Dense Feature Correspondence for Video-Based Endoscope Three-Dimensional Motion Tracking

Ying Wan*, Qiang Wu, and Xiangjian He

Abstract— **This paper presents an improved video-based endoscope tracking approach on the basis of dense feature correspondence. Currently video-based methods often fail to track the endoscope motion due to low-quality endoscopic video images. To address such failure, we use image texture information to boost the tracking performance. A local image descriptor – DAISY is introduced to efficiently detect dense texture or feature information from endoscopic images. After these dense feature correspondence, we compute relative motion parameters between the previous and current endoscopic images in terms of epipolar geometric analysis. By initializing with the relative motion information, we perform 2-D/3-D or video-volume registration and determine the current endoscope pose information with six degrees of freedom (6DoF) position and orientation parameters. We evaluate our method on clinical datasets. Experimental results demonstrate that our proposed method outperforms state-of-the-art approaches. The tracking error was significantly reduced from 7.77 mm to 4.78 mm.**

I. Introduction

Navigated endoscopy (NE) seeks to assist physicians to successfully perform endoscopic interventions (e.g., needle biopsies). Compared to conventional endoscopy giving only 2-D endoscopic video information, NE provides an augmented reality environment on the basis of pre- and intra-operative images to find where an endoscope is currently observing inside body cavities. Endoscope 3-D motion tracking plays a key role in endoscope navigation. It aims to register live endoscopic video sequences to pre-operative volume data, e.g., computed tomography (CT) slices, to locate the endoscope tip in a pre-operative volume coordinate system. Hence, physicians can exactly obtain endoscope position and orientation and perform tumor resection or biopsies successfully. Current tracking approaches consist of two main groups of vision- and sensor-based methods [1], [2].

Video-based endoscope tracking, which is a very active topic in the field of computer assisted interventions, is also the topic of this paper. It usually defines a similarity function between endoscopic video sequences and virtual rendering images generated from pre-operative volume data and performs video-volume registration to maximize the similarity between them [1], [3], [4]. Without using video image texture information, sole video-volume registration approaches have been discussed [1], [3], [4]. To enhance sole registration methods, scale-invariant feature transform (SIFT) has been introduced [2]. Although current video-based approaches work well, more accurate and robust tracking methods are still expected to correctly navigate the endoscope.

This work is motivated by an effective image local descriptor – DAISY which not only computes image features densely, but also was demonstrated to be more robust than SIFT [5]. We use DAISY to detect image dense features. Based on dense features, inter-frame motion information can be computed by camera epipolar geometry. Using inter-frame motion information, we perform video-volume registration to determine continuous endoscope poses. The main contribute of this work not only proposes an improved video-based tracking method but also extend the application of DAISY algorithm to the field of computer assisted interventions.

II. Approaches

Our tracking approach includes three main stages: (1) DAISY feature detection, (2) epipolar geometric analysis, and (3) video-volume registration, as described as follows.

A. DAISY Feature Detection

For each pixel (x, y) in an input image, N orientation maps, $\mathbf{M}_{i,o}, 1 \leq i \leq N$, o indicates the direction of the derivative, can be calculated by the gradient information. We convolve each orientation map with Gaussian kernels Σ_σ of different σ (used to adjust the region size) and obtain convolved orientation map $\mathbf{M}_{i,o}^{\sigma}$. To reduce the computational time, we calculate $\mathbf{M}_{i,o}^{\sigma_2}$ with respect to $\mathbf{M}_{i,o}^{\sigma_1}$ and Σ_σ [5]:

$$\mathbf{M}_{i,o}^{\sigma_2} = \Sigma_\sigma * \mathbf{M}_{i,o}^{\sigma_1}, \quad \sigma = \sqrt{\sigma_2^2 - \sigma_1^2}. \tag{1}$$

After convolution, all the orientation maps at pixel (x, y) can be represented by vector $\mathbf{V}_\sigma(x, y)$ [5]:

$$\mathbf{V}_\sigma(x, y) = \left[\mathbf{M}_{1,o}^{\sigma}, \cdots, \mathbf{M}_{N,o}^{\sigma}\right]^T. \tag{2}$$

Finally, DAISY descriptor $\mathcal{D}(x, y)$ at pixel (x, y) can be

Ying Wan is with School of Computing and Communications, University of Technology, Sydney, Broadway NSW 2007 Australia Ying.Wan@student.uts.edu.au

Qiang Wu is with School of Computing and Communications, University of Technology, Sydney, Broadway NSW 2007 Australia Qiang.Wu@uts.edu.au

