Journal of Medical Robotics Research, (2022) 2241002 (16 pages)

DOI: 10.1142/S2424905X22410021

Journal of Medical Robotics Research
https://www.worldscientific.com/worldscinet/jmrr

Journal of Medical Robotics Research

Real-Time Camera Localization during Robot-Assisted Telecystoscopy for Bladder Cancer Surveillance

Chen Gong*, Yaxuan Zhou†, Andrew Lewis*, Pengcheng Chen*, Jason R. Speich‡, Michael P. Porter§, Blake Hannaford†, Eric J. Seibel*

*Mechanical Engineering, University of Washington, 3900 E Stevens Way NE, Seattle, WA 98195, USA

†Electrical and Computer Engineering, University of Washington
185 W Stevens Way NE, Seattle, WA 98195, USA

‡Center for Research and Education in Simulation Technologies (CREST)
University of Washington, 1959 NE Pacific St., Seattle, WA 98195, USA

§Department of Urology, University of Washington, 1959 NE Pacific St., Seattle, WA 98195, USA

Telecystoscopy can lower the barrier to access critical urologic diagnostics for patients around the world. A major challenge for robotic control of flexible cystoscopes and intuitive teleoperation is the pose estimation of the scope tip. We propose a novel real-time camera localization method using video recordings from a prior cystoscopy and 3D bladder reconstruction to estimate cystoscope pose within the bladder during follow-up telecystoscopy. We map prior video frames into a low-dimensional space as a dictionary so that a new image can be likewise mapped to efficiently retrieve its nearest neighbor among the dictionary images. The cystoscope pose is then estimated by the correspondence among the new image, its nearest dictionary image, and the prior model from 3D reconstruction. We demonstrate performance of our methods using bladder phantoms with varying fidelity and a servo-controlled cystoscope to simulate the use case of bladder surveillance through telecystoscopy. The servo-controlled cystoscope with 3 degrees of freedom (angulation, roll, and insertion axes) was developed for collecting cystoscope videos from bladder phantoms. Cystoscope videos were acquired in a 2.5D bladder phantom (bladder-shape cross-section plus height) with a panorama of a urothelium attached to the inner surface. Scans of the 2.5D phantom were performed in separate arc trajectories each of which is generated by actuation on the angulation with a fixed roll and insertion length. We further included variance in moving speed, imaging distance and existence of bladder tumors. Cystoscope videos were also acquired in a water-filled 3D silicone bladder phantom with hand-painted vasculature. Scans of the 3D phantom were performed in separate circle trajectories each of which is generated by actuation on the roll axis under a fixed angulation and insertion length. These videos were used to create 3D reconstructions, dictionary sets, and test data sets for evaluating the computational efficiency and accuracy of our proposed method in comparison with a method based on global Scale-Invariant Feature Transform (SIFT) features, named SIFT-only. Our method can retrieve the nearest dictionary image for 94–100% of test frames in under 55 ms per image, whereas the SIFT-only method can only find the image match for 56–100% of test frames in 6000–40000 ms per image depending on size of the dictionary set and richness of SIFT features in the images. Our method, with a speed of around 20 Hz for the retrieval stage, is a promising tool for real-time image-based scope localization in robotic cystoscopy when prior cystoscopy images are available.

Keywords: Telecystoscopy; camera re-localization; 3D reconstruction; image retrieval; telemedicine.

JMRR

Received 15 December 2021; Revised 16 March 2022; Accepted 30 March 2022; Published xx xx xx . This paper was recommended for publication in its revised form by editorial board member, NAME.
Email Address: *eseibel@uw.edu*

1. Introduction

Flexible cystoscopy is an important diagnostic procedure performed by urologists in-office for procedures such as evaluating blood in urine (hematuria), removing stents after kidney stone surgery, and investigating urethral strictures [1]. A diagnostic flexible cystoscopy usually begins with these steps: (1) insertion of the cystoscope into the urethra, (2) inflation and flushing of bladder with clear, sterile fluid pressurized through the working channel (throughout the procedure), (3) inspection of urethral wall during scope insertion, (4) insertion through bladder sphincters, (5) identification of common landmark (usually left or right ureteral orifice), and (6) inspection scan of the entire urothelium (bladder surface) with detailed inspection of areas of interest. Flexible cystoscopy is the gold standard for diagnosis and surveillance of bladder cancer, the 6th most common and the most costly cancer in the US [2,3]. Bladder cancer has a recurrence rate of over 50% [4], which requires that patients return to their urologists for follow-up cystoscopies up to 4 times per year for surveillance after initial treatment [5], and a delay in diagnosis of muscle-invasive tumors of 3–6 months can increase risk of death by bladder cancer by 34% [6]. Nearly 90% of urologists in the US practice in metropolitan areas [7], which can burden some patients with travel costs and time off work [3]. Bladder cancer patients in rural and underserved areas would benefit from a telerobotic cystoscopy system placed in geographically distributed clinics or urgent care facilities, set up and overseen by nurses, and operated by urologists located in their own office. Such a telemedicine system would be useful for many diagnostic urologic procedures, but would be especially useful for bladder cancer patients who require frequent in-person visits for cancer surveillance.

Although this vision of telecystoscopy is not yet in practice, the technologies required have already been demonstrated: the first transcontinental telesurgery was successfully completed two decades ago [8], telerobotic flexible endoscopes are being introduced commercially for use with surgeons in the room [9,10], and researchers are developing transurethral surgery robots [11–13]. Introducing teleoperation for bladder inspection is logical because the organ is pliable and not close to critical life-sustaining functions and nurses are well experienced with insertion of urinary catheters. Widespread adoption of clinic-based telecystoscopy will likely begin with a telerobotic platform that can interface with off the shelf, and perhaps single-use cystoscopes [14] which reduces infrastructure overhead. Thus, flexible cystoscopy may serve well as a test case for long-distance teleoperation by urologists in major cities and patients in clinics with nursing and general practitioner support, reducing barriers to timely specialty care.

A major challenge within the teleoperation interface is the accurate pose estimation of the cystoscope within the bladder, since the haptics and proprioception that urologists rely upon for localization will be difficult to simulate in an economical way. Teleoperation of clinical catheter robots has been shown to be improved with the integration of tip-tracking and shape estimation with preoperative 3D anatomical models [15]. Thus, a key feature for developing a telecystoscopy system is the ability to estimate the position and orientation of the cystoscope tip in order to display the pose within a patient-specific model of the bladder and highlight the current Field Of View (FOV) for the urologist during teleoperation (Fig. 1). However, the kinematics of flexible endoscopes can vary widely even between endoscopes of the same make [16] with different amounts of use, and are also dependent on the curvature of the main scope body [17], making accurate forward kinematics estimation of clinical endoscopes difficult without a detailed characterization for each endoscope. Magnetic field- and electromagnetic wave-based localization strategies are

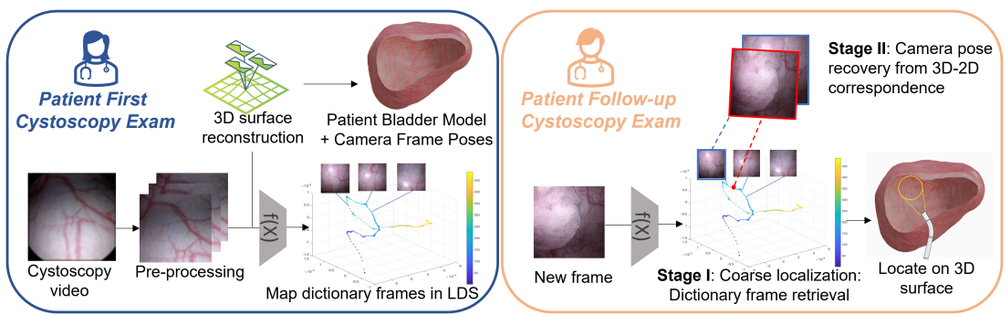

Fig. 1. Process of our localization system for telecystoscopy. (Left): Video from the 1st exam is used to create a 3D bladder model and used image frames are mapped onto a Low-Dimensional Space (LDS) as a dictionary set. (Right): During the 2nd exam, each new image frame is mapped into the same space and its closest neighbor is retrieved from the dictionary (Stage I). Then 3D-2D correspondences among the new image, its retrieved dictionary image, and the 3D reconstructed model are used to recover camera pose associated with the new image (Stage II). The video frame can then be highlighted on the 3D surface and the estimated cystoscope pose can be used for downstream tasks.

widely used in robotic flexible endoscopy [18–20], but these methods require extra sensors, specialized hardware, and sensitive calibration. The cost and operational complexity associated with precise endoscope calibration or additional sensing modalities may be disadvantageous to the adoption of a widely distributed telediagnostic platform. On the other hand, an image-based scope localization approach during teleoperation would not only provide the urologist with a feedback of scope pose within the bladder, it could also ensure thorough examination by calculating a running bladder surface coverage metric, providing positions of areas of interest during the current or subsequent procedure, and enabling stabilization around an area of interest [21].

