Learning 2D Surgical Camera Motion From Demonstrations

Jessica J. Ji[1], Sanjay Krishnan[1], Vatsal Patel[1], Danyal Fer[2], Ken Goldberg[1]

***Abstract*— Automating camera movement during robot-assisted surgery has the potential to reduce burden on surgeons and remove the need to manually move the camera. An important sub-problem is automatic viewpoint selection, proposing camera poses that focus on important anatomical features at the beginning of a task. We use the 6 DoF Stewart Platform Research Kit (SPRK) to simulate camera movements and study camera motion in surgical robotics. To provide demonstrations, we link the platform's control directly to the da Vinci Research Kit (dVRK) master control system and allow control of the platform using the same pedals and tools as a clinical movable endoscope. We propose a probabilistic model that identifies image features that "dwell" close to the camera's focal point in expert demonstrations. Our experiments consider a surgical debridement scenario on silicone phantoms with foreign bodies of varying color and shape. We evaluate the extent to which the system correctly segments candidate debridement targets (box accuracy) and correctly ranks those targets (rank accuracy). For debridement of a single uniquely colored foreign body, the box accuracy is 80% and the rank accuracy is 100% after 100 training data points. For debridement of multiple foreign bodies of the same color, the box accuracy is 70.8% and the rank accuracy is 100% after 100 training data points. For debridement of foreign bodies with a particular shape, the box accuracy is 70.5% and the rank accuracy is 90% after 100 training data points. A demonstration video is available at: `https://vimeo.com/260362958`**

***Index Terms*— Surgical Robotics; Active Perception; Viewpoint Selection**

I. Introduction

Camera positioning and movement is an important skill in manual laparoscopic surgery [1–4], and robot-assisted laparoscopic surgery has the potential to facilitate partial automation of camera control [1, 5–7] to free the surgeon's hands to concentrate on manipulating laparoscopic instruments such as scissors, needle drivers, or electrocautery tools. A critical sub-problem is automatic *viewpoint selection*, i.e., suggesting camera angles and positions that center on important regions at the beginning of a task. The viewpoint selection problem complements existing literature on automated surgical camera movement that focuses on local tracking models, which center the camera around the positions of the tools or follow surgeon eye-gaze [6]. We explore learning to identify features of interest from endoscopic images viewed during demonstrations performed by expert surgeons to yield a predictive model that learns features of interest to guide camera positioning.

We base our learning from demonstration (LfD) approach on a probabilistic model that identifies anatomical features that "dwell" close to the camera's focal point during control by an expert surgeon. We consider stereo planar camera movements consisting of 2D pans with brightly colored phantom anatomy to demonstrate this concept. The technique first coarsely segments an endoscopic image into a set of candidate bounding boxes. We track an expert surgeon's camera movements and score the bounding boxes by how well they are centered in the camera's field of view–bounding boxes closer to the focal point are more relevant than those that are further away. Using the inverse kinematics of the movable camera, the bounding box can be converted into a camera position. For camera trajectories with multiple targets, a trajectory segmentation model is used to detect points at which the camera is stationary. This scoring metric provides a weak, noisy label which can be fed into a robust linear regression model that predicts the score given features of the anatomy in the bounding box. In a future procedure, this regression model can be used to rank a set of candidate bounding boxes by their relevance. An example output of the model is illustrated in Figure 1.

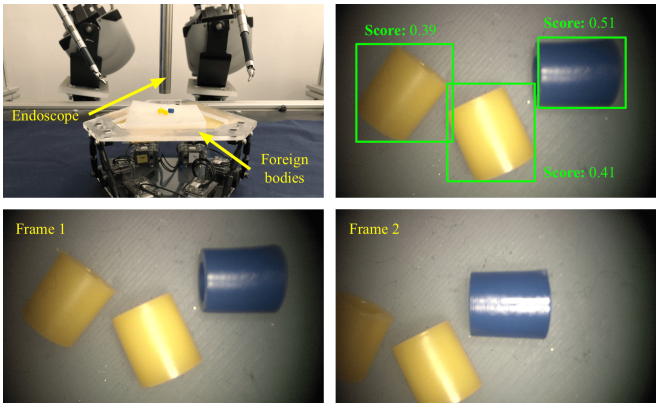

Fig. 1: We construct a debridement task with two colored Ethicon Endo-Surgery phantom "foreign bodies" placed on a planar silicone slab mounted on the SPRK. A surgical demonstrator is instructed to center the camera on the blue foreign body, an example of which is shown in the bottom two frames. The learning from demonstration (LfD) method computes scores for each foreign body and learns a model that captures this preference and synthesizes a viewpoint for a novel scene.

Experimentally, we emulate a movable camera with a physically fixed camera setup and use the 6 DoF SPRK to translate the entire workspace [8]. Existing surgical robotic cameras, such as the one used by the Intuitive Surgical's da Vinci, are cable-driven and consequently have imprecise kinematics [9, 10]. The SPRK has a very particular kinematic chain structure, namely, that the inverse kinematics are precise

Authors are affiliated with:
[1]The AUTOLAB at UC Berkeley; `autolab.berkeley.edu`
[2]UC San Francisco East Bay; `eastbay.surgery.ucsf.edu`

and trivial to compute even if the forward kinematics are non-linear and challenging to reason about. As a result of this structure, the platform can be precisely servoed to a position in the world frame, allowing us to collect clean, consistent demonstration data. We consider a surgical debridement scenario on silicone phantoms with training foreign bodies that vary in color and shape. In each experiment, we categorize a consistent set of these bodies as important, based on color, shape, size, or a combination. It is the surgeon's task to center the camera over all of the desired foreign bodies. For instance, an example task is to find all of the blue foreign bodies, or find all of the circular ones. The surgeon's demonstrations actuate the camera to center on those objects.

