\overrightarrow{r}$, but also correlates with $\bar{I}_1$ and σ_I of candidate image regions at source and target images. The multivariable NCC function is easy to compute through discretely matching image regions. A Markov random field [38] is an efficient approach to minimize Eq. 6 with a discrete NCC term by converting images into four connected graphs. Efficient belief propagation [39,40] is used to minimize the connected graph. The detailed implementation to compute region flow can be found in our earlier work [5].

Figure 6 compares image displacements measured by region flow and optical flow on an image pair separated by non-informative images. Figure 6b illustrates an OC image with overlaid region flow vectors. They represent the actual image displacements between Fig. 6b, c. Three corner points are manually selected and indicated by white boxes in Fig. 6a and green boxes in Fig. 6b. The white squares in Fig. 6c represent corresponding pairs generated by optical flow. They do not match up with the green squares, which roughly represent the positions of the true corresponding pairs.

Feature matching

Figure 7 illustrates the process of feature matching based on region flow. Two sets of SIFT feature points are detected on the *original-sized* colonoscopy image pair, illustrated as white crosses in Fig. 7. The SIFT algorithm [22] is chosen because it usually generates a sufficient number of feature points. This property is useful for colonoscopy tracking, considering that colonoscopy images often lack sufficient visual cues.

Region-to-region matching In this step, corresponding regions are identified using region flow field and a local matching procedure. The corresponding regions of SIFT feature points in the target image are identified using region flow vectors and a local neighborhood search. In Fig. 7a, the green squares joined by the white lines represent corresponding regions containing at least one SIFT feature point in the source image and 0 or more SIFT feature points in the target image. In the implementation, the mapped region is locally adjusted using NCC as a metric to find the best region match.

Point-to-point feature matching In this step, each corresponding region pair is refined to a corresponding point pair. If the target region does not contain a SIFT feature point, it is removed. For target regions with multiple SIFT feature point

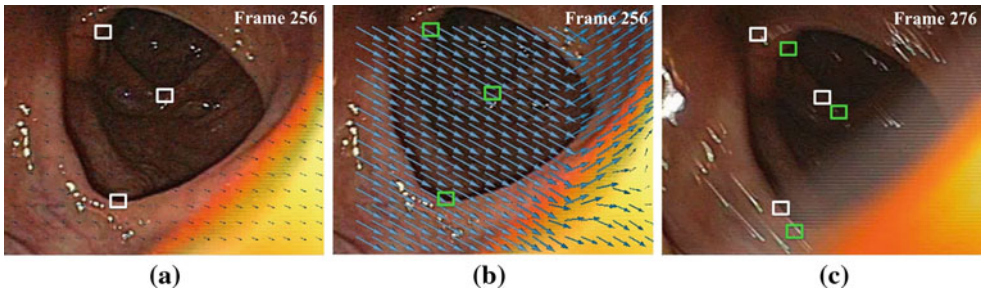

Fig. 6 Region flow versus optical flow for describing large motion, **a** source image with overlaid optical flow vectors, **b** source image with overlaid region flow vectors, **c** target image after a 20 frame non-informative sequences. The lengths of the vectors in the source images represent the magnitude of the motion velocity. Three corner regions are marked by *white squares* in the source image **a** and *green squares* in the source image **b** to assist in the accuracy comparison between optical and region flow. Their corresponding regions are also highlighted as *green* and *white squares* in the target image **c** after application of optical and region flow vectors. Region flow does a better job tracking the image motion because *green squares* are located at approximately the same corner regions, in contrast to the *white squares*

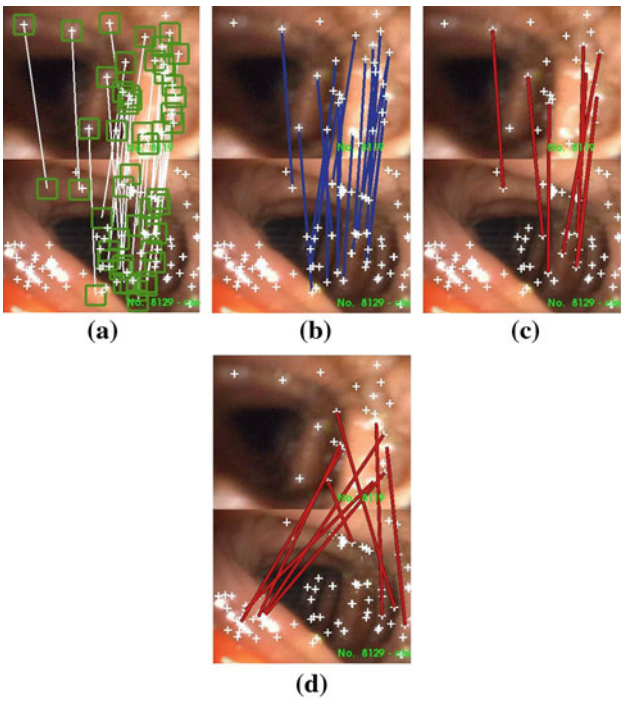

Fig. 7 Corresponding pair computation. *Top* and *bottom* images represent images before and after the non-informative image sequence, **a** region-to-region matching. *Green squares* indicate the matched regions using the region flow field. Local search using NCC is performed to find the best region pair, **b** point-to-point feature matching. Using SIFT descriptor as a metric, the best SIFT feature point pair is determined between source and target regions, **c** false feature match rejection using epipolar geometry, **d** original SIFT feature matching. Note the incorrect feature matches using SIFT only approach

candidates, the candidate with the closest SIFT descriptor (a distance metric) is chosen as the best candidate. Figure 7b illustrates the selected feature point pairs after this step.

False feature match rejection With the chosen feature point pairs, epipolar geometry is built using the RANdom SAmple Consensus (RANSAC) algorithm [25]. Matched feature points should stay at corresponding epipolar lines in the source and target images based on epipolar geometry. Feature pairs that fail to fulfill this condition are removed, as seen in Fig. 7c.

Region flow generates accurate SIFT feature matches because region flow vectors predefine feature matching ranges and limit false feature matches. In comparison, original SIFT feature matching generates significant mismatches in Fig. 7d because the matching size is uncertain and the SIFT feature descriptor is indistinct.

Region flow results in a set of matched SIFT feature points, and accurate large image motion is obtained by measuring relative displacements between the matched points.

Incremental egomotion estimation

Most existing egomotion estimation methods [41–44] fail to accurately estimate rapid camera motion because they assume that image motion should be small. To address this issue, we developed an incremental egomotion estimation strategy to subdivide every large image displacement vector into a sequence of optical flow vectors by iteratively performing point-to-point optical flow computation (bottom of Fig. 3). Rapid camera motion is estimated by combining small camera motion parameters from each iteration.