A standard, image-based approach for camera localization is Simultaneous Localization And Mapping (SLAM). Visual SLAM is common in robotics and utilizes images from monocular, stereo, or RGB + Depth cameras to simultaneously localize robot position and reconstruct the surrounding scene in real time [22]. Visual SLAM has been used primarily in rigid laparoscopic surgery [23–25] and flexible endoscopies [26]. However, the feature detection algorithm and sequential frame matching design in the existing SLAM pipeline does not perform well in many areas of the body due to a lack of texture [24]. Blood vessels on the inner surface of the bladder are a major source of feature points in cystoscope frames, but they are sparse. Structure from Motion (SfM) achieves offline 3D reconstruction through feature detection and matching, triangulation and global optimization of reconstructed 3D points and estimated camera poses, with emphasis on robustness and accuracy, but sacrifices speed. Thus, prior studies used SfM for post-procedure bladder reconstruction [27–32]. SIFT is most generally used in SfM because of the high accuracy for feature point extraction and matching [33], while the computation of SIFT features in SfM is time consuming. Speeded Up Robust Features (SURF) was developed to further reduce the computation load involved in SIFT and provides similar performance at faster speed ($3\times$) through the use of integral images [34]. SURF is primarily applied when high-speed matching is required [35–37], but does not work well under scale or rotation changes, thus, inferior to SIFT for this application. On the other hand, SIFT has limited success with medical images because sparse features and homogeneous backgrounds provide significantly less information for global feature point matching. Low image quality, small FOV, and motion blur in cystoscopy can further increase the difficulty of feature point matching. Accordingly, there will be a high-quality requirement of the captured videos for SIFT-based mapping and localization.

In this work, we propose a two-stage global camera localization method for robot-assisted flexible cystoscopy when a video from a previous procedure is available. Recordings of previous procedures could be available for half a million bladder cancer survivors under routine surveillance cystoscopy, which represents a subset of the 724,000 prevalent cases of bladder cancer in the US [38]. For bladder cancer patients, the initial cystoscope video from their first cystoscopy exam during screening will be available. Our method utilizes the initial video for generation of a 3D bladder model and a dictionary set composed of frames with calculated camera poses in an off-line manner. Then during follow-up exams for surgery or monitoring, our method can estimate camera pose for a new image by first retrieving a prior image with known camera pose and large overlap with the new image frame for coarse localization, and then recovering camera pose from the correspondence information for fine localization in an online manner. Unlike the localization based on continuous frames, this coarse-to-fine paradigm performs a global matching, avoiding accumulated errors and the effects of occasional failures. We investigate the performance of our algorithm in localizing video frames and camera pose captured by a servo-actuated cystoscope inserted within a 2.5D bladder phantom and 3D bladder phantom. By changing the settings of scanning and phantoms, we simulate the change of the bladder condition between the first exam and the follow-up exams which may challenge our image-based localization based on a patient's previous exam. For example, we attached the artificial tumors onto the phantom to simulate the tumor progression. We also vary imaging distance to simulate different extents of bladder distention among different exams. The results showed that our algorithm is reasonably robust to these challenges as well as efficient.

2. Methods

In this section, we describe the real-time re-localization of the cystoscope camera in the bladder with a prior 3D-reconstructed model generated from the available cystoscope video acquired during screening examination, as in the case of a bladder cancer patient returning for surveillance. In the first visit (Fig. 1(Left)), the urologist collects a cystoscope video which fully covers the complete inner surface within the bladder. We first use an off-line 3D reconstruction pipeline [29] to generate a reconstructed 3D model of the bladder inner surface from the video frames. The video frames used for reconstruction are then stored as dictionary set for subsequent re-localization. In the follow-up visits (Fig. 1(Right)), we use the prior 3D-reconstructed surface model as a prior model and estimate the camera pose associated with newly-acquired frame with respect to the coordinate of the prior model.

2.1. *3D reconstruction*

The shape and texture of the urothelium surface within bladder are reconstructed off-line using cystoscope video

frames. The 3D reconstruction pipeline is composed of the following modules:

(1) *Camera calibration and image preprocessing*: Intrinsic parameters of the cystoscope camera are first calculated from frames imaging a calibration target [39]. Then bladder frames are downsampled to avoid redundancy and preprocessed with adjustment of contrast and illumination as well as correction of lens-induced distortion.
(2) *Sparse reconstruction*: An off-line SfM algorithm [40] is used to extract and match SIFT features from frames and then calculate the camera pose at each frame as well as a 3D point cloud model depicting the shape of bladder inner surface.
(3) *Mesh generation*: Poisson surface reconstruction [41] uses recovered 3D point cloud model to generate a watertight mesh model, which better represents the shape of bladder inner surface.
(4) *Texture mapping*: The mesh model surface is then mapped with texture patches cropped from preprocessed frames to generate a textured mesh model [42], which captures both shape and texture of the bladder inner surface. Thus, the output of the 3D reconstruction includes a textured mesh model that can be used as prior 3D model for the bladder and a dictionary set composed of frames used for 3D reconstruction with their corresponding camera poses, all of which are crucial components for the subsequent camera localization step in follow-up cystoscopy visits.

2.2. *Camera localization*

Camera localization is a method for computing the camera pose associated with a camera view under a world coordinate system [43]. If we can estimate the camera pose in the coordinate system of the patient's reconstructed 3D bladder model, we can display the real-time location of camera within the model for visualization and also estimate the camera pose under any chosen world coordinate for robot actuation.

To estimate camera pose quickly and accurately, we have developed a novel two-stage camera localization pipeline (Fig. 1):

(I) *Image retrieval from dictionary set with dimension reduction*: When given a newly acquired image, we first use an efficient and accurate algorithm to retrieve the nearest dictionary image which has the largest overlap with the new image. This step is a coarse localization of the test frame. The camera pose of the retrieved dictionary frame can be directly used as a fallback solution when speed has higher priority than accuracy.
(II) *Camera pose recovery from 3D-2D correspondence*: From Sec. 2.1, we already know the correspondence between feature points on each dictionary image and the reconstructed 3D points on the prior 3D model. Thus, we can use the retrieved dictionary image as a bridge to obtain the correspondence between 3D points on the prior model and 2D SIFT features on the new image, in short, 3D-2D correspondence. Then camera pose of the new image can be calculated from the 3D-2D correspondence and represented under the 3D prior map's coordinate system.

2.2.1. *Stage I: Image retrieval from dictionary set with dimension reduction*

Sampled from continuous video frames during cystoscopy, the dictionary images used for 3D reconstruction have large overlap with their neighbors. Overlap between two images contains correspondence information useful for recovering pose of the camera views associated with the images. Thus with a dictionary set of overlapping images, one can retrieve a dictionary image that has the largest overlap with the newly acquired image for its pose localization. To perform the retrieval efficiently, we apply dimension reduction and map each dictionary frame into an LDS, where Euclidean distance between frames in the LDS indicates similarity or overlap (i.e. frames that are close to each other in a cystoscopy video are close to each other in the LDS, as shown in Fig. 1). Similar to our previous work on retinal images [44], we achieve dimension reduction by Principal Component Analysis (PCA) through Singular Value Decomposition (SVD), which is simple, versatile, and satisfies the real-time requirement for use in teleoperation, unlike other nonlinear methods such as kernel PCA [45] and Isomap [46]. Note that although PCA is known to be sensitive to outliers, occlusions, and corruption in the data, our dictionary images are acquired under expert- or robot-control and selected from the 3D reconstruction pipeline, resulting in good image quality and minimized number of outlier (bad-quality) images, thus ensuring reasonable performance of PCA.

The procedure of dictionary image retrieval is described as follows. We resize all dictionary images to vectors and form the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. The low-dimensional distribution representation of the target image distribution is obtained by implementing PCA on $\mathbf{X}$ as shown in the following equation:

$$\mathbf{Z} = \mathbf{X}\mathbf{W}, \tag{1}$$

where $\mathbf{Z} = [z_1, z_2, z_3, \ldots, z_N]^T \in \mathbb{R}^{n \times l}$, $\mathbf{W} \in \mathbb{R}^{d \times l}$ and $l \ll d$. The image space Ω_1 is mapped to a low-dimensional space Ω_2 with the mapping $\mathbf{W}$. $\mathbf{Z}$ is the low-dimensional representation. We select the top 20 principal components ($l = 20$) to represent each image in low dimension according to the dominant singular values. For

more details of the implementation and acceleration, refer to our previous work [44].

We define newly acquired frames from the follow-up cystoscopy as **T**, which are represented by the test frames in our experiments. To find the nearest dictionary image to each new frame, we use the same mapping matrix **W** to map **T** to its low-dimension representation $\mathbf{z}_T$, as shown in the following equation:

$$\mathbf{z_T} = \tilde{\mathbf{T}}\mathbf{W}, \tag{2}$$

where $\tilde{\mathbf{T}}$ is the vectorized representation of **T**. Finally, we can quickly find a representation **z** with the minimal Euclidean distance to $\mathbf{z_T}$ in the LDS, which corresponds to the dictionary image that has the largest overlap with the new frame.