II. Related Work and Background

One of the seminal projects in robotic laparoscopic camera control is the Automated Endoscopic System for Optimal Positioning (AESOP) at the Johns Hopkins Medical Center [11]. This system provided the surgeon with both a foot-pedal camera control interface as well as a voice command interface. The study concluded that voice commands and tele-operative interfaces at the time were inaccurate and increased automation in camera control was desired. Since then, the community has developed new interfaces such as the EndoAssist device, which is controlled by surgeon head movements [12], and has studied software-based camera motion automation (see surveys [6]). Automatic camera control has also been studied widely in computer graphics [13–15]. In robotics and automation, it has been studied as a form of active vision or active perception [16, 17]. We focus on a specific subproblem of learning viewpoint selection from expert demonstrations.

A. Surgical Camera Movement

Automation strategies can be broadly taxonomized into two groups: (1) reactive, where the camera's motion is triggered by tool motion, and (2) predictive, where an algorithm anticipates future surgeon motion and moves the camera accordingly.

Several autonomous camera systems for minimally invasive surgery have been constructed that use sets of rules to calculate a camera target position and zoom level. For example, Eslamian et al. [18] proposed a heuristic for automated camera movement where the camera tracks the midpoint of a two-arm surgical robot. There are similar surgical approaches that reactively move the camera in response to surgeon's eye-gaze [19, 20]. Both instrument tracking and eye-gaze tracking are important primitives in automated camera movement. However, they neglect longer horizon planning.

Weede et al. [21] proposed a system that applies a Markov model to predict surgeon tool movements and center the camera accordingly on the anticipated end-effector midpoint. In contrast to prior work, we focus on the viewpoint selection problem based on anatomical features, not the tool positions. These viewpoints are learned from demonstrations. To the best of our knowledge, there is limited work in viewpoint selection for endoscopic cameras. The closest work has been in the context of medical imaging with MRIs [22].

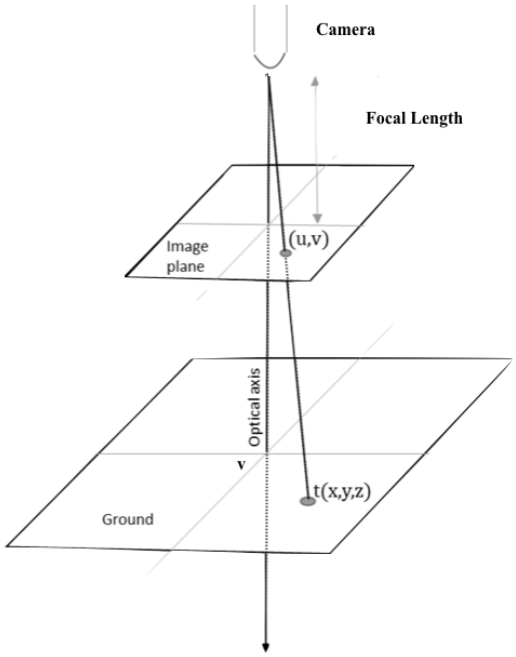

Fig. 2: A viewpoint is the intersection of the optical axis and the planar workspace.

B. Viewpoint Selection

The viewpoint selection problem has been studied in other research areas [23–27]. Much of the work considers accurate 3D models, or registration, of an object of interest. We approach the problem with a model-free approach, where the regions of interest are learned from demonstrations. This is related to work on Region Proposal Networks (RPN), proposed by Ren et al. [28], which combines convolutional features with object detection networks to segment and predict object bounds. In addition, the problem is very related to camera placement and sensor placement problems [29–31]. The area is also termed visual attention modeling in the computer vision community [32]. Our work specifically focuses on surgical hardware (an endoscopic camera) and a surgical task (debridement).

III. Problem Definition

Let $\mathcal{O}_b$ define a global coordinate frame in $\mathcal{SE}(3)$. Let m define a planar workspace in $\mathcal{O}_b$. We consider an ideal pinhole camera model. Let $\vec{o}_a$ be the optical axis of this camera, i.e., a ray originating at the lens. The oriented point at which the optical axis intersects the plane m is denoted by $v \in \mathcal{O}_b$. We call v a *viewpoint*. This corresponds to an ideal pinhole camera model with no distortions. The geometry of this relationship is described in Figure 2.

A. Kinematic Assumptions

We assume that the camera has inverse kinematic mappings. Given a viewpoint v, one can analytically calculate a set of

joint angles to actuate the camera such that the optical axis intersects the plane at v:

$$u = f^{-1}(v)$$

Generally, this kinematic mapping will be redundant since the points lie in a plane. We assume that there is a consistent technique to resolve the redundancies.