We use Eq. 1 to combine global SIFT matching term and local image intensity and gradient constancy terms within a variational equation, and minimizing this equation yields the subdivision of large displacement vectors. The Euler–Lagrange equation is unfortunately invalid because the SIFT matching term is non-differentiable. Instead of keeping the SIFT matching term during the minimization process, we advance it to the initialization phase by performing image region matching on every SIFT feature correspondence. Figure 8 illustrates image region matching. For each feature correspondence, two regions centered at matched feature points, such as Fig. 8b, c near a polyp, are first built in the OC image pair. Optical flow method [28] is employed

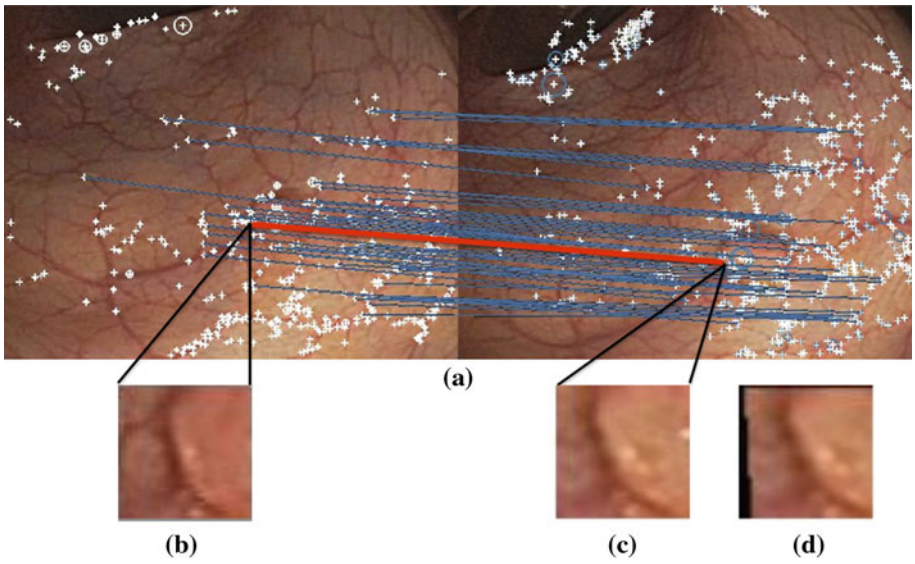

Fig. 8 Region-to-region image matching. **a** SIFT feature matches, where *white crosses* indicate SIFT features and lines link matched feature pairs, **b** an image region centered at a SIFT feature in the *left* image, **c** the image region in the *right* image, **d** the warped image region in terms of the image matching results

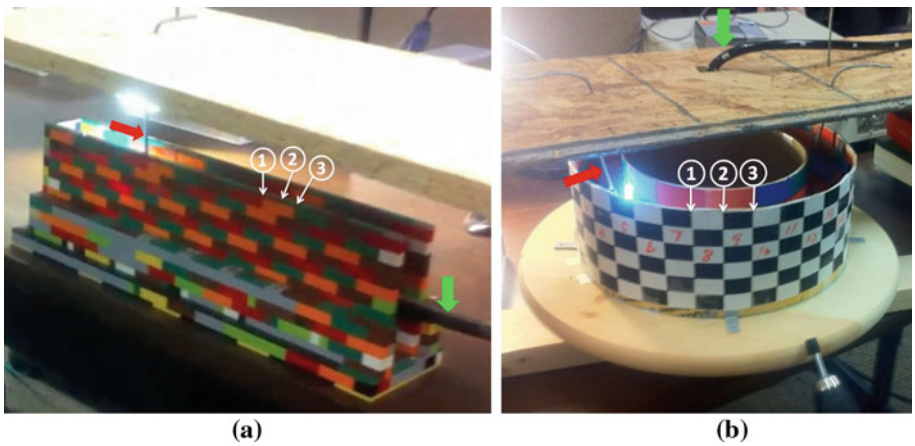

Fig. 9 Phantom design. **a** Straight phantom, **b** curved phantom

to compute image motion between two regions. Figure 8d is the warped equivalent of Fig. 8c using the image matching results. Image matching accurately measures relative displacements between two image regions because Fig. 8b is similar to Fig. 8d. Discrete SIFT matching term is replaced by a term that measures the difference between optical flow vectors and the vector sum of region matching vectors and large image displacement vectors. This replacement makes the Euler–Lagrange equation become valid, and large displacement vectors can be iteratively decomposed into a sequence of optical flow vectors. We modify our egomotion estimation method [5] to incrementally estimate camera translational and rotational parameters at every iteration. The detailed implementation of incremental egomotion estimation can be found in the online supplement of this article.

The camera motion parameters are finally determined by summing all incremental motion parameters. They are used to transform the VC camera, and the OC and VC images are aligned.

Phantom validation

In order to evaluate the accuracy and robustness of our tracking algorithms, we constructed two colon-like phantoms (curved and straight), as seen in Fig. 9. Colonoscope velocity and displacement were the metrics used to compare the algorithm performance to the generated ground truth.

Phantom experiment setup

We designed the two phantoms to satisfy the following requirements:

– *Repeatable.* The experiments must be capable of being performed multiple times under the same conditions and collecting multiple image sequences for statistical analysis.

– *Consistent.* To simulate the actual colonoscopy procedure, the colonoscope must be placed inside the phantom and moved at typical speeds.
– *Controllable.* Colonoscope velocities must be adjustable, with the ability to constrain other free parameters, enabling validation of the most important parameters, camera velocity and displacement.
– *Precise.* The actual camera velocities and displacements must be measurable, enabling reliable evaluation of the tracking system.

The repeatability requirement suggests the use of a motor to move the optical colonoscope. However, this is nontrivial, since the colonoscope has a long flexible tube and a heavy handle. Instead, we motorize the phantoms to generate relative camera motion. We chose LEGO blocks to build the straight phantom since (1) LEGO phantoms can be easily built to fulfill the controllability requirement and (2) LEGO products have high precision, with a manufacturing tolerance as small as 0.01 mm [45]. As a result, the phantoms can be easily built by using LEGO bricks to fulfill the requirements of controllable experiments, accurate ground-truth determination, and consistent navigation.

In the straight phantom experiment, LEGO bricks were used to build a straight-tunnel phantom (Fig. 9a), with the interior size of 105 mm × 32 mm × 384 mm. The colonoscope (indicated by green arrows) was suspended by the iron wire (red arrows), while the straight phantom is driven by a motor. The straight phantom was translated at three speeds of 10, 15, and 20 mm/s. At each speed, five image sequences were collected.

The curved phantom was built using two concentric sheets (thick cardboard) of radii 158.5 and 102.5 mm (Fig. 9b). The height of each sheet is 125 mm. Textured (color squares) patterns coat the inside of the two curved sheets, simulating the effects of LEGO bricks. The size of each colored square is 54 mm × 28 mm. A small wheel of radius 0.6 mm is attached to the end of the drill and used to rotate the turntable. Similar to the straight phantom experiments, the colonoscope is kept stationary by iron wires, while the curved phantom was rotated at three different speeds. The detailed experimental setup can be found in [5].