2.2.2. *Stage II: Camera pose recovery from 3D-2D correspondence*

To recover the camera pose for the test frame **T**, we first extract SIFT features $\mathbf{P_T} = \{(u_T^1, v_T^1), (u_T^2, v_T^2), \ldots, (u_T^i, v_T^i), \ldots\}$ from **T**, where (u_T^i, v_T^i) denotes the pixel-level position of detected SIFT feature point on **T**. Then we can match $\mathbf{P_T}$ with the pre-extracted SIFT features $\mathbf{P_D} = \{(u_D^1, v_D^1), (u_D^2, v_D^2), \ldots, (u_D^i, v_D^i), \ldots\}$ on the retrieved dictionary image. From Sec. 2.1, we already know the correspondence between SIFT feature point (u_D^i, v_D^i) and reconstructed 3D point (x^i, y^i, z^i) in the coordinate system of the reconstructed 3D model. Now using the retrieved dictionary image as a bridge, we can get the 3D-2D correspondence between (u_T^i, v_T^i) and (x^i, y^i, z^i). Each 3D-2D correspondence pair satisfies the projection relation in Eq. (3), where s is a scale coefficient, **K** is the camera intrinsic parameter which is known from 3D reconstruction, and the rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$ and translation vector $\mathbf{t} \in \mathbb{R}^{3\times1}$ form the camera extrinsic parameter.

$$s\begin{pmatrix} u_T^i \\ v_T^i \\ 1 \end{pmatrix} = \mathbf{K}[\mathbf{R} \quad \mathbf{t}]\begin{pmatrix} x^i \\ y^i \\ z^i \\ 1 \end{pmatrix}. \tag{3}$$

We solve this equation iteratively using Random Sample Consensus (RANSAC) to find the camera extrinsic parameter $[\mathbf{R}\ \mathbf{t}]$. In each iteration, three 3D-2D correspondence pairs are sampled randomly to form an equation group based on the projection relation in Eq. (3). The solution of the equation group $[\tilde{\mathbf{R}}\ \tilde{\mathbf{t}}]$ are then used to calculate the reprojection error in the test image and count number of inliers based on a chosen threshold. The final $[\mathbf{R}\ \mathbf{t}]$ is selected from the $[\tilde{\mathbf{R}}\ \tilde{\mathbf{t}}]$ with the maximum number of inliers among all the iterations. Lastly, camera pose can be represented as follows:

$$\begin{bmatrix} \mathbf{R^T} & -\mathbf{R^T t} \\ \mathbf{0^T} & 1 \end{bmatrix} \tag{4}$$

which indicates the position and orientation of the camera in the coordinate system of the reconstructed prior 3D model.

3. Experiments

3.1. *Robotic cystoscope actuation*

We built hardware systems for actuator-controlled cystoscope movement during acquisition of videos, which has been shown to improve extraction of features on bladder phantom in our prior work [47].

3.1.1. *Hardware setup for 2.5D bladder phantom*

A linear actuator was attached to the thumb lever of a Karl Storz (Tuttlingen, Germany) HD-View Flexible Digital Cystoscope (Fig. 2(Top)) for servo-controlled angulation of scope tip with desired bending angle. The cystoscope FOV is 100°. The Actuonix (Saanichton, BC, Canada) L12-P Micro Linear Actuator servo is controlled with an Actuonix Linear Actuator Control Board via manual control via potentiometer and digital control from an Arduino Mega. The Control Board provides analog position sensor feedback from the servo. The distal end of the cystoscope is affixed to a raised platform on an independent phantom plate and the cystoscope shaft is kept straight for all experiments.

A 2.5D bladder phantom was made by 3D-printing a bladder-shaped cross-section and then taping a high-resolution, wide-FOV panorama of bladder urothelium [29] to the interior surface of the 2.5D model. The bladder contour is designed based on a bladder's sagittal cross-section (Fig. 2(Bottom Left)). This 2.5D phantom serves as a simplified test case for evaluating our localization algorithm with limited surface curvature distortions. The size of our phantom's cross-section (100×85 mm) is about 3 times larger than that of uninflated adult bladders (83×40 mm on average). The enlarged size guarantees the ideal imaging distance between scope tip and the phantom wall even when the bending angle is large, thus allowing for unconstrained angulation.

To acquire the ground truth angulation for recorded videos, we modeled kinematics for the tip angulation on our specific cystoscope. Prior research on robot-controlled endoscopes [16,17,48,49] describes flexible endoscope angulation in free space as a linear relation between thumb tip and angulation with two additional factors: hysteresis and dead-band. Hysteresis, or backlash, is when the output of a system does not change immediately as the input changes direction. Dead-band is an area around the center of thumb lever travel where angulation does not change. However, this model does

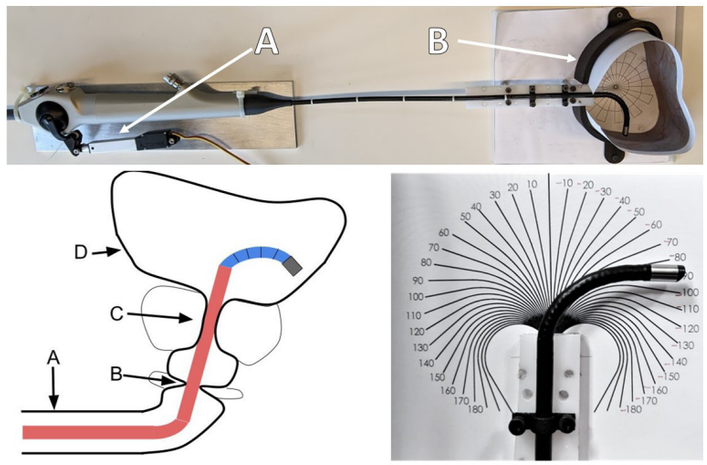

Fig. 2. (Top) 2.5D bladder phantom experiment setup: A — linear actuator for cystoscope angulation, B — 2.5D bladder phantom. The 2.5D bladder model printed model approximates surface curves that may be seen in cystoscopies. Note that the scope tip is bent with an angulation of 90° in the picture. (Bottom Left): Simplified sketch of cystocope in male anatomy: A — Urethra, B – External Urethral Sphincter, C — Verumontanum at Prostate, D — Anterior wall. The flexible cystoscope body is shown in red and controlled angulation area in blue. (Bottom Right): Cystoscope angulation measurement.

not account for curvature of the endoscope body or external contact with the endoscope.

The distal end of the cystoscope was attached below the angulation section and aligned over an angulation scale (Fig. 2(Bottom Right)). Potentiometer input to the controller was used to step the thumb lever through 3 cycles of angulation. At each step, angulation and linear sensor values were recorded. The linear actuator itself was similarly tested and was found to not exhibit hysteresis. Thumb lever angles were calculated from servo position sensor data.

3.1.2. *Hardware setup for 3D bladder phantom*

To further study the performance of our camera localization method, we expand the previous 2.5D phantom experiment setup to 3D phantom experiment setup which better simulates the scenario in clinical cystoscopy. A 3-DoF cystoscopy robot (Fig. 3(Left)) was developed to actuate the same Karl Storz cystoscope and consists of three modules.

(A) *Flexible cystoscope angulation*: The cystoscope's distal section can be deflected from $-210°$ to $+140°$. The flexible cystoscope shaft is 370 mm long, and the steerable distal section is 60 mm long and 5.5 mm in diameter. A linear servo is used to actuate angulation at the cystoscope's thumb lever.

(B) *Linear insertion*: A ball screw provides the translation action and has a working range of 30 cm. This module consist of a NEMA-17 stepping motor, the ball screw, and a linear bearing, and a slider carriage, which carries the cradle.

(C) *Cradle with roll module*: The cradle for the 3-DoF robot holds the cystoscope and provides rotation along the cystoscope's roll axis. The cradle consists of a 3D-printed body, a small drive pulley linked to a NEMA-17 stepping motor, a driven pulley fixed in a ball bearing, a timing belt, and a mounting point for the angulation servo. A removable clamping ring is mounted on the pulley to fix the cystoscope to the robotic mechanism.

A 3D bladder phantom made by the UW Medicine Center for Research and Education in Simulation Technologies (CREST) is used in the experiments, as shown in Fig. 3 (center). The phantom was created by capturing patient data through MRI and CT scanning. The bladder is digitally recognized and isolated by segmentation software and a digital file is created and 3D-printed as a mold. The resulting part represents the bladder volume as a positive form. This form is used as a mandrel to apply layers of platinum-cured and low-durometer silicone material (PlatSil silicone rubber, Polytek Development Corp., Easton, PA) to create the bladder wall. Attention is given to how the layers will be represented by the lighting and imaging from the cystoscope. Many semi-transparent layers are applied to capture depth of the tissue, highlight topology, and represent blood vessels within the phantom. The silicone form is cut and demolded from the mandrel and sealed with adhesive to make the cut line watertight. For simplicity of robot fixation and water filling of the phantom, we kept the 3D phantom inverted during data collection to avoid spilling the water. The robot was fixed on a flat table top above the phantom with some elevation. Although such positioning does not influence the performance of our method, in the future we do plan to improve our robot hardware and the sealing accessory of the phantom so that we can distend the bladder to a larger size and manipulate the scope to view the phantom in more optimal perspectives.