B. Problem Definition

Given a set of "candidate" viewpoints $V = \{v_1, ..., v_k\}$, we seek to assign a score to each viewpoint in terms of its value to a particular task. We learn this scoring function from expert demonstrations.

A demonstration trajectory is a sequence of 2D planar motions $x_i \in SE(2)$ represented as rigid transformations to the workspace:

$$d = [\, x_1, ..., x_T \,]$$

In the ideal pinhole camera model, a translation of the camera corresponds to a translation of the workspace and a rotation of the camera corresponds to a rotation of the workspace. We consider a planar debridement task where there are some true foreign bodies to remove from the phantom and spurious ones that must be left alone. The goal is to learn which foreign bodies are indeed candidates for debridement and automatically synthesize the camera movement to focus on those bodies. From these demonstrations, we infer a model $\pi_\theta : V \mapsto [0, 1]$ that ranks a set of candidate viewpoints based on the likelihood that an expert would have selected it.

IV. Probabilistic Viewpoint Learning

Given a set of demonstrations, our algorithm fits a model that ranks candidate viewpoints to optimally match behavior observed in the expert demonstrations.

A. Region Proposals

First, the segmentation algorithm coarsely segments the initial image into a set of candidate closed contours in the image. This segmentation is designed to over-predict bounding boxes and is implemented with standard OpenCV filters[1]. We first generate a set of thresholded images with different HSV thresholds. On each of these images, we use a Canny edge detector to identify contours after the filters [33]. A bilateral filter is used to de-noise the detected contours. All remaining closed contours are added to a set of candidates, which we prune for containment and size thresholds.

Each of these bounding boxes defines a candidate viewpoint v_i. The centroid of each contour can be translated into the base coordinate frame. Known movements of the camera can be translated to known movements of the centroids under the ideal pinhole camera model.

[1] https://opencv.org/

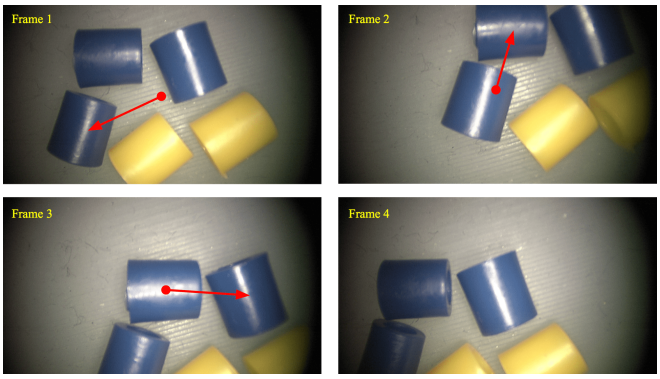

Fig. 3: Three blue and two yellow objects of the same shape and size are placed on a planar silicone slab mounted on the SPRK within view of the camera. The example trajectory above progress from left to right and top to bottom, beginning with the top left initial view and centering on each blue inclusion. The red arrows denote the direction of camera movement from the current frame to the next.

B. Tracking Model

We use this insight to analyze the camera movement demonstrations along their trajectories (Figure 3). Each demonstration is a sequence of rigid transformations:

$$d = [\, x_1, ..., x_T \,]$$

Each of these transformations can be applied to the centroids to get a trajectory of that contour throughout the demonstration:

$$v_i = [\, x_1 \cdot v_i, ..., x_T \cdot v_i \,]$$

For each demonstration and each candidate viewpoint in the demonstration, we get a 2D spatial trajectory of camera focal points:

$$c = [\, c_1, ..., c_T \,]$$

In addition to the spatial trajectories, we also record the clock time between transformations.

C. Scoring Model

The goal of the scoring model is to quantify the value of each of the candidate viewpoints. We do so by comparing the geometric relationship of each trajectory of these viewpoints to the trajectory of camera focal points $c[t]$. For a candidate viewpoint i, let δ_t be defined as:

$$\delta_i = \|v_i[t] - c[t'']\|_2$$

The *instantaneous focusing effort* is defined as the ratio between the displacement from the focal point at some time t compared to the initial displacement:

$$\gamma_i[t] = \frac{\delta_i[t]}{\delta_i[0]}$$

This ratio is greater than 1 when in the current timestep the candidate viewpoint is further away from the focal point than it initially was, and less than 1 when the candidate

viewpoint is closer. The *maximal focusing effort* is defined as the minimum ratio over all t:

$$\gamma_i^* = \min_t \gamma_i[t]$$

This notion of focusing effort aims to quantify an inclusion's maximum importance or relevance at some point in the trajectory. To account for scaling differences and noise, the final scoring model is a negative exponential of this term:

$$\ell(c, v_i) = e^{-\gamma_i^*}$$

The exponential term acts as a "notch" filter that greatly down-weights viewpoints that are not centered.

1) Note About Filtering Noise: For each camera position we also record dwell time, or the duration of time during which the camera remains static at a position. From the set of camera focal points, we filter out all points with dwell times of less than $k = 5$ seconds, where k is a tunable parameter. This filtering greatly reduces the number of spurious or irrelevant motions in the demonstrations.

D. Predictive Model

Each candidate viewpoint is now scored with a label $\ell(c, v_i)$ that indicates how well it was centered in the demonstration. The predictive model attempts to correlate this score with image properties. In other words, given a new scene, the model anticipates which candidate viewpoints an expert might center the camera on.