Note that our phantoms are rigid while the actual colon is a deformable organ. The color and texture of phantom images are also quite different from the actual colonoscopy images, although the phantom images are recorded from a real colonoscope (Fig. 12). However, extracting ground-truth from a deformable phantom is difficult, especially if we are to simulate the complicated motion of a human colon. If the ground-truth cannot be reliably obtained, then validating the accuracy of our tracking algorithms will not be possible, defeating the primary goal of the phantom experiments. Our phantom design supports interior navigation, permits typical colonoscope speeds, and most importantly allows accurate measurement of camera velocity and displacement and the ability to perform multiple trials of the same experiment. We can thus evaluate our tracking system with confidence, unlike prior phantom experiments [8,19,21] in bronchoscopy tracking. Although their phantom shapes are more faithful to the bronchi, they are also rigid, and the color and texture of the endoscope images vary from the actual bronchoscopy images.

The goal of our phantom validation is the evaluation of large camera motion caused by the appearance of non-informative images. Thus, we only chose high-velocity colonoscopy sequences, about 20 mm/s. We created large intervals of varying durations with as few as 19 informative images in a 430-image straight phantom sequence and 13 images in a 430-image curved phantom sequence. All other images were eliminated to simulate non-informative image interruptions. In the straight phantom, 19 images are selected when the iron wire highlighted by a red arrow arrives at the positions shown by the three white circles in Fig. 9a. The selected positions are the boundaries between LEGO bricks or LEGO tilts and are observed from an external video camera. The camera displacement between two adjacent selected images is 16 mm (half brick length). Similarly, 13 phantom images are chosen in the curved phantom when the colonoscope stays at the boundaries between white and black checkerboards in Fig. 9b. The camera displacement between two adjacent curved images is 23.88 mm.

Validation results

Figure 10 shows the tracking results by using region flow and incremental egomotion estimation on the straight phantom image sequences. Here, the camera translational velocity is measured in mm/interval, not mm/s. By the same token, camera rotational velocity is measured by degree/interval. Here, we chose the angle between the z axis of the estimated camera and the medial axis of the straight phantom to evaluate the camera rotational velocity error. Figure 10a indicates that the maximum translational velocity error is under 5 mm/interval on both original and calibrated phantom image sequences (five each) in the straight phantom, after 19 phantom images have been tracked. The average translational velocity error is <3 mm/interval on the original image sequences and <4 mm/interval on the calibrated image sequences. Figure 10b shows that the maximum rotational velocity error is <3.2 degree/interval on both original and calibrated phantom image sequences in the straight phantom. The average rotational velocity error is <1.1 degree/interval on the original image sequences and <2.1 degree/interval on the calibrated image sequences. In Fig. 10c, the average displacement error is <7 mm on the original image sequences and <8 mm on the calibrated image sequences.

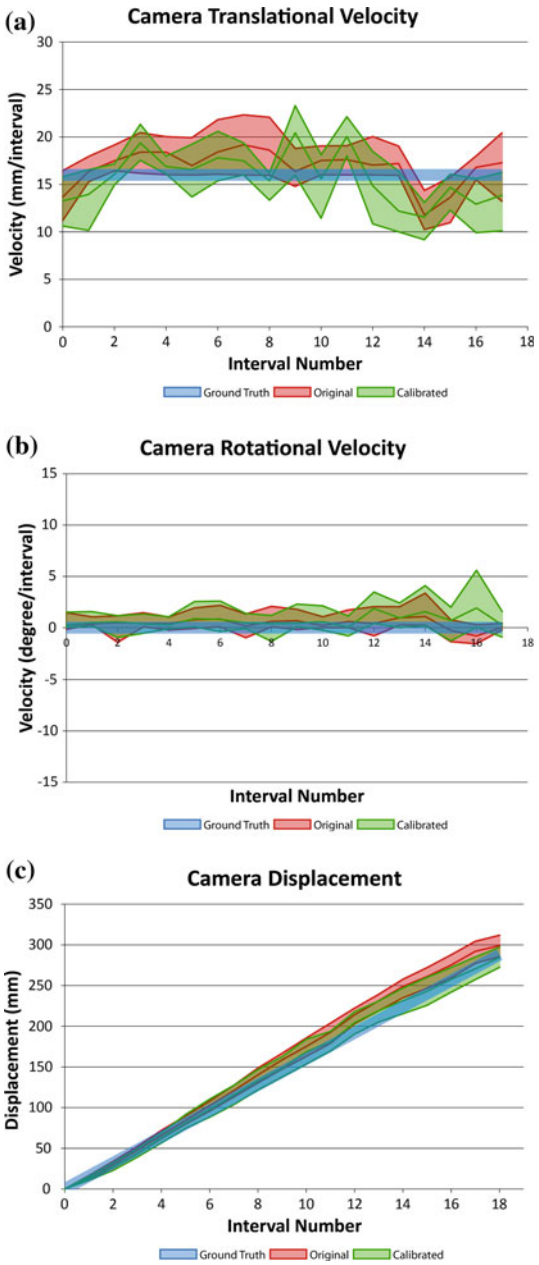

Fig. 10 Comparison between ground-truth and estimated camera motion parameters on the original and calibrated straight phantom image sequences. **a** Camera translational velocity curves, **b** camera rotational velocity curves, and **c** camera displacement curves. The *blue line* represents the ground-truth, and *red* and *green bands* indicate the estimated motion parameters on the original and calibrated phantom image sequences, respectively. The *bottom* and *upper curves* in each band indicate the minimum and maximum motion parameters of five trials, and the *center curve* represents the average motion parameters

The maximum displacement error is <13 mm on both original and calibrated image sequences. The velocity and displacement information in straight phantom experiments is also summarized in Table 1 of the online supplement of this article.

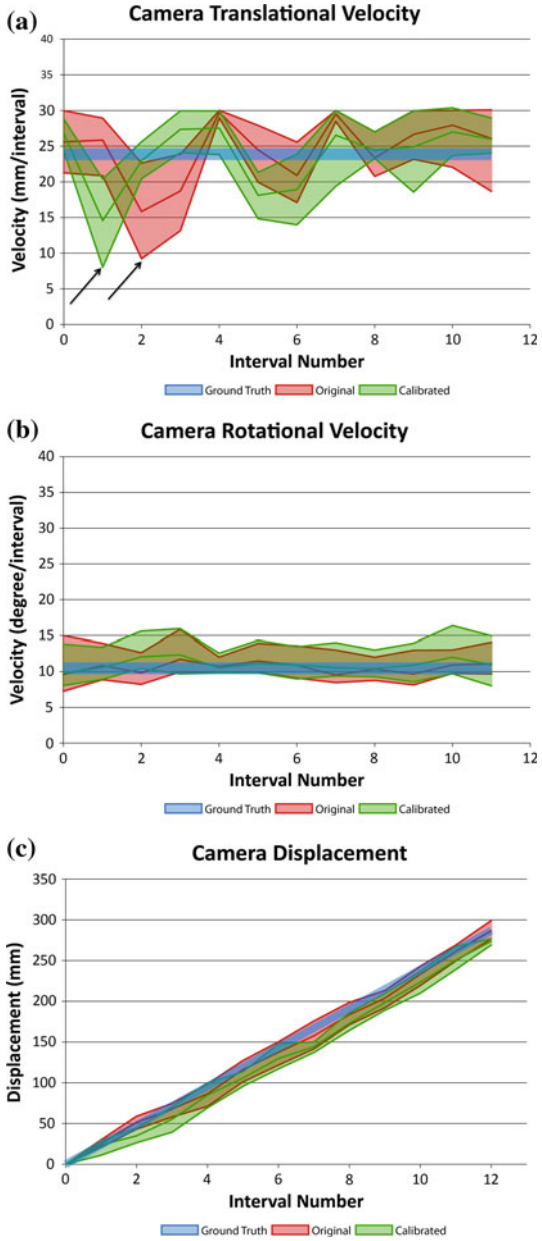

Fig. 11 Comparison between the ground-truth and estimated camera motion parameters on the original and calibrated curved phantom images, **a** camera translational velocity curves, **b** camera rotational velocity curves, and **c** camera displacement curves