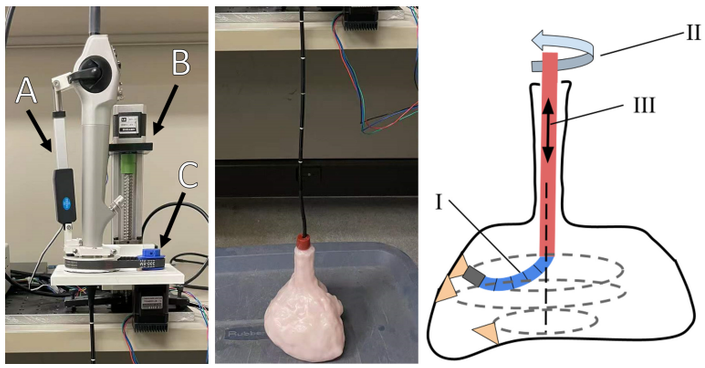

Fig. 3. 3D bladder phantom experiment setup. (Left) The 3 DoF cystoscope robot with three actuation modules: A — cystoscope angulation control, B — cystoscope insertion control, and C — cystoscope roll control. (Center) The cystoscope inserted into the 3D bladder phantom. During data collection, the phantom was filled with water and placed in a container among bags of rice to preserve position and shape. (Right) Data collection process for 3D phantom. I — The bend angle is adjusted to a sufficiently overlapping view ($> 20\%$) with the previous scan. II — The roll axis is actuated through one revolution clockwise and immediately counterclockwise while a video is recorded. The dashed lines represent the trajectory of the cystoscope tip during video recording. III — When the cystoscope hits the walls during a scan, the insertion length is changed and a new set of dictionary and test videos is collected.

3.2. *Experiments in 2.5D phantom*

3.2.1. *Dataset*

Cystoscope videos were captured with controlled saw-tooth-profile trajectories with constant velocity in both directions for 3 cycles, starting and finishing in a downward orientation. Serial data recording included a timestamp in milliseconds, control output, and servo sensor value. The amplitude of the trajectories was set so that the scope tip throughout the trial was not too close to blur bladder features. Trajectory speeds were either slow, medium, or fast (7°/s, 25°/s, 60°/s, respectively). Videos from the cystoscope were saved as MP4s with a frame rate near 27 Hz and a resolution of 720×720.

We focus on testing the robustness of our method when the video captured in the follow-up cystoscopies differs from the dictionary set in several ways: addition of tumors, distance changes between the cystoscope and bladder surface, and angulation speed changes.

Added Tumors I and II: Considering there may be new tumors emerging on the bladder surface between two clinical exams that may interfere with matching to a prior image, we add tumors to the test videos in two ways: (I) attaching a bladder tumor in Ta stage [50] with tape when collecting the test videos (Fig. 7(1st row)) and (II) digitally adding five different types of tumors in the test video frames (Fig. 7(2nd row)). Ta grade papillary tumors were used to test our image matching performance when the original image is obscured with a body of different structure, as would be the case with papillary tumors that grow into the bladder cavity. The tumors were retrieved from online image searches of surveillance cystoscopy and were scaled to our images based on the relative sizes of surrounding vasculature in source images. Digitally placed tumors were added onto test image frames in a random position and rotation.

Imaging Distance Change: During clinical cystoscopy, the bladder is enlarged with water and the enlarged volume may vary between exams by as much as 30%. The inspection distance between the cystoscope tip and the bladder surface will also vary between any two procedures. These factors will cause the imaged area of the same camera location to change between exams, which increases the difficulty of image retrieval from dictionary set (Stage I). We simulate these changes by localizing test frames with an FOV 30% smaller than the dictionary set by translating the cystoscope towards the bladder wall.

Movement Speed Change: During bladder screening and tumor inspection, there may be overly fast movement of the cystoscope leading to motion blur in the frames, which makes traditional tracking difficult. To test our performance during fast movement, we conducted an experiment with test videos with medium and fast movement speed, where the dictionary set is formed from slow speed video.

3.2.2. *Evaluation*

Quantitative evaluation and comparison with SIFT-only matching: We quantitatively show the performance of the test frame localization by evaluating the success rate of the coarse localization in Stage I and the registration accuracy based on SIFT feature matching between the test frame and the retrieved dictionary frame. The performance of image pair registration is determined by the overlap size and the SIFT feature extraction and matching (influenced by image quality), the former is an indicator of the image retrieval performance and the latter is

a crucial step in the 3D-2D correspondence in Stage II of our camera localization method. Since we do not have a ground truth to directly evaluate the camera pose recovery of test videos yet, we use the registration accuracy to partially evaluate our pipeline. Since cystoscope frames have relatively small size and large inter-frame overlap, they are unlikely to be significantly affected by the nonlinear deformation due to the nonplanar bladder surface, which allows for modeling the geometrical relationship between each test frame and its retrieved dictionary frame as a homography transformation for registration.

For comparison, we also present the registration performance with SIFT-only method without our coarse localization. The SIFT-only matching method extracts SIFT feature points from each test video frame to try to match them to features from all of the dictionary images with homography transformation. It takes $\mathcal{O}(n)$ time for each test frame to register with an overlapped dictionary frame globally, where n is the number of dictionary frames. A k-d tree can be used to accelerate the matching process with a time complexity $\mathcal{O}(\log(n))$ [51]. With the coarse localization in our pipeline, the computation time of the global registration is reduced to $\mathcal{O}(1)$.

We selected 25 test video frames in each test case, sampled randomly and distributed uniformly. To measure the registration accuracy between the test and dictionary frame, we use Target Registration Error (TRE) for comparison. Unlike entropy-based or similarity measures, TRE measures the result intuitively in pixels and is independent of different regularization methods [44,52]. For each test frame, five corresponding landmarks were selected by a trained observer. Two trained observers independently selected the corresponding landmarks from the test frame and the retrieved dictionary image. To obtain TRE for each image pair, we first calculate the homography transformation between the test frame and the retrieved dictionary image from matched SIFT features. Then we use the calculated homography to transform the landmark on retrieved dictionary image to the test frame. Lastly, we compute the distance between the transformed landmark points and local landmark points on the test frame. The root mean square of distances for all landmarks and test frames is calculated as the final TRE. A smaller TRE indicates a more accurate homography, which is usually caused by larger overlap and smaller perspective change between the image pair.

Angulation recovery based on image retrieval: To determine the accuracy of our angulation recovery without precisely aligning sensor and video data in each dictionary set, we compare the pixel distance from each test image to its correctly matched image. We then use a linear approximation to determine the tip angulation error from this pixel distance, or the scale between the increment of pixels in localization Δd and angulation $\Delta\alpha$. Taking the arc length as the distance between frames when $\Delta\alpha$ is small (1°), we get a linear relationship between $\Delta\alpha$ and Δd: $\Delta d = K \times \Delta\alpha$, where $K = 20$ pixels per degree.

To fully demonstrate our localization algorithm, we temporally aligned the video frames of the 1st cystoscopy exam video and the medium speed trial with the hysteresis-compensated sensor data corresponding to each video. The test frame's angulation was interpolated between the angles associated with its two nearest dictionary images. Since there is large overlap between close dictionary frames, we assume linear movement between continuous dictionary frames and interpolate accordingly.

3.3. *Experiments in 3D phantom*

3.3.1. *Dataset*

The same Karl Storz cystoscope was inserted into the water-filled 3D bladder phantom during the experiment, thus camera intrinsic parameters and other camera-related parameters are assumed to be unchanged from those in the 2.5D phantom data, except camera trajectory. Figure 10 shows several examples of the video frames collected from the inner surface of the 3D bladder phantom. The vessel features are much denser and thicker than the printed clinical bladder images shown in Fig. 7, and extra features are included, such as fixed bubbles, seams, shadows from surface topology, and floating particles.

Scanning of the 3D bladder phantom was performed in a series of circle trajectories enabled by rotating the cystoscope along its roll axis. Each circle trajectory has a fixed bend angle and the bend angles of different circles increase in the series, as sketched in Fig. 3(Right). The scanning is performed in a slow and constant moving speed of one circle per minute. The cystoscope was only able to image about half of the bladder surface with this simple trajectory before the distance from the bladder wall became too small. Full imaging of the bladder during cystoscopies requires larger distension of the bladder through pressurized fluid filling and precise, coordinated actuation of the cystoscope with respect to the anatomy that our current robotic platform is not yet capable of.

To further test the robustness of our method in the 3D phantom, two parameters are varied during data collection for two groups of experiments.

Tip Bending Angle Change: The first group of experiments aim to evaluate the performance when there is limited overlap between the dictionary images and the test images. Since our scanning is performed layer by layer, we control the view overlap by changing the bending angle of the cystoscope tip. Test scans are recorded at bending angles between those of the

dictionary scans at the same insertion depth within the bladder. The test images have 10–25% vertical shifting with the dictionary images, and they are divided into levels of tip bending I and II. Note that these test videos still contain perspective changes and other potential local deformations because they are separate scans.

Insertion Depth Change: The second group of experiments aim to evaluate performance with changes in the imaging distance during cystoscope scanning which simulates the bladder volume variation between different exams. We set different insertion depths of the cystoscope to change the distance during the test video scanning. Three different insertion levels I, II, III are used which are 2.5, 5 and 10 mm from the insertion depth used in the dictionary video. With the insertion depth change, there is also trajectory shifting between the test and dictionary scanning.

3.3.2. *Evaluation*

Similar to the 2.5D phantom case, we also compare our method with SIFT-only matching by the success rate and mean TRE. Within each level of changed tip bending angle and insertion depth, 100 test frames are sampled and coarse-localized with the dictionary set.

Due to lack of reliable ground truth for camera poses, our camera pose recovery is qualitatively demonstrated. We visualize the trajectory of recovered camera poses (both translation and orientation) for the test video frames in tip bending angle II with respect to the reconstructed 3D model. Since the test videos are acquired by scanning the bladder phantom in circles as in Fig. 3 (Right), we can visually evaluate the quality of the recovered camera poses.