Each bounding box is featurized by the image patch that it bounds. We featurize the content in each viewpoint with color frequency and contour properties f_i. Using OpenCV contour approximation, which can be applied to regular and irregular shapes, we approximate a contour shape for each bounding box with $\varepsilon = 0.005$ where ε is a tunable parameter representing the maximum distance from the true contour to its approximated contour. We swept over the epsilon parameters and chose the value that resulted in the closest visual approximations. Each bounding box is featurized by its area, perimeter, height-width ratio, and average angle of its approximated contour shape. We also include features that count the number of pixels whose color is dominated by red, blue, and green respectively.

In all of our experiments, we use a robust linear regression model, also called an elastic-net model8:

$$\pi_\beta = \arg\min_\beta \sum_{v_i \in V} \|\beta^T f_i - \ell(c, v_i)\|_2^2 + \lambda \|\beta\|_2 + \alpha \|\beta\|_1$$

This results in a model π that can score a given bounding box based on the value anticipated from experts demonstrations.

V. Demonstration System

In this section, we describe the setup we used to collect demonstrations.

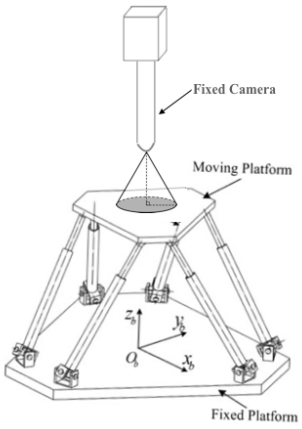

Fig. 4: The SPRK translates the workspace under a fixed endoscope camera.

A. Fixed Camera Moving Workspace

Instead of moving the camera, we emulate idealized camera motions by translating the entire workspace with the SPRK [8]. As in Figure 4, the SPRK consists of two platforms one fixed and one moving. The moving upper platform defines a plane in m in $\mathcal{O}_b$. Above the platform, we place a fixed endoscope camera.

We integrated this system with the dVRK. The dVRK is a development platform provided by Intuitive Surgical to advance research in teleoperated robotic surgical systems. It consists of hardware from the first-generation da Vinci "classic" and open-source electronics and software developed by WPI and Johns Hopkins University. The robot hardware consists of two robotic laparoscopic arms, termed *Patient Side Manipulators* (PSMs), and the Surgeon Console for teleoperating with a stereo viewer, two master controllers, termed "Master Tool Manipulators" (MTMs), and a set of foot pedals.

In classical da Vinci setups, the MTMs control both arm and camera movement. The surgeon presses down on a foot pedal to switch between camera and arm control. This interface is desirable, since the endoscopic camera is typically mounted to a standard da Vinci arm. The system calculates the change in the surgeon's hand position from some designated start position and applies the same pose transformation to the camera lens tip.

B. Masters To SPRK Connection

We emulate this interface with a fixed camera and a moving SPRK. Pressing the camera foot pedal triggers a message that activates the moving platform. One of the critical differences in moving the platform rather than actuating an arm is that rather than controlling the camera lens tip, we actuate the platform. The SPRK has a kinematic chain structure such that the inverse kinematics are trivial to evaluate, but the forward kinematics are non-linear and challenging to compute. Accordingly, given the relative changes in hand pose from the right MTM (Figure 5), we can process these changes into

absolute positions for the SPRK that effect the same delta on the platform, up to a configurable scaling parameter.

Mirroring classical da Vinci setups, when the camera pedal is released, streaming of poses to the SPRK is paused, and the camera viewpoint remains static. The surgeon is free to re-position the right MTM for ergonomic purposes or use the arm in other surgical tasks.

We use the procedure below to map relative poses from the right MTM into absolute SPRK positions. We denote $T_{i,j}$ as a pose from frame i to frame j and define the following:

- *mw* - World frame of the right MTM. Poses for the right MTM are interpreted with respect to this frame.
- *mi* - Frame of the initial right MTM pose.
- *mc* - Frame of the current right MTM pose.
- *cr* - Frame of the right MTM at most recent camera pedal release.
- *cp* - Frame of the right MTM at most recent camera pedal press.
- *sw* - World frame of the SPRK. Poses for the platform are interpreted with respect to this frame.

We use $T_{mi,cp}$ and $T_{mi,cr}$ in our pose calculations to ensure that all MTM pose changes executed while the camera pedal is released have no effect on the SPRK position. Both $T_{mi,cp}$ and $T_{mi,cr}$ are initialized to the identity transformation. Given a new right MTM pose $T_{mw,mc}$, with every camera pedal press, we perform the following update:

$$T_{mi,cp} \leftarrow T_{mi,cp} T_{cp,mw} T_{mw,cr}$$
$$T_{mw,cp} \leftarrow T_{mw,mc}$$

With every camera pedal release, we execute this update:

$$T_{mw,cr} \leftarrow T_{mw,mc}$$

For a given new published MTM pose $T_{mw,mc}$ not associated with a camera pedal press or release, we find $T_{sw,mc}$ by evaluating:

$$T_{sw,mc} = T_{sw,mw} T_{mw,mi} T_{mi,cp} T_{cp,mw} T_{mw,mc}$$

where $T_{sw,mw}$ is a 180° rotation. The resulting pose $T_{sw,mc}$ is scaled according to operator preference and comfort and sent to the SPRK. This formulation is based on our prior work [34].