Large camera motion parameters are more challenging to estimate in the curved phantom image sequences because the colonoscope moves 23.88 mm between two adjacent images. Figure 11a shows that there is a significant translational velocity error at frame 1 in the original phantom image sequences and at frame 2 in the calibrated sequences (indicated by black arrows) in the second trial. This was investigated, as shown in Fig. 12. There is a vertical curve highlighted by a red ellipse in the left phantom image, which is located on the inner wall of the curved phantom. Some SIFT feature points were detected on the vertical curve, while

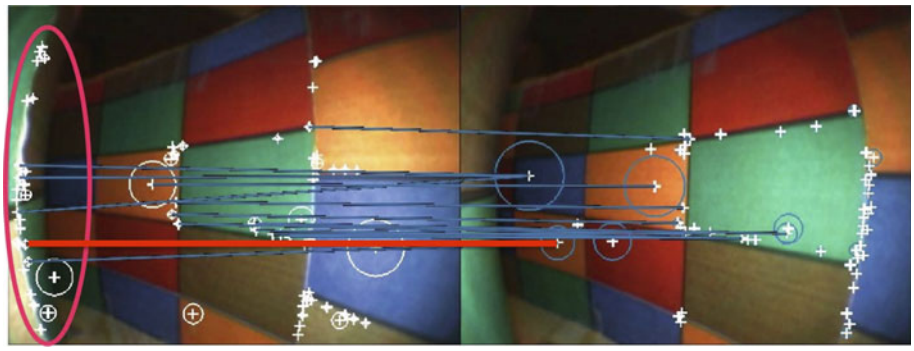

Fig. 12 SIFT feature matches between two phantom images. Some SIFT feature points inside a *red ellipse* are detected, along a *vertical curve* in the *left* image, while all these points disappear in the *right* image because the *vertical curve* is occluded. During incremental egomotion estimation, several false SIFT feature matches from the elliptical region were chosen, causing significant estimation error. An example false SIFT feature correspondence connected by a *red line* is shown

they disappear from the right image because the vertical curve no longer exists due to large camera motion. These cause false SIFT feature correspondences that are used by the incremental egomotion estimation procedure (for instance, a false SIFT feature correspondence is shown by the red line in Fig. 12). The image motion vector calculated from this false feature match would point to the image center corresponding to the image motion when the colonoscope moves backward. However, the colonoscope is currently moving forward. Thus, all false feature matches from the vertical curve reduce the estimated camera motion parameters during egomotion estimation. As a result, there is a significant drop of the estimated camera velocity at frame 1, indicated by a black arrow in Fig. 11a. For the same reason, the camera velocity is underestimated at frame 2 on the calibrated image sequence.

With the exception of the second curved phantom image sequence, the camera velocities are estimated reasonably well in all the remaining sequences. The average translational velocity error is <3 mm/interval in both original and calibrated phantom image sequences. The maximum velocity error is <8 and 7 mm/interval on the original and calibrated image sequences, respectively. Similar to the straight phantom, the rotational velocity error was measured by the angle between the Z axis of the estimated camera and the tangent direction of the medial axis of the curved phantom. The average rotational velocity error is <1.5 degree/interval in both original and calibrated phantom image sequences. The maximum velocity error is <6 degree/interval and 7 degree/interval on the original and calibrated image sequences, respectively. The average camera displacement error is <8 mm on the original phantom image sequences and <7 mm on the calibrated sequences. The maximum camera displacement error is <14 mm on both original and calibrated curved phantom image sequences. Similar to straight phantom experiments, we have summarized the validation results in Table 2, in the online supplement of this article.

From the phantom experimental results, we can draw the following conclusions:

1. Both straight and curved phantom results demonstrate that our approach can accurately recover large camera motion. The average velocity error is 3 mm of 16 mm in the straight phantom and 3 mm of 23.88 mm in the curved phantom. If the colonoscope is moving at the average speed of 10 mm/s, our method can accurately recover the tracking system after $\frac{16\text{ mm}\times 30\text{ frames/s}}{10\text{ mm/s}} = 48$ frames are excluded in the straight phantom, and $\frac{23.88\text{ mm}\times 30\text{ frames/s}}{10\text{ mm/s}} \approx 72$ frames are eliminated in the curved phantom.
2. The accuracy of large camera motion estimation is dependent on the amount of the colonoscope's movement because large camera motion will cause a large portion of SIFT features to be occluded.
3. There is no significant variance in results between the original and calibrated colonoscopy image sequences, indicating that camera calibration is not required.

Clinical data evaluation

We randomly selected 30 image sequences from 22 patients at the WRAMC virtual colonoscopy training data archive of the National Cancer Institute. These datasets were specifically collected for the training purpose in colonoscopy research, and the CRADS zero score [46] might be excluded in this study. Each patient underwent OC and VC examination, and OC and VC reports recorded polyp size and location. We collected 10 ascending colonoscopy image sequences, five transverse sequences, nine descending sequences, two sigmoidal sequences, and four rectal sequences. Each image sequence contains at least one non-informative image sequence. Fifteen image sequences have polyps, while the other 15 sequences are devoid of polyps. Only qualitative