Xiangjian He is with School of Computing and Communications, University of Technology, Sydney, Broadway NSW 2007 Australia Xiangjian.He@uts.edu.au

formulated as the concatenation of vector $\mathbf{V}_{\sigma}(x, y)$ [5]:

$$\mathcal{D}(x, y) = \begin{bmatrix} \mathbf{V}_{\sigma_1}^T(x, y) \\ \mathbf{V}_{\sigma_1}^T(\mathbf{p}_1(x, y, L_1)), \cdots, \mathbf{V}_{\sigma_1}^T(\mathbf{p}_j(x, y, L_1)) \\ \vdots \\ \mathbf{V}_{\sigma_q}^T(\mathbf{p}_1(x, y, L_Q)), \cdots, \mathbf{V}_{\sigma_q}^T(\mathbf{p}_j(x, y, L_q)) \end{bmatrix}^T, \tag{3}$$

where $\mathbf{p}_j(x, y, L_q))$ is the pixel with distance L_q from pixel (x, y) along direction j to region q whose radius depends on the standard deviations of the Gaussian kernels.

B. Epipolar Geometric Analysis

After computing DAISY descriptor $\mathcal{D}(x, y)$ for each pixel on $k-1$ and k (previous and current) video images, we match these descriptors to get matched pair $(\mathbf{p}_{k-1}^a, \mathbf{p}_k^i)$ [2], point index a, and compute fundamental matrix $\mathbf{F}$ by:

$$(\mathbf{p}_k^a)^T \mathbf{F} \mathbf{p}_{k-1}^a = 0, \tag{4}$$

which can be solved given enough corresponding points [6].

Combining $\mathbf{F}$ with intrinsic endoscopic camera matrix $\mathbf{Q}$, essential matrix $\mathbf{E}$ can be determined by: $\mathbf{E} = \mathbf{Q}^T \mathbf{F} \mathbf{Q}$. Finally, we obtain inter-frame motion matrix $\Delta\tilde{\mathbf{T}}_k$ including translation unit vector $\Delta\tilde{\mathbf{t}}_k = (\Delta\tilde{t}_k^x, \Delta\tilde{t}_k^y, \Delta\tilde{t}_k^z)$ and rotation matrix $\Delta\tilde{\mathbf{R}}_k$ between the previous and current video images by sequentially solving the following two equations:

$$\mathbf{E}^T \Delta\tilde{\mathbf{t}}_k = \mathbf{0}, \tag{5}$$

$$\Delta\tilde{\mathbf{R}}_k \mathbf{E}^T = \left[\Delta\tilde{\mathbf{t}}_k\right]_\times^T = \begin{bmatrix} 0 & -\Delta\tilde{t}_k^z & \Delta\tilde{t}_k^y \\ \Delta\tilde{t}_k^z & 0 & -\Delta\tilde{t}_k^x \\ -\Delta\tilde{t}_k^y & \Delta\tilde{t}_k^x & 0 \end{bmatrix}^T. \tag{6}$$

Note that essential matrix $\mathbf{E}$ is determined only up to an arbitrary scale factor. We empirically determine such a factor.

C. Video-Volume Registration

After obtaining relative motion parameters $\Delta\tilde{\mathbf{T}}_k$, we perform video-volume registration to determine current endoscope pose $\mathbf{T}_k$ with 6DoF position and orientation at frame k. We utilize a modified mean squared error similarity measure ($MoMSE$) to characterize the similarity between endoscopic sequence $\mathbf{I}_k$ and virtual rendering image $\mathbf{I}_V$ [1].

Let $\mathbf{I}_V(\mathbf{T}_k)$ be a virtual rendering image that is generated using rendering parameters $\mathbf{T}_k = \mathbf{T}_{k-1}\Delta\mathbf{T}_k$. We optimize relative motion matrix $\Delta\mathbf{T}_k$ to find the most similar virtual rendering image $\mathbf{I}_V(\mathbf{T}_{k-1}\Delta\mathbf{T}_k)$ corresponding to video image $\mathbf{I}_k$. The process of video-volume registration involved with $\Delta\mathbf{T}_k$ can be formulated as the following optimization:

$$\Delta\mathbf{T}_k^* = \arg\max_{\Delta\mathbf{T}_k} MoMSE(\mathbf{I}_k, \mathbf{I}_V(\mathbf{T}_{k-1}\Delta\mathbf{T}_k)). \tag{7}$$

The maximization process is implemented on the basis of the Powell optimization method [7]. During such an optimization process, its initialization of $\Delta\mathbf{T}_k$ plays an important role, possibly increasing the tracking performance. Setting $\Delta\mathbf{T}_k$ as an identity matrix might be an easy way [1]. However, without using image texture information, the optimization process easily gets trapped in local minima, in turn, resulting in the tracking failure. In this study, we use relative motion matrix $\Delta\tilde{\mathbf{T}}_k$ to initialize $\Delta\mathbf{T}_k$. $\Delta\tilde{\mathbf{T}}_k$ was estimated using dense feature correspondence. Compared to previous registration methods, we introduce DAISY features since it is more dense than SIFT. These features are robust to image motion blurring and illumination changes which usually happen in endoscopic videos. Hence our method can improve the performance of video-based endoscope 3-D motion tracking. This is the major point of this work.