4. Results

4.1. *Hysteresis model in 2.5D setup*

Tip angulation kinematic data (Fig. 4) shows a hysteresis of 6.5° at the thumb lever. No discernible dead-band is observed. The resulting parallel hysteresis model is $\alpha = 5 \times \theta \pm 16$, where α is the angulation angle in degrees and θ is the thumb lever angle difference from center. The angulation estimation looks for inflection points, at which it maintains its estimate and either: switches the model when the thumb lever has continued in the new direction past the 6.5° horizontal gap; or returns to the original model when the lever angle movement matches the original direction and passes the initial inflection point. The 32° vertical gap between models at a given thumb lever angle represents the imperfect precision of kinematic estimation when the direction of the thumb lever movement is unknown or not modeled.

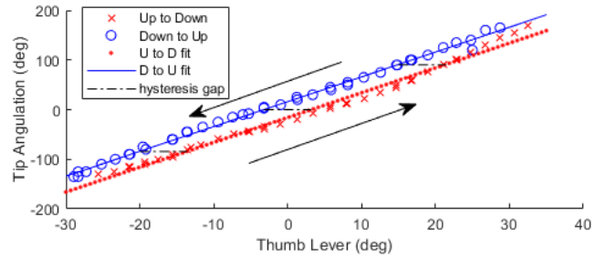

Fig. 4. Hysteresis model of cystoscope angulation. When the direction of the cystoscope changes, the estimated value is held constant until the sensor value returns to the point of change or crosses the "hysteresis gap", the horizontal distance between the parallel lines.

4.2. *3D reconstruction*

To evaluate the accuracy of reconstruction, we first align the ground truth model of the phantom and the reconstructed model in Meshlab [53]. We then use Meshlab to calculate Hausdorff distance, which represents the upper bound of accuracy of all reconstructed points.

4.2.1. *Reconstruction of 2.5D phantom*

Using 156 frames as input, the offline 3D reconstruction takes 560 s (9.3 min) on average. After the bladder phantom reconstruction is aligned with ground truth as in Fig. 5, the Hausdorff distance is calculated to be 0.0319 (normalized over diagonal of bounding box), i.e. the error is bounded within 3.2% of the size of the phantom. Refer to our prior work [32] for detailed instructions on model alignment for evaluation of shape reconstruction.

4.2.2. *Reconstruction of 3D phantom*

Using 548 frames as input, the offline 3D reconstruction of the 3D phantom takes 1928 s (32 minutes) on average. After the bladder phantom reconstruction is aligned with ground truth, the Hausdorff distance is calculated to be is 0.0290 (normalized over diagonal of bounding box), i.e. error is bounded within 3% of the size of the phantom.

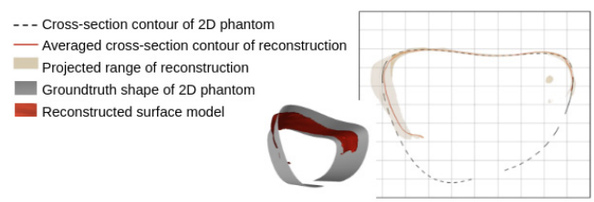

Fig. 5. Comparison between reconstructed model of cystoscope scanned surface and the ground truth shape of 2.5D bladder phantom. Inset plot shows the reconstructed surface model (red) aligned with the 3D surface ground truth surface shape of the phantom (gray).

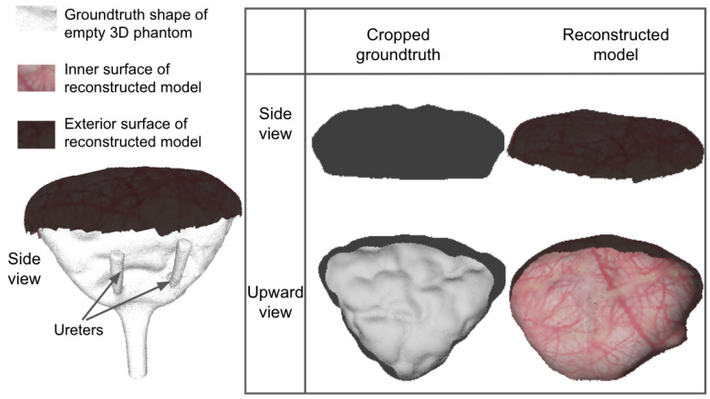

Fig. 6. (Left): Comparison between reconstructed model of cystoscope scanned surface and the ground truth shape of 3D bladder phantom. (Right): Side view and upward view of cropped ground truth shape and reconstructed surface model of 3D bladder phantom.

Note that the ground truth shape of 3D phantom is acquired from a 3D scan of the mold that was used to make the 3D phantom. Since the phantom slightly expands when filled with water, it is expected that the reconstructed surface model is actually larger than the original model, as shown in Fig. 6. This means that the reconstruction may have better accuracy than what is shown by the calculated Hausdorff distance.

4.3. *Camera localization*

4.3.1. *Localization results on 2.5D phantom*

The performance of our localization approach and SIFT-only approach among sampled test frames is defined by success rate, runtime, and average TRE of successful matches, as shown in Table 1. The success rate of the SIFT-only control method is defined as the percentage of successful matching pairs with TRE less than 15 pixels. The success rate of Stage I in our method is defined as the percentage of test frames matched with a correct dictionary image with recognizable overlap. We also perform SIFT-based fine registration (denoted as Reg. in Table 1) on the test frame and its retrieved dictionary image to calculate TRE. Thus success rate of our method followed by the fine registration is also calculated as for the SIFT-only method.

Since we also match SIFT features in registration, the TRE of SIFT-only among successful cases are similar with ours over the 2.5D phantom. In every experiment, our method has a significant improvement over SIFT-only in success rate. Except when changing FOV via imaging distance, our success rate is over 96%. Our method can reach an accuracy of less than 10-pixel TRE with an average observer variability of 2.98 ± 1.64. When the FOV changes, the success rate of registration is 80%, while most test images can be matched with a correct dictionary image in Stage I (96% success rate).

The coarse localization by Stage I of our method improves the running speed of fine registration of test images to a correct match among hundreds of dictionary images to around 20 times faster than using a SIFT-only global registration method. Matching the test frames in the LDS happens at about ~ 20 fps, which is over $100\times$ faster than SIFT-only. Several success and failure examples with challenging test video conditions are shown in Fig. 7. Failure cases of the subsequent fine registration indicate there are insufficient matching SIFT points after the correct image retrieval.

In the coarse localization, we select the top 20 PCA coefficients to form the low-dimensional representation of the dictionary frames in LDS. Figure 8 shows a distance map calculated from 200 dictionary frames (continuously sampled from a video sequence) after they are

Table 1. Localization performance per image frame over 2.5D phantom.

	SIFT-only		Ours		
Changes from dictionary	Success rate	Runtime	Success rate (Stage I/Reg.)	Average TRE (Pix)	Runtime (Stage I/Reg.)
Tumor I	80%	6263 ms	100/96%	5.6	53/368 ms
Tumor II	84%	6482 ms	100/100%	3.43	42/332 ms
Distance	56%	6977 ms	96/80%	8.4	51/291 ms
Speed (Med)	72%	6087 ms	100/100%	6.7	55/304 ms
Speed (Fast)	76%	6115 ms	100/100%	5.9	47/312 ms

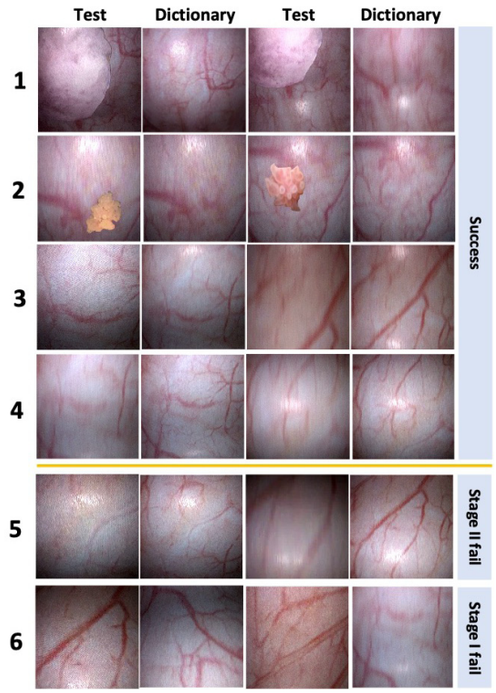

Fig. 7. Test frames and matched dictionary images of success and failure examples of our algorithm within the 2.5D phantom. *Rows 1-2*: our algorithm identified an image match with the dictionary even with a physical tumor (1) and digitally added tumor (2) taking up much of the frame. *Rows 3-4*: challenging examples registered when FOV changes or frames exhibit motion blur. *Row 5*: Stage I (coarse localization) succeeds while SIFT-based registration fails. *Row 6*: Stage I failures. All failure examples are from changing FOV trial.