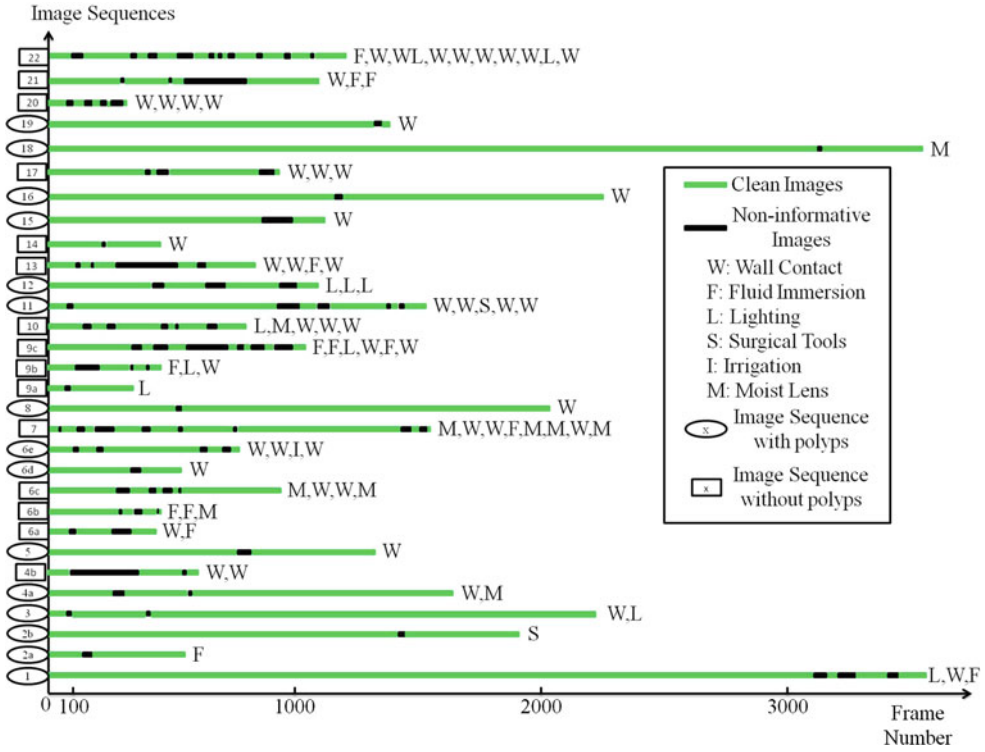

Fig. 13 Experimental results of tracking 30 clinical image sequences containing non-informative images sequences. *Horizontal* axis represents the number of tracked frames, and *vertical* axis indicates the 30 image sequences. The sequences with polyps are marked by *ellipses*, while those without polyps are marked by *rectangles*. *Green bars* denote clear image sequences, and *black bars* represent non-informative image interruptions in the sequence. *Symbols* at the end of each sequence show (in order) the causes of the non-informative image interruptions (in order) in the corresponding colonoscopy sequence (with a legend detailing the causes). In our experiments, we have been able to track more than *3500* OC images (sequence 1) and continuously over *10* non-informative sequences (sequence 22)

evaluation of the tracking accuracy is possible on the clinical sequences, since the ground truth of the colonoscope is unknown. Polyps are good landmarks to qualitatively evaluate system performance. An image sequence is considered to be successfully tracked if polyps are simultaneously located at similar regions of the OC and VC images. By the same token, colon folds can be used as landmarks on image sequences without polyps. Tracking is considered to be successful if the number of colon folds traversed is the same in both OC and VC images.

Figure 13 illustrates additional information on the causes behind the non-informative sequences and the successful use of our system in handling these sequences. We observe the following:

- We tracked more than *3,630* images with multiple non-informative image sequences, such as sequence 1. The average number of tracked images is 1,185.
- Our system tracked up to *10* non-informative image sequences (sequence 22) within a single colonoscopy image sequence. The average number of non-informative image sequences is 2.9. Assuming the frame rate is 30frame/sec, the non-informative images appear every $(1,185/30)/2.9 \approx 14$ s. Accordingly, recovery from non-informative image interruptions is essential.
- Our system successfully tracked a 273-image non-informative image sequence (sequence 4b). The number of non-informative images per non-informative image sequence varied from 5 to 273.
- In these experiments, our system encountered non-informative images due to wall contact, fluid immersion, irrigation, lighting, presence of surgical tools, and moist lens, as detailed in table 3 in online supplement.
- We noticed that the image sequence without polyps (marked by squares in Fig. 13) contained more non-informative image sequences than those with polyps (marked by ellipses). The average number of non-informative image sequence increases to 4 in the selected image sequences, with a non-informative sequence appearing every 6.4 s, on average. Most likely, this is due to the higher colonoscope velocity in these image sequences without polyps.

We next look at tracking results from the use of our system on three types of non-informative images: wall contact, fluid immersion, and polyp removal (appearance of surgical tools

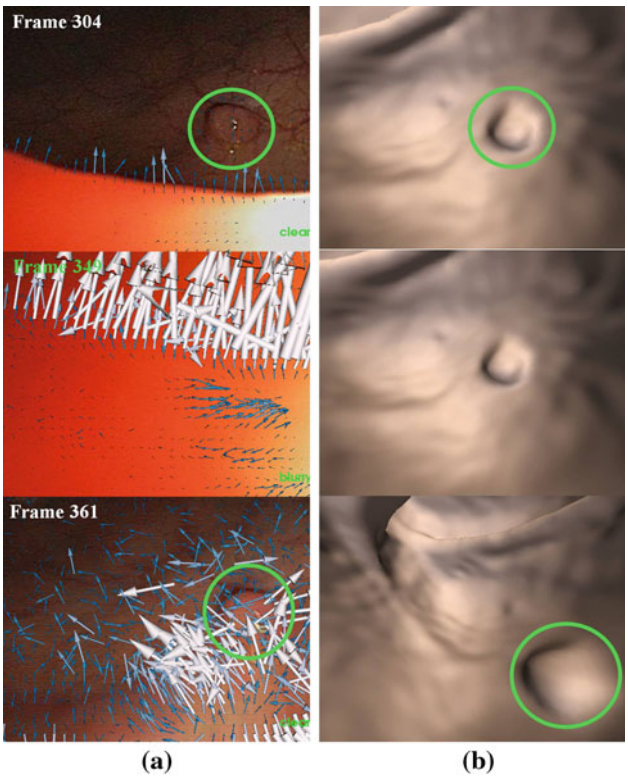

Fig. 14 Wall contact caused non-informative images. A 520-rectum image sequence with a rounded polyp is selected, where the polyp is highlighted by *green circles*. Wall contact causes a sequence of 55 non-informative images. OC images are shown in the *left column*, and VC images are displayed in the *right column*. *Top row*: the OC–VC image pair before the non-informative images; *center row*: the image pair during non-informative images; *bottom row*: the image pair after non-informative sequence

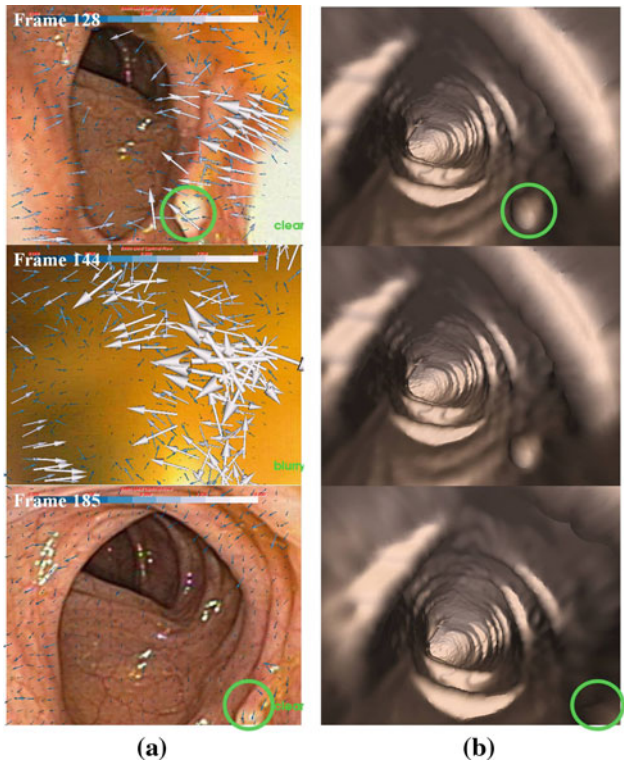

Fig. 15 Fluid immersion caused non-informative images. A 535-descending-image sequence after polyp removal is used to evaluate our tracking strategy. The colon compression produces fluid immersion, which results in 35 non-informative images as shown in the *center* image of the *left column*. In spite of fluid-induced non-informative images and complicated colon deformation, our tracking algorithm is able to recover colonoscopy, as seen by the areas of the polyp marked by *green circles* in the *top row* (before non-informative images) and the *bottom row* (after)

in the images). We conclude with an image sequence with eight non-informative image sequences, to demonstrate the robustness of our tracking algorithms.