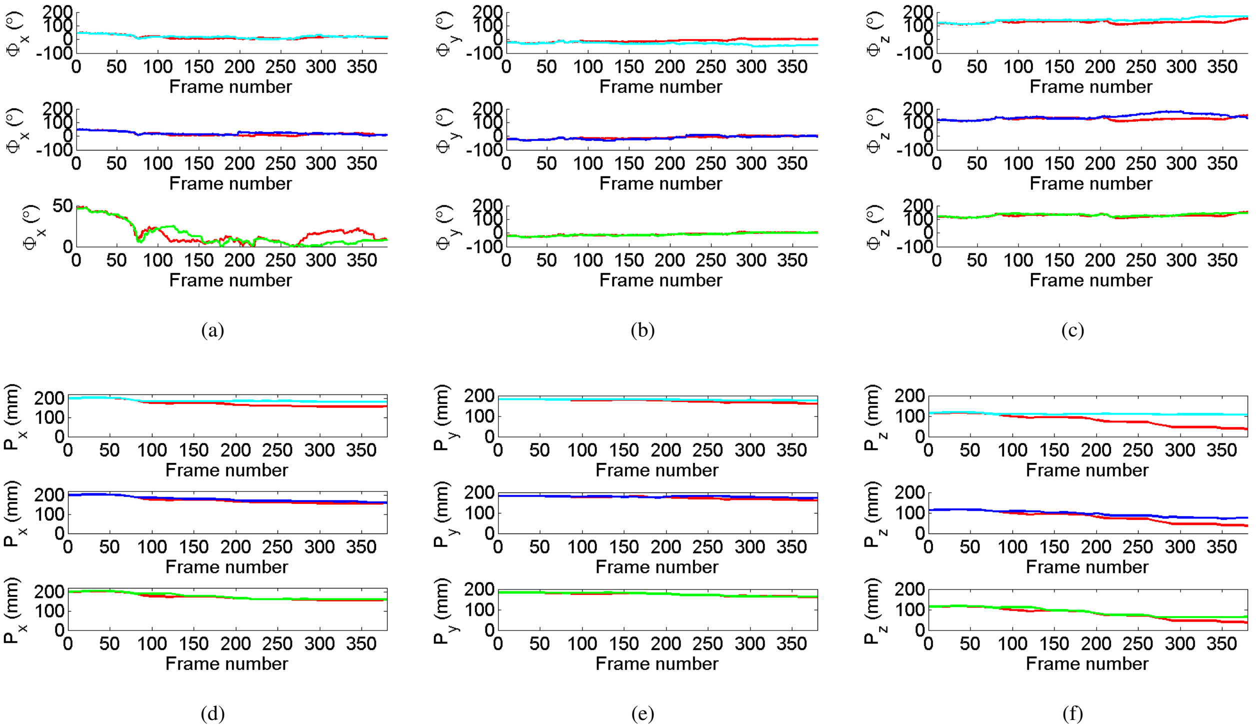
Algorithm 1: DAISY-based endoscope motion tracking

```
Input: CT images and endoscopic video sequences;
Output: A series of motion estimation $\{\mathbf{T}_k\}$;
❶ Initialize motion parameters $\mathbf{T}_1$ at frame $k = 1$;
❷ Detect and store DAISY features at frame $k = 1$;
for $k = 2$ to $K$ ($K$: Total video frames) do
    ❸ DAISY feature detection at frame $k$ (Eqs.1∼3);
      Store DAISY feature at frame $k$;
      Matching DAISY features at frames $k-1$ and $k$;
    ❹ Epipolar geometric analysis (Eqs.4∼6);
      Obtain relative motion $\Delta\tilde{\mathbf{T}}_k$;
    ❺ Volume-video registration (Eq.7);
      Generate virtual rendering image $\mathbf{I}_V(\mathbf{T}_{k-1}\Delta\mathbf{T}_k)$;
      Optimize $\Delta\mathbf{T}_k$ and obtain optimal $\Delta\mathbf{T}_k^*$;
    ❻ Determine current estimate: $\mathbf{T}_k = \Delta\mathbf{T}_k^*\mathbf{T}_{k-1}$;
    ❼ Store $\mathbf{T}_k$ and go to the next iteration $k = k + 1$;
end
return Motion estimation $\{\mathbf{T}_k\}_{k=1}^K$;
```