C. System Parameters

We characterize the visual field that the SPRK is capable of supporting. The endoscope camera supports stereo 1920x1080 images–corresponding to a 25 mm x 16 mm field of view. The range of motion of the SPRK corresponds to 17 mm in both dimension, effectively tripling the field of view with movement.

Field of View With and Without SPRK

Field of View (no movement)	448 mm^2
Effective Field of View	1386 mm^2

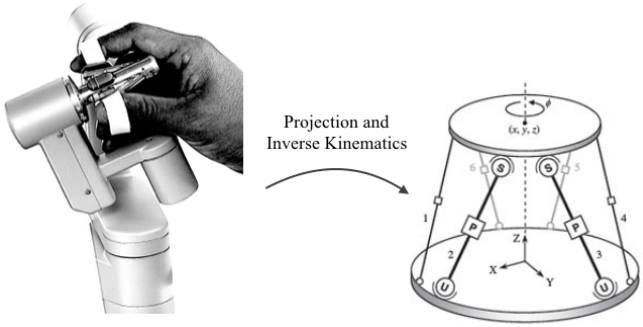

Fig. 5: The emulated camera movement system fully integrates with the da Vinci research kit master manipulators and foot pedal system.

VI. Experiments

We evaluate our algorithm on three planar viewpoint selection tasks in which we learn the characteristics of desired foreign bodies to debride. In our experiments, we measure both the error in learning the scoring metric and contextualize this learning error in terms of hand-labeled ground truth. We present multiple accuracy metrics due to the inherent sensitivity of computer vision algorithms to specularity. Due to the specularity, some single objects can misclassified as two objects with two bounding boxes, each defining a candidate viewpoint. Therefore, we look at both the number of candidate viewpoints ranked properly, even if there are multiple, as well as the number of candidate viewpoints that correspond to single objects.

Mean Squared Error: We first measure how well the model predicts our scoring metric (the maximal focusing effort) from features of image. This quantifies the extent to which the learning model can infer relevance purely from the image features.

Box Accuracy: Bounding box accuracy is measured by the percentage of foreign objects in a demonstration with accurate bounding boxes, averaged over the total number of demonstrations in the test set. An object's viewpoint is accurate if the object has one unique bounding box that encompasses the entire object.

Rank Accuracy: We measure rank accuracy as the percentage of demonstrations with the correct relative ranking of predicted scores among the viewpoints in the workspace. The relative ranking is critical for performance on real world tasks. For instance, in a task such as debridement, we would expect scores of tissue to be removed to be higher than that of healthy tissue and expect the camera to center on the higher-scored tissue.

A. Phantom Setups

The learning tasks are inspired by debridement scenarios on silicone tissue phantoms. A roughly planar silicone slab is mounted on the SPRK with phantom foreign bodies that vary in terms of color, size, and shape (Figure 1). The materials for this experiment are taken from the Ethicon

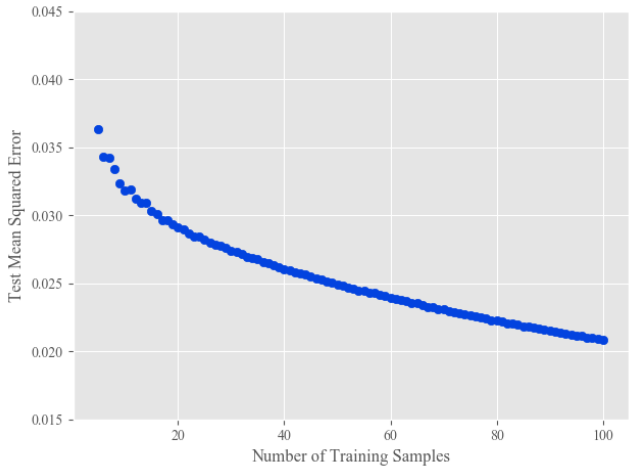

Fig. 6: Experiment 1 test error on a consistent held-out dataset of 25 candidate objects for varying training set sizes. The figure uses a statistical re-sampling estimator to illustrate the variance in the accuracy, where the model is retrained on random subsets of the data and averaged.

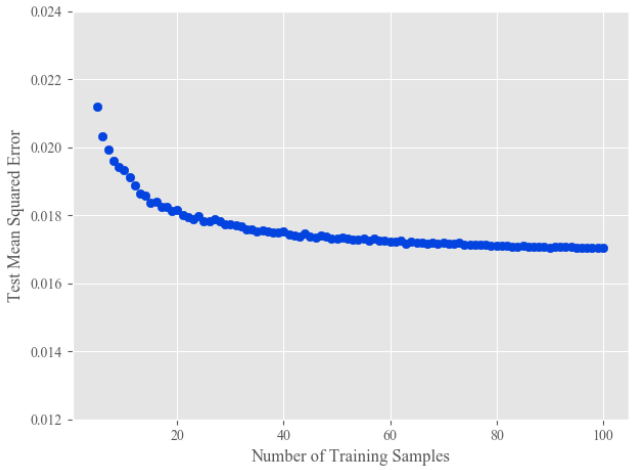

Fig. 7: Experiment 2 test error on a consistent held-out dataset of 25 candidate objects for varying training set sizes.