Wall contact

In Fig. 14, a 520-rectum colonoscopy image sequence corresponding to sequence 6d in Fig. 13 was chosen to illustrate the tracking results from wall contact caused non-informative images. OC images are shown in the left column, and VC images are displayed on the right. There are 55 non-informative images starting from frame 306 to 360 in this sequence. Temporal volume flow picks frame 304 and 361 to recover the tracking system. The top row illustrates the co-aligned OC–VC image pair at frame 304 before non-informative images, and the bottom row shows the tracking results at frame 361 after non-informative OC images. The center row shows the non-informative images due to wall contact. Despite the fact that colon folds present in the top parts of VC images while they disappear in the OC images due to colon deformation, the most important feature, a rounded polyp highlighted by green circles, is still seen (displaced) in both OC and VC images. These results also mean that polyps can be used by the gastroenterologist as landmarks to maintain spatial context between the OC and VC images.

Fluid immersion

Figure 15 shows the tracking results on the sequence 2a in the descending colon when the colonoscope is being withdrawn. It contains 535 images, and 35 non-informative images occur from 135 to 170 due to the colonoscope being immersed in fluids. In this work, we are only interested in recovering the tracking system from the interruption in the withdrawal phase because colon surgery often happens in this phase. In OC images of Fig. 15, the polyp regions are marked by green circles. The top row indicates that OC and VC images are co-aligned in terms of polyp locations, though the colon deformation causes significant shape variance of colon folds. After non-informative images, the colon folds in the OC images enlarge because of deformation (bottom row), which pushes

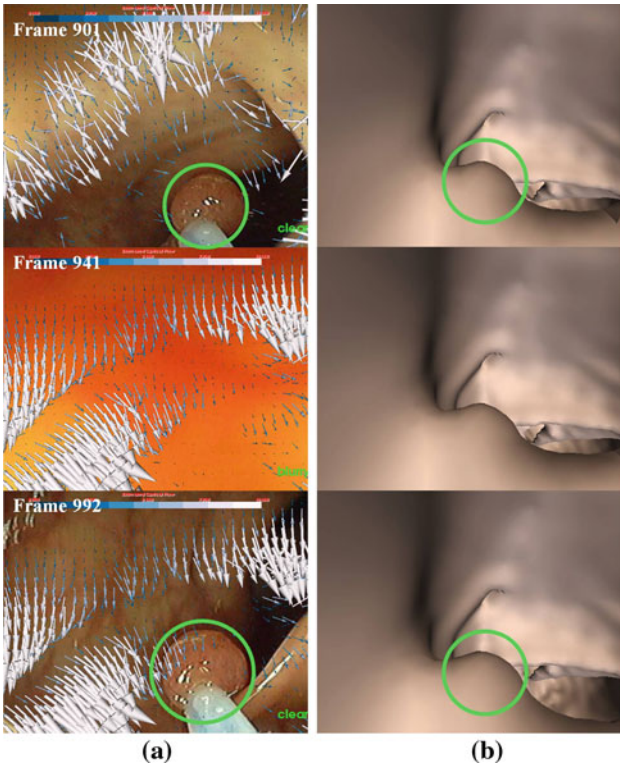

Fig. 16 Surgical tool operation caused non-informative images. A 1443 sigmoid image sequence with five non-informative image sequences is used to evaluate our tracking system. Here, we use the fourth non-informative image sequence as an example. In this sequence, the gastroenterologist uses a snare to remove the polyp, which causes the camera to touch the colon wall, generating 80 non-informative images. However, the tracking system continues to perform well, as evidenced by the corresponding polyp pairs in OC and VC images

the polyp regions to the bottom (green circle, lower right). The polyp in the tracked VC image also moves to the bottom.

Structural changes caused by surgical tools

Sequence 11 was chosen to illustrate tracking over non-informative images due to the presence of surgical tools in the optical field. Surgical tools used in the colonoscopy examination include snare, biopsy forceps, and a measuring tool. They are all observed in our selected image sequences. In Fig. 16, the snare is used to remove a polyp. This video segment contains 1,443 OC images with five non-informative image sequences. We used the fourth non-informative image sequence as an example. The gastroenterologist snared a polyp back and forth to remove it from the folds. The movement makes the colon collapse and produce 80 wall contact–like non-informative images, as seen in the top column of Fig. 16a. Using temporal volume flow calculations, OC image frames 901 and 992 were chosen to estimate camera motion. Due to small image displacements between OC images, the camera motion is accurately estimated, using polyp location as the landmark for verification.

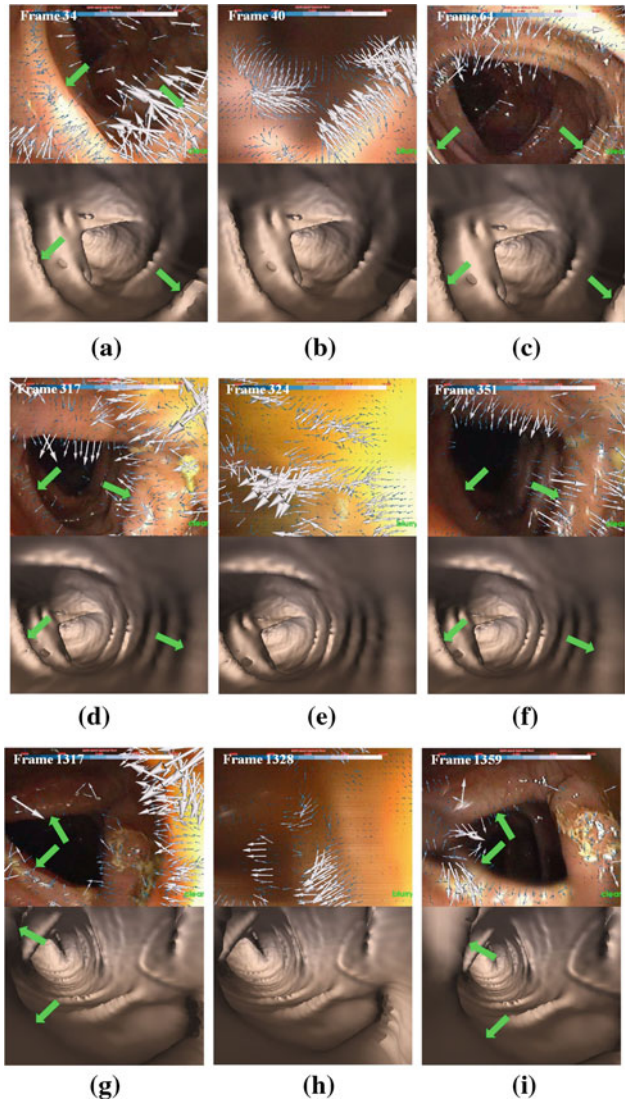

Fig. 17 Tracking results on a descending colonoscopy image sequence with eight non-informative image sequences. **a**–**c** Recovery results at No.1 non-informative image sequence; **d**–**f** recovery results at No. 4 non-informative sequence; **g**–**i** results at No.7 sequence. Here, colon folds indicated by *green arrows* are chosen to determine the relative image motion between two images interrupted by non-informative images, so as to qualitatively evaluate the accuracy of tracking failure recovery

Multiple non-informative image sequences

Sequence 7 was selected to demonstrate that our algorithm can successfully track through multiple non-informative image sequences. This sequence has eight non-informative image sequences, and the total number of non-informative images is 158. Such frequent appearance of non-informative images is partly because this sequence does not contain polyps. The dominant motion is fast withdrawal. We chose non-informative image sequences No. 1, No. 4, and No 7 for illustration, which correspond to the three rows in Fig. 17.