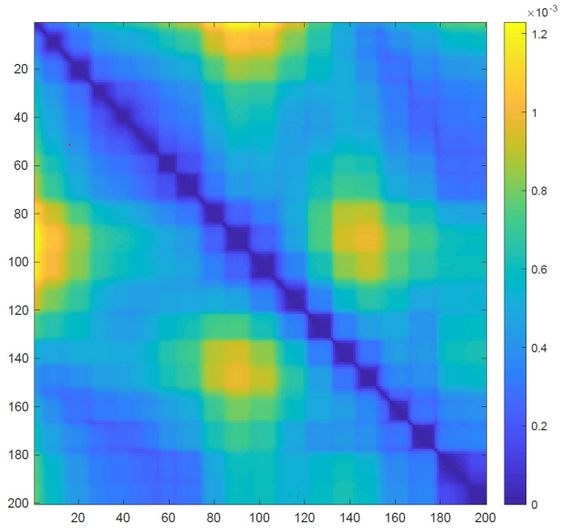

Fig. 8. Distance map of 200 dictionary frames in LDS. Dark values indicate small distances in LDS and larger overlap between the image pair; light values indicate large distances in LDS and smaller or no overlap between the image pair.

Table 2. Angulation prediction success rate.

Tumor I	Tumor II	FOV change	Speed (Med)	Speed (Fast)
96.5%	98.2%	85.1%	99.7%	97.3%

mapped into the LDS. In the distance map, the intensity of a square at row i and column j indicates distance between low-dimensional representations of frame i and frame j in LDS. We can see that the shortest distance values is gathered near the diagonal axis. It confirms that in the LDS, each image is still closest to its adjacent frames, which should have the largest overlap with the image.

Table 2 shows the percentage of robot angles computed from frame localization with an error less than 5° compared to the pixel-based linearization of the averaged robot trajectory (i.e. within 100 pixels of the sawtooth trajectory in each trial). Each test video is downsampled to 370 frames. We only take the results in Stage I when the subsequent fine registration fails. In most cases, the robot tip position can be correctly estimated with a success rate over 96%. Changing the test video FOV (imaging distance) by about 30% increases the difficulty and the success rate is only 85.1%. The angulation trajectory reconstructed from the medium speed trial is seen in Fig. 9. The RMS trajectory error over the 23-s long trial is 9.4° and the RMS error between the coarse and fine estimates is only 1.3°.

4.3.2. *Localization results on 3D phantom*

Table 3 shows the success rate, runtime, and the average TRE of successful matches of our coarse localization + fine registration approach and SIFT-only approach among different test videos. The success rate and TRE are defined to be the same as in the 2.5D phantom case. Except for the Insertion Depth III test, our success rate is over 99% in all cases. Our method reaches an accuracy of less than 3-pixel TRE with an average observer variability of 1.32 ± 1.02. With sufficient distinctive feature points, SIFT-only method in these experiments has a high success rate, however, it is very time consuming with a runtime of each test frame around 60–75 times slower than our method. The coarse localization (Stage I) is over $1000\times$ faster than SIFT-only method. In the case of insertion depth III, our success rate is 4% lower than the SIFT-only method. The SIFT-only method can sometimes find the correct match with the overlap of selected matched pairs less than ours, especially in insertion depth change, thus we have a smaller TRE among success matches in these cases.

Several success and failure examples under different types of test videos are shown in Fig. 10.

Figure 11 visualizes an example of the camera localization results. In this example, the dictionary images are

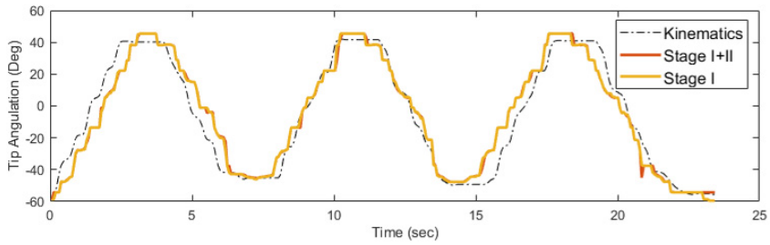

Fig. 9. Angulation trajectory computed from our localization method during the medium speed trial. The dictionary images are paired with kinematics estimates synchronized with video recording and the localized trajectory is compared to kinematics estimates from the same trial.

Table 3. Localization performance per image frame over 3D phantom.

	SIFT-only			Ours		
Changes from dictionary	Success rate	Average TRE (Pix)	Runtime	Success rate (Stage I/Reg.)	Average TRE (Pix)	Runtime (Stage I/Reg.)
Tip bending I	100%	1.86	38,676 ms	100%/100%	1.81	43 ms/602 ms
Tip bending II	100%	2.53	37,123 ms	99%/99%	2.20	41 ms/619 ms
Insertion I	100%	2.56	38,965 ms	100%/100%	2.37	46 ms/634 ms
Insertion II	99%	5.09	39,012 ms	99%/99%	2.82	43 ms/645 ms
Insertion III	98%	5.12	37,841 ms	94%/94%	1.98	42 ms/622 ms

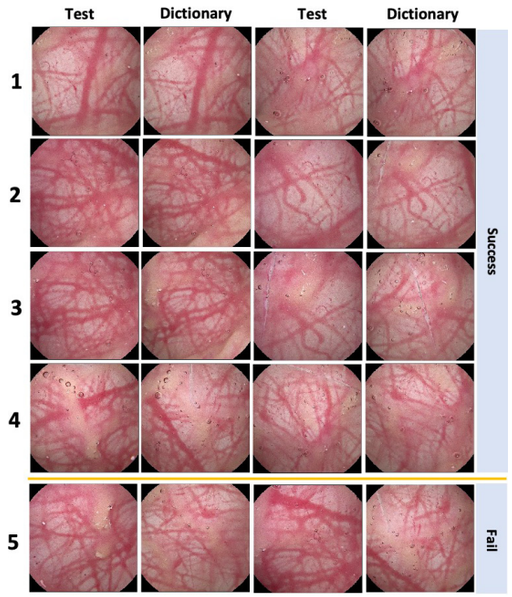

Fig. 10. Test frames and retrieved dictionary images of success and failure examples of our algorithm within the 3D phantom. *Row 1*: Success examples in tip bending angle change; *Rows 2–4*: Success examples in insertion depth change; *Row 5*: Failure cases in insertion depth change.

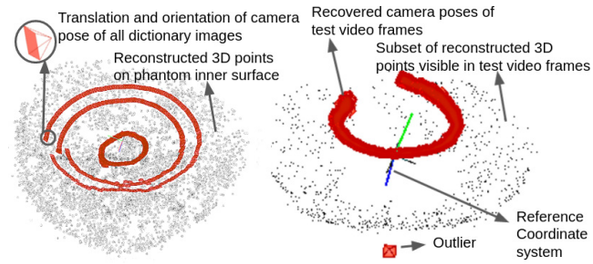

Fig. 11. (Left): Visualization of reconstructed 3D point cloud and camera poses of all dictionary images. (Right): The subset of reconstructed 3D points that are visible in test video frames and the therefrom recovered camera poses of test video frames.

acquired in three circles with different tip bending angles and the same insertion length to achieve 3D reconstruction. Figure 11(Left) shows the camera poses (denoted by solid red frustums) of all dictionary images the point cloud of the reconstructed 3D model (denoted by blacks points). The test video in tip bending angle II has a tip bending angle between the top two largest angles used in the dictionary set. With the two-stage camera localization pipeline, we found the subset of 3D points from the reconstructed 3D point cloud that are visible in test frames. This subset appears to be a ring (Fig. 11(Right)). We then extracted 3D-2D correspondence based on the matching relation among test image, its corresponding retrieved dictionary image and the reconstructed 3D point cloud. Finally, the camera poses are recovered as shown in Fig. 11(Right), which appears to be a circle trajectory with camera facing towards the phantom wall. There is only one outlier below the point cloud whose recovered camera pose is clearly wrong.

5. Discussion

Our two-stage camera localization method can provide pixel-level accuracy in several clinically relevant test cases. Compared to tracking between continuous frame for relative pose recovery, localizing every frame globally for absolute pose recovery avoids accumulated errors and the effects of failure cases, which occurs more frequently in surgical videos than ordinary tracking tasks. Low-dimensional mapping in Stage I was shown to significantly improve the efficiency of image retrieval and can be used for coarse localization in challenging conditions that might be encountered in surveillance telecystoscopy. As shown in Fig. 9, our coarse localization step has a mean error of less than 10° including the error in the kinematic ground truth. So the coarse localization can be independently used when high speed is required or feature matching in Stage II fails. The coarse localization using cystoscopes with 100° FOV should provide sufficient accuracy for presenting pose estimates and maintain sufficient overlap with the prior map to teleoperators.

PCA used in our Stage I is sensitive to outliers, occlusions, and corruption in the data. Robust Principal Component Analysis (RPCA) was introduced to address this issue [54,55]. In general, RPCA is more expensive than PCA, requiring an iterative optimization to decompose the original matrix. In our pipeline, dictionary images are selected during the 3D reconstruction algorithm to be good-quality frames. With few enough outliers in the dictionary set, RPCA is not necessary. However, it is important to keep RPCA as an option for data with outliers and corruption. The distance map in Fig. 8 shows that our dimension reduction process keeps the position relationship of the input dictionary frames in this near-clinical data while providing a more efficient searching space. If this were not the case, dark areas would appear away from the main diagonal within one circle of scanning, indicating that the dimension reduction did not sufficiently separate disparate images within the LDS. In such a case, new images may be mismatched to the wrong area of the bladder.