TABLE I: Quantitative tracking errors of position and orientation estimated by different approaches (unit: mm, degree)

Cases	Deguchi et al. [1]		Luo et al. [2]		our method	
A	32.8	31.9	16.9	29.3	10.3	20.9
B	6.93	23.4	5.07	16.5	3.30	10.4
C	4.34	10.1	4.07	9.29	2.51	7.80
D	15.3	45.6	7.75	24.1	4.82	15.6
E	13.8	23.8	5.06	17.1	2.95	10.8
Average	14.6	27.0.	7.77	19.3	4.78	13.1

Our proposed method using DAISY features for endoscope 3-D motion tracking is summarized in **Algorithm 1**.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 3: Compare estimated 6DoF position and orientation parameters in $x-$, $y-$, and $z-$directions to ground truth (*red* line) on Case A. As we can see, *green* line that shows the position and orientation estimated by our propose method was more overlapping on *red* line than *cyan* and *blue* lines indicated estimated results from other methods. This means that the tracking performance of our proposed method outperforms the other two methods of Deguchi et al. [1] and Luo et al. [2].

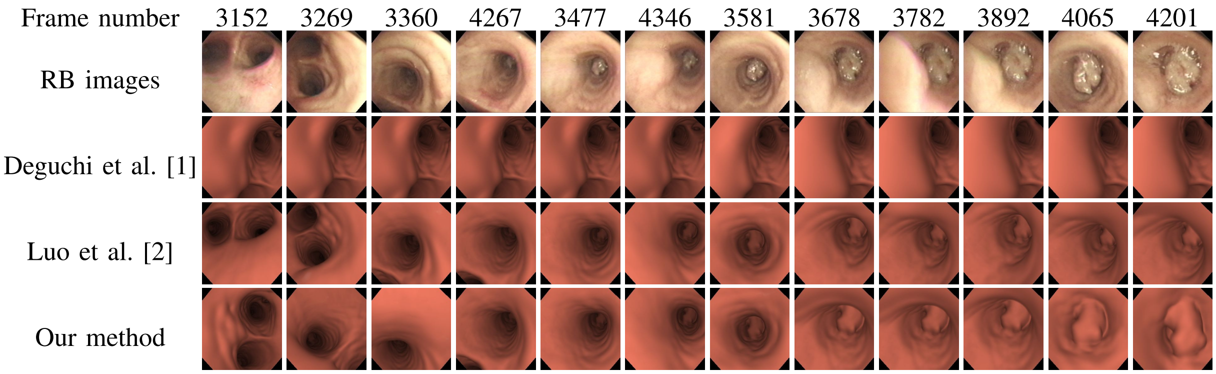


Fig. 4: Another example of visual comparison on Case E.

III. Results

We evaluated our method on five cases of clinical datasets. Each dataset includes a volume of CT images and its corresponding endoscopic video sequences. We manually generated ground truth for these datasets. We can compute position and orientation errors, κ, ϕ by: $\kappa = \|\mathbf{t} - \mathbf{t}_G\|$, $\phi = \arccos((trace(\mathbf{R}\mathbf{R}_G^T) - 1)/2)$, where $\mathbf{t}$ and $\mathbf{R}$ are estimated position and orientation, $\mathbf{t}_G$ and $\mathbf{R}_G$ are ground truth. We investigate three tracking approaches: (1) a sole video-volume registration method [1], (2) a SIFT-driven framework [2], and (3) our method, as discussed in Section II.

Figs. 1 and 2 displays an example of tracking errors using different methods evaluated on Case A. Fig. 3 further compares the estimated position and orientation to ground truth. These figures prove that our method outperforms others. Table I quantifies tracking errors of the three methods. The position error was greatly reduced from 7.77 mm to 4.78 mm and orientation error was also improved from 19.3° to 13.1°. Moreover, we visually inspect the tracking results. We