Endo-Surgery Inc. training kit and are standard laparoscopic training materials. In each experiment, we classify a uniform set of these phantom objects as important, based on color, shape, size, or a combination thereof. It is the surgeon's task to center the camera over all of the important bodies in the task. The surgeon's demonstrations actuate the emulated movable camera to center on those deemed important.

B. Experiment 1. Viewing One Colored Foreign Body using Single Camera Movement

In the first experiment, we considered the task of centering on a single foreign body. We placed one blue and one to two yellow objects of the same size and shape on the phantom within the view of the camera (Figure 1). The demonstrator was instructed to center on the blue object.

We used the following parameters for the regression model: $\lambda = 0.48, \alpha = 0$. We swept over model parameters and selected the values that achieved the lowest mean squared error on a held-out randomized set comprising 20% of all collected observations. The model was trained on a set of 100 candidate viewpoints and tested on a set of 25 candidate viewpoints. The training and test errors given varying training set sizes is illustrated in Figure 6. Using the full training set yields 0.021 training mean squared error and 0.021 test mean squared error. With respect to hand labeled ground truth, the system achieved:

Experiment 1: Accuracy

Box Accuracy	Rank Accuracy
80%	100%

C. Experiment 2. Viewing Multiple Colored Foreign Bodies from Camera Movement Trajectory

In the next experiment, we considered an extension of the previous task to center on multiple blue objects. In this experiment, one to four blue and one to four yellow objects were placed on the phantom (Figure 3). The demonstrator moved the camera along a trajectory and centered on each blue object.

We applied the robust linear regression model with $\lambda = 1.33, \alpha = 0$. We swept over model parameters and selected the values that achieved the lowest mean squared error on a held-out randomized set comprising 20% of all collected observations. The model was trained on a set of 100 candidate viewpoints and tested on a held-out set of 25 candidate viewpoints. The training and test errors given varying training set sizes is illustrated in Figure 7. Using the full training set yields 0.016 training mean squared error and 0.017 test mean squared error. With respect to hand labeled ground truth, the system achieved:

Experiment 2: Accuracy

Box Accuracy	Rank Accuracy
70.8%	100%

D. Experiment 3. Viewing One Circular Foreign Body using Single Camera Movement

The previous experiments illustrate the technique for selecting the foreign body by color. The goal of this experiment was to use demonstrations to learn a model for scoring candidate viewpoints containing objects of different configurations. We placed one circular and one to three rectangular objects of any color (yellow, blue, or orange) on the phantom within the view of the camera (Figure 8). The demonstrator was instructed to center on the circular object.

We applied the robust linear regression model with $\lambda = 0, \alpha = 10^{-25}$. We swept over model parameters and selected the values that achieved the lowest mean squared error on a held-out randomized set comprising 20% of all collected observations. The model was trained on a set of 100 candidate viewpoints and tested on a held-out set of 25 candidate viewpoints. The training and test errors given varying training

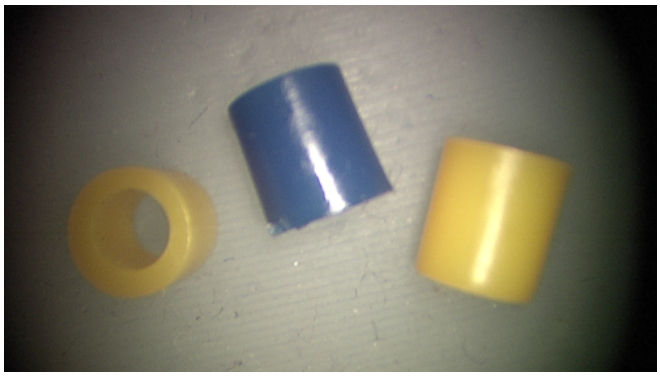

Fig. 8: One circular object is placed on a planar silicone slab among two rectangular objects and mounted on the SPRK within the view of the camera.

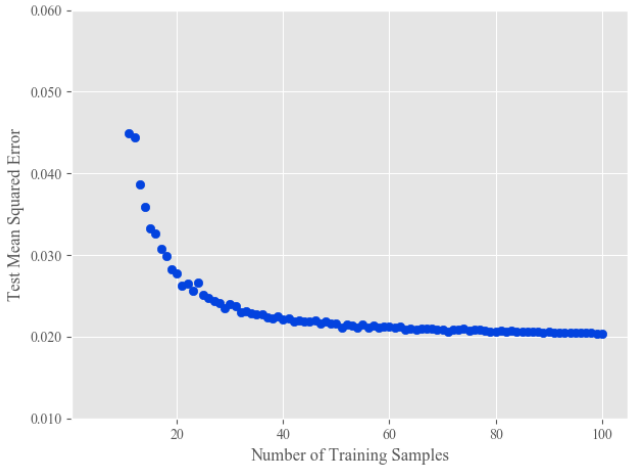

Fig. 9: Experiment 3 test error on a consistent held-out dataset of 25 candidate objects for varying training set sizes.

set sizes is illustrated in Figure 9. Using the full training set yields 0.021 training mean squared error and 0.021 test mean squared error. With respect to hand labeled ground truth, the system achieved:

Experiment 3: Accuracy

Box Accuracy	Rank Accuracy
70.5%	90%

VII. Discussion and Future Work

We explored learning from endoscopic images viewed during demonstrations performed by experts to yield a predictive model that suggests camera positions at the beginning of a procedure. We address limitations and several avenues of future work to consider.