Hysteresis modeling of our cystoscope shows that image-based pose estimation is needed for providing a capable and reliable teleoperation interface for robotic cystoscopy since real time, accurate forward kinematic estimation may be difficult. To inspect the entire urothelium, urologists will deflect a flexible cystoscope against the bladder. Although this achieves viewing angles in retroflexion, this also introduces significant difficulties in estimating the pose of the scope with traditional kinematic approaches. In the 2.5D phantom case, we use cystoscope-specific kinematics to provide ground truth angulation data for the image dictionary in Fig. 9 after alignment, and a camera pose estimate is derived from the 3D reconstruction in the 3D phantom case. In the next step, a reliable ground truth of camera poses will be collected from extra sensors, for example, attaching electromagnetic tracking sensors on the cystoscope tip, to quantitatively evaluate the reconstructed camera pose trajectory.

We tested our approach in both 2.5D phantom and 3D phantom of the bladder. The 2.5D phantom is simple in shape so that our initial single-DOF robotic cystoscope can cover the whole phantom while recording videos that simulate cystoscopy. It is also rigid and open so that we can measure the ground truth trajectory of the camera easily and evaluate the accuracy of kinematics. The 3D phantom is an effort to evaluate our method in an environment with a more realistic shape (by using a 3D phantom made of distensible and deformable material). However, the acquisition of ground truth for camera trajectory/poses is much harder in this case because the phantom is close and we can't rely on measurement from the electromagnetic tracker due to the large error observed. Also, the manually designed vessel features on it are not realistic enough and cannot well present the robustness of our method over image degradations. Thus, we describe the experimental results on both the 2.5D and 3D phantoms to present the performance and potential of our method as clearly as possible. The scanning of 3D phantom is performed with the phantom filled with water, which more closely simulates clinical conditions. The captured videos therefore contain bubbles and floating debris. The deformable phantom material can cause local distortions during the scanning. Perspective changes between different scans will cause the features and illuminations from the same region to appear different. With homogeneous vessel features from a larger surface in 3D phantom, there will be more local optima interference in the global localization. The spatially dense, hand-painted vessels in the 3D bladder (Fig. 10) also provide more SIFT features than the printed human bladder image in the 2.5D phantom, thus allowing the SIFT-only method to achieve higher matching accuracy in the 3D tests than in the 2.5D tests. However, the increasing runtime factor of the SIFT-only case, from $100\times$ to $1000\times$ between 2.5D and 3D tests, makes it hard to use in teleoperation where reasonable computational complexity is important. Although the dictionary set in the 3D case only covers a portion of the bladder phantom, an increase in the dictionary size should not greatly affect the runtime of our method.

For camera pose recovery in Stage II, we also experimented with using the 2D-2D feature correspondences between the test image and its retrieved dictionary image to calculate the transformation between the two images and then recover the camera pose of test image. We observed that using 3D-2D correspondences for camera pose recovery has better reliability than using 2D-2D correspondences. This is reasonable since the global bundle adjustment in the reconstruction step provides 3D points that are calculated to be more

globally consistent with all collected images. Thus the 3D-2D correspondences are much more well-constrained and less subject to noise, compared to 2D-2D correspondences. The trajectory of the recovered test frame poses shown in Fig. 11 qualitatively indicate the reliability of camera pose recovery from 3D-2D correspondences, as the trajectory of the source test video is a similar circle scan at a constant tip bend angle.

Future development of image-based localization using the 3D phantom can investigate new approaches to maintain robust tracking. For example, multiple dictionary images can be retrieved for each test frame and their matching relationship with the test frame can be studied to find more reliable 3D-2D correspondences. Utility of the 3D reconstruction and real-time image matching can provide new user interfaces in teleoperation of medical robotics. Our 3D reconstruction results demonstrate reasonably accurate reconstruction of shape and texture of the bladder, which is crucial for accurate display of the bladder during teleoperation. Once camera pose of a new image is recovered, the newly acquired image can be mapped onto the 3D surface model and highlighted on the model for the operator. Not only will this help situational awareness during telecystoscopy, this could also be implemented during manual cystoscopy for training urology residents. If examined image patches are shown in contrast with unexamined areas, trainees can visualize completeness during the procedures and a real-time completeness metric can be calculated.

Additional testing is required to demonstrate efficiency and accuracy with more realistic cystoscopy videos with a wide range of bladder cancer tumors and natural anatomical variation. The experiments conducted on these phantoms provide higher image quality than a real cystoscopic video from a human bladder containing urine and water/saline. In addition, the bladder surface deformation during scanning is also not considered in the performance evaluation. When using clinical videos, the 3D reconstruction and localization performance may be affected by image degradation. With the proposed two-stage framework, both the coarse localization and camera pose recovery in our pipeline may be improved with deep-learning-based approaches [56,57]. Moreover, our localization method could be especially useful when combined with other estimation technologies. For instance, if applying continuous frame tracking, our coarse localization can provide a quick and accurate estimate to regain tracking when continuous localization fails. Finally, a Kalman filter could be used to combine our global localization with continuous frame tracking and endoscope kinematics to make a more robust teleoperation system.

6. Conclusion

Our coarse localization algorithm is shown to be 100–1000× faster than a SIFT-only dictionary matching approach in the context of a two-stage camera localization pipeline that could be used for bladder cancer surveillance where 3D bladder models can be reconstructed after a primary exam. In the follow-up visits, our algorithm can efficiently estimate a flexible cystoscope's tip pose at around 20 Hz in bladder phantoms. We believe that our algorithm will be able to perform well in more realistic scenarios and could help make telecystoscopy a compelling option for urologists and their patients.

Acknowledgment

Co-authors Chen Gong, Yaxuan Zhou and Andrew Lewis contributed equally in this journal article. The authors thank urologists Smita De and Lee White for their insight and advice. Funding is provided by the UW Mechanical Engineering Department, UW CoMotion Innovation Fund, UW CREST, and NSF 1631146 PFI:BIC with INTERN program in collaboration with VerAvanti Inc. Additional thanks to Professors Audrey Bowden for the 3D reconstruction pipeline and Steven L. Brunton for algorithm advice and Karl Storz endoscopes for loaning the cystoscopy system.

References

1. J. Engelsgjerd and C. Deibert, Cystoscopy (2020).
2. R. L. Siegel, K. D. Miller, H. E. Fuchs and A. Jemal, Cancer Statistics, 2021, *CA: A Cancer J. Clin.* **71** (2021) 7–33.
3. M. Mossanen and J. L. Gore, The burden of bladder cancer care: Direct and indirect costs, *Curr. Opin. Urol.* **24** (2014) 487–491.
4. K. Chamie, M. S. Litwin, J. C. Bassett, T. J. Daskivitch, J. Lai, J. M. Hanley and B. R. Konety, C. S. Saigal and Urologic Diseases in America Project, Recurrence of high-risk bladder cancer: A population-based analysis, *Cancer* **119**(17) (2014) 3219–3227.
5. National Comprehensive Cancer Network, Bladder Cancer (Version 4.2021), https://www.nccn.org /professionals/physician_gls/pdf/bladder.pdf.
6. B. K. Hollenbeck, R. L. Dunn, Z. Ye and J. M. Hollingsworth, Delays in diagnosis and bladder cancer mortality, *Cancer* **116** (2010) 5235–5242.
7. The State of the Urology Workforce and Practice in the United States, Technical Report, American Urology Association (2019).
8. J. Marescaux and F. Rubino, Transcontinental robot-assisted remote telesurgery, feasibility and potential applications, *Ann. Surg.* **235**(4) (2006) 487–492.
9. C. F. Graetzel, A. Sheehy and D. P. Noonan, Robotic bronchoscopy drive mode of the Auris Monarch platform, *Int. Conf. Robotics and Automation (ICRA)* (IEEE, Montreal, Canada, 2019), pp. 3895–3901.
10. L. Yarmus, J. Akulian, M. Wahidi, A. Chen, J. P. Steltz, S. L. Solomon, D. Yu, F. Maldonado, J. Cardenas-Garcia, D. Molena, H. Lee and A. Vachani, A prospective randomized comparative study of three guided bronchoscopic approaches for investigating pulmonary nodules: The precision-1 study, *Chest* **157** (2020) 694–701.
11. N. Sarli, G. Del Giudice, S. De, M. S. DIetrich, S. D. Herrell and N. Simaan, Preliminary porcine in vivo evaluation of a telerobotic system for transurethral bladder tumor resection and surveillance, *J. Endourol.* **32**(6) (2018) 516–522.
12. N. Sarli, G. Del Giudice, S. De, M. S. Dietrich, S. D. Herrell and N. Simaan, TURBot: A system for robot-assisted transurethral