The non-informative images in sequence No.1 are due to the moist camera lens, as shown in Fig. 17b. Considering this sequence does not contain any polyps, we chose colon folds for qualitative evaluation. The selected folds are indicated by green arrows. In terms of relative displacements between marked folds, the camera motion between two OC images in Fig. 17a, c is moving toward the image center. We notice that the virtual camera also moves forward as two marked colon folds move toward image boundaries.

In non-informative image sequence No. 4, the colonoscope is immersed in fluids, and it is adjusted to return to normal. Since marked colon folds have small image displacements in OC images, they are well tracked, as seen in Fig. 17d, f.

The No. 8 non-informative sequence is caused by the colonoscope touching the colon wall. From the virtual colon model, we note that there is a small curved turn at the current camera location. This turn is wiped out in the OC images, which causes significant visual differences between OC and VC images in Fig. 17g. However, note that marked colon folds move to right in the OC images, and the corresponding folds exactly follow this motion by comparing Fig. 17g, i.

Conclusions and future work

In this paper, we presented a multistage framework to recover colonoscopy tracking failure from non-informative images. Our study indicates that every *14 s* on average a non-informative image will appear in a typical colonoscopy video stream. They are generated more frequently when the current OC images contain no polyps, most likely because the velocity of the colonoscope is higher when polyps are not encountered. The average interval of their appearance is reduced to *6.4 s*. Therefore, continuously tracking over non-informative image interruptions is critical to show that our methods are viable in clinical environments. Sensors that can be externally tracked can supplement our methods to overcome intractable displacements where no visual features can be found to correspond across an interval of non-informative frames.

Colonoscopy tracking over non-informative image interruption is challenging. The exclusion of non-informative images artificially causes motion gaps. We thus need to find an image pair containing the same visual contents before and after non-informative images. Our temporal volume flow algorithm serves this purpose and densely matches two temporal volumes interrupted by non-informative images to search for two images with the maximum amount of similar visual contents. Unfortunately, the selected image pair often contains large visual motion; the lack of distinctive visual cues in the OC images further complicates the visual motion computation. Region flow was developed to resolve this issue by measuring large image displacements between all image regions of the selected image pair. Combining the global region flow field and local SIFT [22] features, we can accurately estimate image displacements between two selected images. Finally, every large image displacement vector was subdivided into a sequence of small optical flow vectors through the incremental egomotion estimation. An optical flow-based approach [5] is then used to estimate camera motion during every subdivision step, and the combination of all small camera motion parameters yields the final camera motion parameters that are used to transform the virtual camera.

The strategy described in this paper was validated using straight and curved phantoms. The phantom results demonstrated that the average tracking error was 3 mm of 16 mm after 48 images were excluded in the straight phantom, and also 3 mm of 23.88 mm after 72 images were removed in the curved phantom. Moreover, there was no significant difference between the results with and without camera calibration. Thirty colonoscopy image sequences with at least one non-informative image sequence from 22 patients were used to qualitatively evaluate the robustness of the tracking framework. Our clinical results indicated that the proposed strategy was sufficient to track over different types of non-informative image interruptions, such as wall contact, fluid immersion, or surgical tool operations. The proposed algorithm can also track an image sequence with eight non-informative images, up to 358 non-informative images.

Our method successfully extended the number of tracked OC images from a few hundred [5] to a few thousand. However, our tracking system still fails to track the entire OC video stream, which is mainly caused by drift tracking errors and colon deformation. We are investigating image registration strategies to understand the similarity of OC and VC images. Temporal volume flow is another potential strategy to tackle this issue. We plan to explore a probability function to measure the degree of occlusion and our TVF approach to find an image pair with the minimum amount of occlusion.

Computational efficiency is also an important issue to be studied in the future. Our unoptimized program spends about 2–5 min of processing time per non-informative image sequence. The processing time is determined by the number of shifted temporal volumes used to find the best image pair (for TVF computation) as well as the number of SIFT feature points to compute patch flow for incremental egomotion estimation. These time-consuming computations currently prevent the applicability of our system to clinical practice. We are studying an additive operator splitting scheme [47] to reduce the computational cost of temporal volume flow and subdivision egomotion estimation. A fast normalized cross-correlation approach [48] is also being developed to accelerate region flow computation.

As for colon deformation, we are studying visual odometry [49] to reconstruct the camera trajectory and compare it

against the colon centerline to measure the deformation. The scaling effect of anus-to-cecum distance between OC and VC measurements found in Duncan [4] will be considered in the study of colon deformation. Moreover, clinical image sequences used in our study were cropped for colonoscopy training, resulting in some incomplete OC videos, which could introduce some bias. We chose video segments with important anatomical features, such as polyps and colon folds visible in both OC and VC images, to evaluate our tracking algorithm. The image sequences selected in our experiments span all colon segments, ensuring a comprehensive evaluation of our tracking system, as well as across different artifacts: colon fold shapes, fluid accumulation, etc. The validation on the complete OC video sequences including CRADS zero score [46] will be studied in the future.