First, we consider a limitation demonstrated by two failure modes illustrated in Figure 10. We encountered general bounding box imprecision when filtering with OpenCV as well as bounding box and contour approximation errors due to specularities. For a small subset of configurations, the light source of the camera, even on the lowest brightness setting,

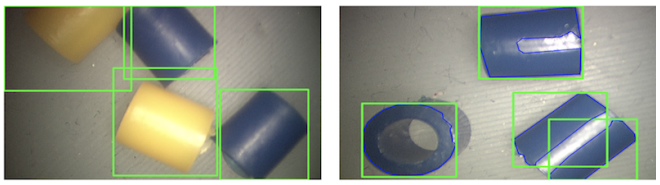

Fig. 10: Two failure modes encountered during image segmentation. The left figure depicts general bounding box imprecision when filtering with OpenCV. The right figure illustrates bounding box and contour approximation errors as a result of light specularities.

caused one inclusion to appear as two. The split bounding boxes consequently resulted in erroneous predictions. We believe that different light sources, more precise filtering through OpenCV, and the use of region proposal networks could improve our current segmentation and the model's accuracy.

Second, we believe the current featurization limits the robustness of this learning model. We featurized inclusions using characteristics such as color, contour approximations, area, and perimeter. The next step towards generalizing the model is to leverage deep neural networks for image featurization as they might better capture more complex features that are difficult or inefficient to feature engineer.

Third, this work focused on 2D planar camera and object viewpoints. An important challenge for future work is extending the algorithm to $\mathcal{SE}(3)$ and incorporating 3D views of the objects themselves. Adding supplementary cameras and rotating the SPRK can provide the system with the additional freedom needed to more closely emulate the clincial surgical robotic camera on the da Vinci.

Our experiments considered a surgical debridement scenario on silicone phantoms with training inclusions that vary in color and shape, and results suggest that we can learn an accurate model from relatively noisy data. Identifying camera viewpoints from demonstrations has applications for identifying points of interest for initiating other manipulation tasks such as incision closure or anastomosis in the future.

Acknowledgements

This research was performed at the AUTOLab at UC Berkeley in affiliation with the Berkeley AI Research (BAIR) Lab, the Real-Time Intelligent Secure Execution (RISE) Lab, the CITRIS "People and Robots" (CPAR) Initiative, by the Scalable Collaborative Human-Robot Learning (SCHooL) Project, NSF National Robotics Initiative Award 1734633, and in affiliation with UC Berkeley's Center for Automation and Learning for Medical Robotics (Cal-MR). The authors were supported in part by donations from Siemens, Google, Honda, Intel, Comcast, Cisco, Autodesk, Amazon Robotics, Toyota Research Institute, ABB, Samsung, Knapp, Loccioni, and by a major equipment grant from Intuitive Surgical and by generous donations from Andy Chou and Susan and Deepak Lim. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors. We thank our colleagues who provided helpful feedback and suggestions, in particular Brijen Thananjeyan, Carolyn Chen, Jeff Mahler, Daniel Seita, Matthew Matl.

REFERENCES

[1] L. R. Kavoussi, R. G. Moore, J. B. Adams, and A. W. Partin, "Comparison of robotic versus human laparoscopic camera control", *The Journal of Urology*, vol. 154, no. 6, 1995.

[2] W. R. Chitwood Jr, L. W. Nifong, W. H. Chapman, J. E. Felger, B. M. Bailey, T. Ballint, K. G. Mendleson, V. B. Kim, J. A. Young, and R. A. Albrecht, "Robotic surgical training in an academic institution", *Annals of surgery*, vol. 234, no. 4, 2001.

[3] A. W. Partin, J. B. Adams, R. G. Moore, and L. R. Kavoussi, "Complete robot-assisted laparoscopic urologic surgery: A preliminary report.", *Journal of the American College of Surgeons*, vol. 181, no. 6, 1995.

[4] P. J. Roch, H. M. Rangnick, J. A. Brzoska, L. Benner, K.-F. Kowalewski, P. C. Müller, H. G. Kenngott, B.-P. Müller-Stich, and F. Nickel, "Impact of visual–spatial ability on laparoscopic camera navigation training", *Surgical endoscopy*, vol. 32, no. 3, 2018.

[5] A. Casals, J. Amat, and E. Laporte, "Automatic guidance of an assistant robot in laparoscopic surgery", in *Robotics and Automation (ICRA), 1996 IEEE International Conference on*, IEEE, 1996.

[6] A. Pandya, L. A. Reisner, B. King, N. Lucas, A. Composto, M. Klein, and R. D. Ellis, "A review of camera viewpoint automation in robotic and laparoscopic surgery", *Robotics*, vol. 3, no. 3, 2014.

[7] M. Wilson, J. McGrath, S. Vine, J. Brewer, D. Defriend, and R. Masters, "Psychomotor control in a virtual laparoscopic surgery training environment: Gaze control parameters differentiate novices from experts", *Surgical endoscopy*, vol. 24, no. 10, 2010.