bladder tumor resection, *IEEE/ASME Trans. Mechatron.* **24**(4) (2020) 1452–1463.
13. R. J. Hendrick, C. R. Mitchell, S. Duke Herrell and R. J. W. Iii, Hand-held transendoscopic robotic manipulators: A transurethral laser prostate surgery case study, *Int. J. Robot. Res.* **34**(13) (2015) 1559–1572.
14. A. Wong, Y. Phan, H. Thursby and W. Mahmalji, The first UK experience with single-use disposable flexible cystoscopes: An in-depth cost analysis, service delivery and patient satisfaction rate with AmbuⓇaScope 4 Cysto, *J. Endoluminal Endourol.* **4** (2021) e29–e44.
15. A. Schwein, B. Kramer, P. Chinnadurai, S. Walker, M. O'Malley, A. Lumsden and J. Bismuth, Flexible robotics with electromagnetic tracking improves safety and efficiency during in vitro endovascular navigation, *J. Vasc. Surg.* **65**(2) (2017) 530–537.
16. E. D. Rozeboom, R. Reilink, M. P. Schwartz, P. Fockens and I. A. M. J. Broeders, Evaluation of the tip-bending response in clinically used endoscopes, *Int. J. Med. Robot. Comput. Assist. Surg.* **9** (2013) 240–246.
17. B. Bardou, F. Nageotte, P. Zanne and M. De Mathelin, Improvements in the control of a flexible endoscopic system, in *Proc. — IEEE Int. Conf. Robotics and Automation* (Institute of Electrical and Electronics Engineers Inc., 2012), pp. 3725–3732.
18. L. Sliker, G. Ciuti, M. Rentschler and A. Menciassi, Magnetically driven medical devices: A review, *Expert Rev. Med. Devices* **12**(6) (2015) 737–752.
19. J. Li, E. S. Barjuei, G. Ciuti, Y. Hao, P. Zhang, A. Menciassi, Q. Huang and P. Dario, Magnetically-driven medical robots: An analytical magnetic model for endoscopic capsules design, *J. Magn. Magn. Mater.* **452** (2018) 278–287.
20. F. Bianchi, A. Masaracchia, E. Shojaei Barjuei, A. Menciassi, A. Arezzo, A. Koulaouzidis, D. Stoyanov, P. Dario and G. Ciuti, Localization strategies for robotic endoscopic capsules: A review, *Expert Rev. Med. Devices* **16**(5) (2019) 381–403.
21. C. Fang, W. Sang, J. D. J. Gumprecht, G. Strauss and T. C. Lueth, Image-guided steering of a motorized hand-held flexible rhino endoscope in ENT diagnoses, *2012 IEEE Int. Conf. Robotics and Biomimetics, ROBIO 2012 — Conf. Digest* (2012), pp. 1086–1091.
22. C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid and J. J. Leonard, Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age, *IEEE Trans. Robot.* **32**(6) (2016) 1309–1332.
23. O. G. Grasa, E. Bernal, S. Casado, I. Gil and J. M. M. Montiel, Visual slam for handheld monocular endoscope, *IEEE Trans. Med. Imaging* **33**(1) (2014) 135–146.
24. N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon and J. M. M. Montiel, Live tracking and dense reconstruction for handheld monocular endoscopy, *IEEE Trans. Med. Imaging* **38**(1) (2019) 79–89.
25. C. Xie, T. Yao, J. Wang and Q. Liu, Endoscope localization and gastrointestinal feature map construction based on monocular slam technology, *J. Infect. Publ. Health* **13**(9) (2020) 1314–1321.
26. C. Xie, T. Yao, J. Wang and Q. Liu, Endoscope localization and gastrointestinal feature map construction based on monocular slam technology, *J. Infect. Publ. Health* **13**(9) (2020) 1314–1321.
27. T. D. Soper, M. P. Porter and E. J. Seibel, Surface mosaics of the bladder reconstructed from endoscopic video for automated surveillance, *IEEE Trans. Biomed. Eng.* **59**(6) (2012) 1670–1680.
28. A. Ben-Hamadou, C. Daul and C. Soussen, Construction of extended 3d field of views of the internal bladder wall surface: A proof of concept, *CoRR*abs/1607.04773 (2016).
29. K. L. Lurie, R. Angst, D. V. Zlatev, J. C. Liao and A. K. Ellerbee Bowden, 3d reconstruction of cystoscopy videos for comprehensive bladder records, *Biomed. Opt. Express* **8**(4) (2017) 2106–2123.
30. Q. Pentek, S. Hein, A. Miernik and A. Reiterer, Image-based 3d surface approximation of the bladder using structure-from-motion for enhanced cystoscopy based on phantom data, *Biomedizinische Technik. Biomed. Eng.* **63**(4) (2018).
31. N. O. Falcon, S. Ranjbar, E. Cisneros, B. Vu, A. Schoppe, P. Sanchez, Y. Jin, J. Ye, Y. Feng, D. Kaushik and R. L. Hood, Innovative computer vision approach to 3D bladder model reconstruction from flexible cystoscopy, in *Therapeutics and Diagnostics in Urology 2019*, Vol. 10852 (SPIE, 2019), pp. 18–26.
32. Y. Zhou, R. L. Eimen, E. J. Seibel and A. K. Bowden, Cost-efficient video synthesis and evaluation for development of virtual 3d endoscopy, *IEEE J. Transl. Eng. Health Med.* **9** (2021) 1800711.
33. D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vis.* **60**(2) (2004) 91–110.
34. H. Bay, A. Ess, T. Tuytelaars and L. Van Gool, Speeded-up robust features (surf), *Comput. Vis. Image Understand.* **110**(3) (2008) 346–359.
35. A. S. Vemuri, K.-C. Liu, Y. Ho, H.-S. Wu and M.-C. Ku, Endoscopic video mosaicing: Application to surgery and diagnostics, *Living Imaging Workshop* (2011), pp. 1–2.
36. D. K. Iakovidis, E. Spyrou and D. Diamantis, Efficient homography-based video visualization for wireless capsule endoscopy, *13th IEEE Int. Conf. BioInformatics and BioEngineering* (IEEE, 2013), pp. 1–4.
37. R. Richa, B. Vágvölgyi, M. Balicki, G. Hager and R. H. Taylor, Hybrid tracking and mosaicking for information augmentation in retinal surgery, *Int. Conf. Medical Image Computing and Computer-Assisted Intervention* (Springer, 2012), pp. 397–404.
38. Cancer stat facts: Bladder Cancer (National Cancer Institute, Surveillance, Epidemiology, and End Results (SEER) Program) (2018).
39. C. Wengert and M. Reeff, Fully automatic endoscope calibration for intraoperative use, in *Bildverarbeitung in der Medizin*, Informatik aktuell (Springer, Berlin, 2006), pp. 419–423.
40. C. Zach and M. Pollefeys, Practical methods for convex multi-view reconstruction, Lect Notes Comput Sci, Vol. 6314 (2010).
41. M. Kazhdan, M. Bolitho and H. Hoppe, Poisson surface reconstruction, *Symp. Geom Process*, Vol. 7 (2006).
42. M. Waechter, N. Moehrle and M. Goesele, Let there be color! large-scale texturing of 3d reconstructions, in *Proc. ECCV* (2014).
43. Y. Wu, F. Tang and H. Li, Image-based camera localization: An overview, *Vis. Comput. Ind. Biomed. Art* **1**(8) (2018).
44. C. Gong, N. B. Erichson, J. P. Kelly, L. Trutoiu, B. T. Schowengerdt, S. L. Brunton and E. J. Seibel, Retinamatch: Efficient template matching of retina images for teleophthalmology, *IEEE Trans. Med. Imaging* **38**(8) (2019) 1993–2004.
45. B. Schölkopf, A. Smola and K.-R. Müller, Kernel principal component analysis, *Int. Conf. Artificial Neural Networks* (Springer, 1997), pp. 583–588.
46. J. B. Tenenbaum, V. De Silva and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* **290** (5500) (2000) 2319–2323.
47. P. Chen, C. Gong, A. Lewis, Y. Zhou, E. J. Seibel and B. Hannaford, Real-time flexible endoscope navigation within bladder phantom having sparse non-distinct features is enhanced with robotic control, in *SPIE Medical Imaging* (2022).
48. R. Reilink, A. M. Kappers, S. Stramigioli and S. Misra, Evaluation of robotically controlled advanced endoscopic instruments, *Int. J. Med. Robot. Comput. Assist. Surg.* **9** (2013) 240–246.
49. H. F. Talari, R. Monfaredi, E. Wilson, E. Blum, C. Bayne, C. Peters, A. Zhang and K. Cleary, Robotically assisted ureteroscopy for kidney exploration, in *Proc. SPIE Int. Soc. Opt. Eng.*, Vol. 10135 (International Society for Optics and Photonics, 2017), p. 1013512.
50. P. A. Geavlete (ed.), Endoscopic treatment of bladder tumors, in *Endoscopic Diagnosis and Treatment in Urinary Bladder Pathology*, Chap. 4 (Academic Press, San Diego, 2016), pp. 83–203.

51. M. Brown and D. G. Lowe, Automatic panoramic image stitching using invariant features, *Int. J. Comput. Vis.* **74**(1) (2007) 59–73.
52. J. M. Fitzpatrick, J. B. West and C. R. Maurer, Predicting error in rigid-body point-based registration, *IEEE Trans. Med. Imaging* **17** (5) (1998) 694–702.
53. Meshlab software, https://github.com/cnr-isti-vclab/meshlab/releases/tag/v2016.12, Accessed on December 2021.
54. E. J. Candès, X. Li, Y. Ma and J. Wright, Robust principal component analysis?, *J. ACM* **58**(3) (2011) 1–37.
55. T. Bouwmans, S. Javed, H. Zhang, Z. Lin and R. Otazo, On the applications of robust pca in image and video processing, *Proc. IEEE* **106**(8) (2018) 1427–1457.
56. Y. Wang, H. Yao and S. Zhao, Auto-encoder based dimensionality reduction, *Neurocomputing* **184** (2016) 232–242.
57. M. Niethammer, R. Kwitt and F.-X. Vialard, Metric learning for image registration, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (2019), pp. 8463–8472.