Acknowledgments The authors thank the Walter Reed Army Medical Center at National Cancer Institute for providing optical colonoscopy videos and corresponding CT scans. These data were used for virtual colonoscopy study, and all subjects gave their informed consent prior to their inclusion in the study.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. NCI: colon and rectal cancer (2010) National Cancer Institute. http://www.cancer.gov/cancertopics/types/colon-and-rectal
2. Baxter N, Rabeneck L (2010) Is the effectiveness of colonoscopy "good enough" for population-based screening? J Natl Cancer Inst 102:70–71
3. Summers RM, Swift JA, Dwyer AJ, Choi JR, Pickhardt PJ (2009) Normalized distance along the colon centerline: a method for correlating polyp location on ct colonography and optical colonoscopy. AJR Am J Roentgenol 193(1):1296–1304
4. Duncan JE, McNally MP, Sweeney WB, Gentry AB, Barlow DS, Jensen DW, Cash BD (2009) Ct colonography predictably overestimates colonic length and distance to polyps compared with optical colonoscopy. AJR Am J Roentgenol 193(5):1291–1295
5. Liu J (2011) From pixel to region to temporal volume: a robust motion processing framework for visually-guided navigation. Ph.D. thesis, University of North Carolina at Charlotte
6. Hwang S, Oh J, Lee J, Tavanapong W, de Groen PC, Wong J (2007) Informative frame classification for endoscopy video. Med Image Anal 11–2:110–127
7. Oh J, Hwang S, Cao Y, Tavanapong W, Liu D, Wong J, de Groen P (2009) Measuring objective quality of colonoscopy. IEEE Trans Biomed Eng 56:2190–2196
8. Mori K, Deguchi D, Akiyama K, Kitasaka T, Maurer CR Jr, Suenaga Y, Takabatake H, Mori M, Natori H (2005) Hybrid bronchoscope tracking using a magnetic tracking sensor and image registration. In: Proceedings of 8th MICCAI, pp 543–555
9. Deligianni F, Chung A, Yang GZ (2006) Non-rigid 2d–3d registration with catheter tip em tracking for patient specific bronchoscope simulation. In: Proceedings of 9th MICCAI, pp 281–288
10. Rai L, Helferty J, Higgins W (2008) Combined video tracking and image-video registration for continuous bronchoscopic guidance. Int J Comput Assist Radiol Surg 3(3–4):315–329
11. Mori K, Deguchi D, Sugiyama J, Suenaga Y, Toriwaki J Jr, Maurer CM, Takabatake H, Natori H (2002) Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images. Med Image Anal 6(3):321–336
12. Bricault I, Ferretti G, Cinquin P (1998) Multi-level strategy for computer-assisted transbronchial biopsy. In: Proceedings of 1th MICCAI, pp 161–268
13. Helferty JP, Sherbondy AJ, Kiraly AP, Higgins WE (2005) System for live virtual-endoscopic guidance of bronchoscopy. In: Proceedings of IEEE CVPR, p 68
14. Helferty JP, Higgins WE (2002) Combined endoscopic video tracking and virtual 3d ct registration for surgical guidance. In: Proceedings of IEEE ICIP, pp 961–964
15. Rai L, Merritt SA, Higgins WE (2006) Real-time image-based guidance method for lung-cancer assessment. In: Proceedings of IEEE CVPR, pp 2437–2444
16. Deguchi D, Mori K, Suenaga Y, Hasegawa J, Toriwaki J, Batake HT, Natori H (2003) New image similarity measure for bronchoscope tracking based on image registration. In: Proceedings of 6th MICCAI, pp 399–406
17. Nagao J, Mori K, Enjouji T, Deguchi D (2004) Fast and accurate bronchoscope tracking using image registration and motion prediction. In: Proceedings of 7th MICCAI, pp 551–558
18. Higgins WE, Helferty JP, Lu K, Merritt SA, Rai L, Yu KC (2007) 3d ct-video fusion for image-guided bronchoscopy. Comput Med Imaging Graph 32:159–173
19. Helferty JP, Sherbondy AJ, Kiraly AP, Higgins WE (2007) Computer-based system for the virtual-endoscopic guidance of bronchoscopy. Comput Vis Image Underst 108(1–2):171–187
20. Deligianni F, Chung A, Yang GZ (2004) Patient-specific bronchoscope simulation with pq-space-based 2d/3d registration. Comput Aided Surg 9(5):215–226
21. Deligianni F, Chung A, Yang GZ (2006) Non-rigid 2d/3d registration for patient specific bronchoscopy simulation with statistical shape modelling. IEEE Trans Med Imaging 25(11):1462–1471
22. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
23. Tuytelaars T, Gool LV (2004) Matching widely separated views based on affine invariant regions. Int J Comput Vis 59(1):61–85
24. Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of the British machine vision conference, pp 384–393
25. Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. Int J Comput Vis 60(1):63–86
26. Liu R, Li Z, Jia J (2008) Image partial blur detection and classification. In: Proceedings of the IEEE CVPR (2008) June 27–28. Anchorage, Alaska
27. Brox T, Bregler C, Malik J (2009) Large displacement optical flow. In: Proceedings of the IEEE CVPR, pp 41–48
28. Brox T, Bruhn A, Papenberg N, Weickert J (2004) High accuracy optical flow estimation based on a theory for warping. In: Proceedings of 8th ECCV, vol 4, pp 25–36
29. Brox T, Malik J (2010) Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans Pattern Anal Mach Intell 33(3):500–513
30. Lindeberg T (1993) Scale-space theory in computer vision, 1st edn. Springer, Berlin
31. Young DM (1971) Iterative solution of large linear systems (Computer science and applied mathematics), 1st edn. Academic Press, London
32. Tuytelaars T, Mikolajczyk K (2008) Local invariant feature detectors: a survey, 1st edn. Now Publishers Inc, Hanover
33. Kadir T, Zisserman A, Brady M (2004) An affine invariant salient region detector. In: Proceedings of the European conference on computer vision, pp 404–416

34. Horn B, Schunck B (1981) Determining optical flow. Artif Intell 17(3):185–203
35. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of international joint conference on artificial intelligence, pp 281–288
36. Papenberg N, Bruhn A, Brox T, Didas S, Weickert J (2006) Highly accurate optic flow computation with theoretically justified warping. Int J Comput Vis 67(2):141–158
37. Ryan TW (1981) The prediction of cross-correlation accuracy in digital stereo-pair images. Ph.D. thesis, University of Arizona
38. Li SZ (2001) Markov random field modeling in image analysis, 2nd edn. Springer, Berlin
39. Felzenszwalb PF, Huttenlocher DP (2006) Efficient belief propagation for early vision. Int J Comput Vis 70(1):41–54
40. Liu C, Yuen J, Torralba A (2011) Sift flow: dense correspondence across different scenes and its applications. IEEE Trans Pattern Anal Mach Intell 33(5):978–994
41. Bruss AR, Horn BKP (1983) Passive navigation. Comput Vis Graph Image Process 21:3–20
42. Reiger J, Lawton D (1985) Processing differential image motion. J Opt Soc Am A 2(2):354–359
43. Heeger D, Jepson A (1992) Subspace methods for recovering rigid motion 1: algorithm and implementation. Int J Comput Vis 7(2):95–117
44. Lim J, Barnes N (2009) Estimation of the epipole using optical flow at antipodal points. Comput Vis Image Underst 114(2):245–253
45. LEGO-Group T (2010) Company profile: an introduction to the lego group. http://www.lego.com
46. Zalis ME, Barish MA, Choi JR, Dachman AH, Fenlon HM, Ferrucci JT, Glick SN, Laghi A, Macari M, McFarland EG, Morrin MM, Pickhardt PJ, Soto J, Yee J (2005) Ct colonography reporting and data system: a consensus proposal. Radiology 236:3–9
47. Weickert J, ter Haar Romeny BM, Viergever MA (1998) Efficient and reliable schemes for nonlinear diffusion filtering. IEEE Trans Image Process 7:398–410
48. Luo J, Konofagou E (2010) A fast normalized cross-correlation calculation method for motion estimation. IEEE Trans Ultrason Ferroelectr Freq Control 57:1347–1357
49. Nister D, Naroditsky O, Bergen J (2004) Visual odometry. In: Proceedings of IEEE CVPR, pp 652–659