[8] V. Patel, S. Krishnan, A. Goncalves, and K. Goldberg, "SPRK: A low-cost stewart platform for motion study in surgical robotics", *International Symposium on Medical Robotics (ISMR)*, 2018.

[9] J. Mahler, S. Krishnan, M. Laskey, S. Sen, A. Murali, B. Kehoe, S. Patil, J. Wang, M. Franklin, P. Abbeel, and K. Goldberg, "Learning accurate kinematic control of cable-driven surgical robots using data cleaning and gaussian process regression", in *Conference on Automation Science and Engineering*, 2014.

[10] D. Seita, S. Krishnan, R. Fox, S. McKinley, J. Canny, and K. Goldberg, "Fast and Reliable Autonomous Surgical Debridement with Cable-Driven Robots Using a Two-Phase Calibration Procedure", in *International Conference on Robotics and Automation*, 2018.

[11] M. Allaf, S. Jackman, P. Schulam, J. Cadeddu, B. Lee, R. Moore, and L. Kavoussi, "Laparoscopic visual field", *Surgical Endoscopy*, vol. 12, no. 12, 1998.

[12] J. Gilbert, "The endoassist™ robotic camera holder as an aid to the introduction of laparoscopic colorectal surgery", *The Annals of The Royal College of Surgeons of England*, vol. 91, no. 5, 2009.

[13] M. Christie, P. Olivier, and J.-M. Normand, "Camera control in computer graphics", in *Computer Graphics Forum*, Wiley Online Library, vol. 27, 2008.

[14] L.-w. He, M. F. Cohen, and D. H. Salesin, "The virtual cinematographer: A paradigm for automatic real-time camera control and directing", in *Conference on Computer graphics and interactive techniques*, ACM, 1996.

[15] C. Ware and S. Osborne, "Exploration and virtual camera control in virtual three dimensional environments", *SIGGRAPH computer graphics*, vol. 24, no. 2, 1990.

[16] S. Chen, Y. Li, and N. M. Kwok, "Active vision in robotic systems: A survey of recent developments", *International Journal of Robotics Research*, vol. 30, no. 11, 2011.

[17] R. Bajcsy, "Active perception", *Proceedings of the IEEE*, vol. 76, no. 8, 1988.

[18] S. Eslamian, L. A. Reisner, B. W. King, and A. K. Pandya, "An autonomous camera system using the da vinci research kit",

[19] G. P. Mylonas, A. Darzi, and G. Zhong Yang, "Gaze-contingent control for minimally invasive robotic surgery", *Computer Aided Surgery*, vol. 11, no. 5, 2006.

[20] S. Ali, L. Reisner, B. King, A. Cao, G. Auner, M. Klein, and A. Pandya, "Eye gaze tracking for endoscopic camera positioning: An application of a hardware/software interface developed to automate aesop.", *Studies in health technology and informatics*, vol. 132, 2008.

[21] O. Weede, H. Mönnich, B. Müller, and H. Wörn, "An intelligent and autonomous endoscopic guidance system for minimally invasive surgery", in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, IEEE, 2011.

[22] K. Mühler, M. Neugebauer, C. Tietjen, and B. Preim, "Viewpoint selection for intervention planning.", in *EuroVis*, 2007.

[23] F. Deinzer, J. Denzler, and H. Niemann, "Viewpoint selection–planning optimal sequences of views for object recognition", in *International Conference on Computer Analysis of Images and Patterns*, Springer, 2003.

[24] G. Leifman, E. Shtrom, and A. Tal, "Surface regions of interest for viewpoint selection", in *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012.

[25] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich, "Viewpoint selection using viewpoint entropy.", in *VMV*, vol. 1, 2001.

[26] F. Arai, T. Sugiyama, P. Luangjarmekorn, A. Kawaji, T. Fukuda, K. Itoigawa, and A. Maeda, "3d viewpoint selection and bilateral control for bio-micromanipulation", in *International Conference on Robotics and Automation*, IEEE, vol. 1, 2000.

[27] Y. Motai and A. Kosaka, "Hand–eye calibration applied to viewpoint selection for robotic vision", *IEEE Transactions on Industrial Electronics*, vol. 55, no. 10, 2008.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[29] B. Triggs and C. Laugier, "Automatic camera placement for robot vision tasks", in *Robotics and Automation (ICRA), 1995 IEEE International Conference on*, IEEE, vol. 2, 1995.

[30] S. Sakane and T. Sato, "Automatic planning of light source and camera placement for an active photometric stereo system", in *Robotics and Automation (ICRA), 1991 IEEE International Conference on*, IEEE, 1991.

[31] X. Chen and J. Davis, "Camera placement considering occlusion for robust motion capture", *Computer Graphics Laboratory, Stanford University, Tech. Rep*, vol. 2, no. 2.2, 2000.

[32] A. Borji and L. Itti, "State-of-the-art in visual attention modeling", *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, 2013.

[33] J. Canny, "A computational approach to edge detection", in *Readings in Computer Vision*, Elsevier, 1987.

[34] J. Liang, J. Mahler, M. Laskey, P. Li, and K. Goldberg, "Using dvrk teleoperation to facilitate deep learning of automation tasks for an industrial robot", in *Conference on Automation Science and Engineering*, 2017.