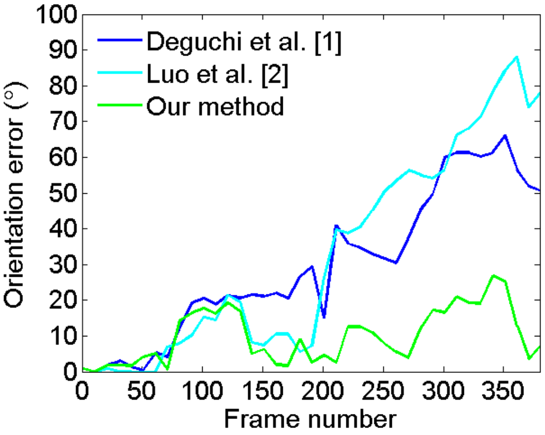


Fig. 1: Orientation error of Case A

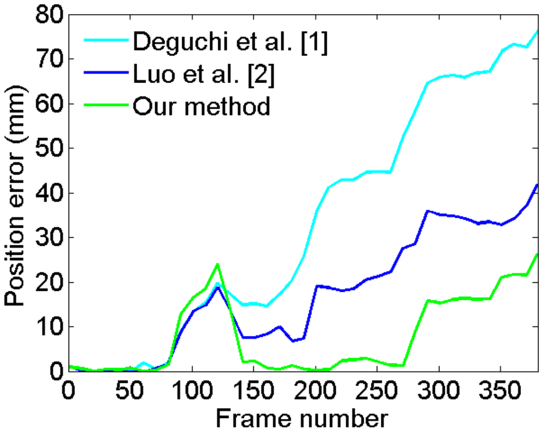


Fig. 2: Position error of Case A

used all the estimated position and orientation parameters to generate virtual rendering images. We manually visualize whether endoscopic video images resemble to virtual rendering image. The more similar of these images, the more robust performance of tracking methods. Fig 4 illustrates the visual comparison of the tracking results of different methods. They further prove that our method is more accurate and robust than other two approaches.

IV. Discussion

This work aims to improve the performance of video-based endoscope motion tracking methods. Based on clinical evaluation, our proposed method provides more accurate and robust tracking than previous methods. We contribute such an improvement to the introduction of dense feature correspondence. Since SIFT only provides a number of key points that might be insufficient, unstable, or easily disappeared in continuous video images, we introduced DAISY, a dense computation of descriptors, which can tackle the insufficiency of point features, since DAISY has the ability to update unstable or disappeared points by new detected points. Hence our method outperforms the SIFT-based approach [2].

Although our proposed method significantly improves the tracking accuracy compared to previous approaches, it still possibly fails to continuously track the endoscope motion. Beyond image dense texture information that is very useful for boosting endoscope motion tracking, a similarity function also plays an important role in video-volume registration. A robust similarity measure should be able to accurately characterize the difference between endoscopic video and virtual rendering images, even if low-quality video images, e.g., motion blurring, inter-reflection, or illumination changes, appear frequently in endoscopic video sequences. The mean squared error-based similarity measure that was used in this work sometimes can not successfully distinguish the difference under low-quality endoscopic video images. Inaccurately computing the similarity easily collapses the optimization trapped into local minima. One of our future work is to explore a new similarity function, which not only can adapt itself to image illumination changes, but also can accurately characterize the difference of video and virtual images for improving the performance of video-volume registration. Additionally, we clarify that our method currently can not track the endoscope motion in real time.

V. Conclusions

This paper proposed an improved video-based endoscope 3-D motion tracking on the basis of dense feature correspondence using the DAISY descriptor algorithm. Since DAISY features are more dense than SIFT, image texture information can be more stably or sufficiently used in computing relative motion parameters that can boost the tracking performance. Compared to previous methods, the tracking position and orientation errors was reduced from at least (7.77 mm, 19.3°) to (4.78 mm, 13.1°). The future work includes development of a new similarity measure for video-volume registration to improve video-based endoscope 3-D motion tracking, reduction of processing time, and more clinical validation.

References

[1] Deguchi, D., et al.: Selective image similarity measure for bronchoscope tracking based on image registration. Medical Image Analysis 13(4), 621–33 (2009)

[2] Luo, X., et al.: Development and comparison of new hybrid motion tracking for bronchoscopic navigation. Medical Image Analysis 16(3), 577–596 (2012)

[3] Deligianni, F., et al.: Nonrigid 2-D/3-D registration for patient specific bronchoscopy simulation with statistical shape modeling: Phantom validation. IEEE TMI 25(11), 1462–1471 (2006)

[4] Helferty, J. P., et al.: Computer-based system for the virtual-endoscopic guidance of bronchoscopy. CVIU 108(1-2), 171–187 (2007)

[5] Tola, E., et al.: DAISY: An efficient dense descriptor applied to wide-baseline stereo. IEEE TPAMI 32(5), 815–830 (2010)

[6] Hartley, R., et al.: Multiple View Geometry in Computer Vision. Cambridge University Press, (2004)

[7] Berghen, F. V., et al.: CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: experimental results and comparison with the DFO algorithm. JCAM 181(1), 157–175 (2005)