Learning to Localize, Grasp, and Hand Over Unmodified Surgical Needles

Albert Wilcox*, Justin Kerr*, Brijen Thananjeyan, Jeff Ichnowski, Minho Hwang, Samuel Paradis, Danyal Fer, Ken Goldberg

* equal contribution

Abstract—Robotic Surgical Assistants (RSAs) are commonly used to perform minimally invasive surgeries by expert surgeons. However, long procedures filled with tedious and repetitive tasks such as suturing can lead to surgeon fatigue, motivating the automation of suturing. As visual tracking of a thin reflective needle is extremely challenging, prior work has modified the needle with nonreflective contrasting paint. As a step towards automation of a suturing subtask without modifying the needle, we propose HOUSTON: Handoff of Unmodified, Surgical, Tool-Obstructed Needles, a problem and algorithm that uses a learned active sensing policy with a stereo camera to localize and align the needle into a visible and accessible pose for the other arm. To compensate for robot positioning and needle perception errors, the algorithm then executes a high-precision grasping motion that uses multiple cameras. In physical experiments using the da Vinci Research Kit (dVRK), HOUSTON successfully passes unmodified surgical needles with a success rate of 96.7% and is able to perform handover sequentially between the arms 32.4 times on average before failure. On needles unseen in training, HOUSTON achieves a success rate of 75 – 92.9%. To our knowledge, this work is the first to study handover of unmodified surgical needles. See https://tinyurl.com/houston-surgery for additional materials.

I. INTRODUCTION

Robotic Surgical Assistants (RSAs) currently rely on human supervision for the entirety of surgical tasks, which can consist of many very repetitive subtasks such as suturing. Automation of surgical subtasks may reduce surgeon fatigue [37], with initial results in surgical cutting [23, 34], debridement [23, 30], suturing [7–10, 28, 31–33], hemostasis [26], and peg transfer [12, 13, 24, 27].

This paper considers automation of the bimanual regrasping subtask [7] of surgical suturing, which involves passing a surgical needle from one end effector to another. This handover motion is performed in between stitches during suturing, and is a critical step, as accurately positioning the needle in the end effector affects the stability of its trajectory when guided through tissue. Because varying cable tension in the cables driving the arms causes inaccuracies in motions, a high precision task such as passing a needle between the end effectors is challenging [12, 13, 21, 24, 25].

The task is also difficult because 3D pose information is critical for successfully manipulating needles, and surgical needles are challenging to perceive with RGB or active depth sensors due to their reflective surface, thin profile [17], and self-occlusions (Figure 2). Prior work has mitigated this by painting needles [7, 31] and using color segmentation, but this solution is not practical for clinical use.

To manipulate unmodified surgical needles, we combine recent advances in deep learning, active sensing, and visual servoing. We present HOUSTON: Handoff of Unmodified, Surgical, Tool-Obstructed Needles, a problem and algorithm for using stereo vision with coarse and fine-grained control policies (Figure 1) to sequentially localize, orient, and handover unmodified surgical needles. We present a localization method using stereo RGB with a deep segmentation network to output a point cloud of the needle in the workspace. This point cloud is used to define a coarse robot policy that uses visual servoing to reorient the needle towards the grasping arm, and then refines this iteratively to correct for positioning errors. Finally, it executes a learned visual servoing grasping policy to complete the handover.

This paper makes the following contributions:

1) A perception pipeline using stereo RGB to accurately estimate the pose of surgical steel needles in 3D space, enabling needle manipulation without active depth
sensors and painted needles.

2) A visual servoing algorithm to perform coarse reorientation of a surgical needle for grasping.

3) A needle grasping policy that performs fine control of the needle learned from a small set of human demonstrations to compensate for robot positioning and needle sensing inaccuracies.

4) Combination of the pose estimator (1), the servoing algorithm (2) and needle controller (3) to perform bimanual surgical needle regrasping, where physical experiments on the da Vinci Research Kit (dVRK) [15] suggest a success rate of 96.7% on needles used in training, and 75 – 92.9% on needles unseen in training. On sequential handovers, HOUSTON successfully executes 32.4 handovers on average before failure.

II. RELATED WORK

A. Automation in Surgical Robotics

Automation of surgical subtasks is an active area of research with a rich history. Prior literature has studied automation of tasks related to surgical cutting [19, 23, 34], debridement [16, 23], hemostasis [26], peg transfer [12, 13, 24], and suturing [7, 9, 28, 31–33]. While automated suturing has been studied in prior work [28, 31], suturing without modifications such as painted fiducial markers is an open research problem. Recent work studies robust and general approaches to specific subproblems within suturing, including the precise manipulation of surgical needles during suturing from needle extraction [32] to bimanual regrasping [7], which is the focus of this work.

Needle manipulation is also studied by [32], where the approach studies the extraction of needles from tissue phantoms to compute robust grasps of the needle even in self-occluded configurations. Bimanual needle regrasping was studied in detail by Chiu et al. [7], with impressive results on simulation-trained policies that take needle end effector poses as input. We extend their problem definition to consider multiple handoffs of unmodified needles and end effectors, which requires perception of the needle and robot pose from images without color segmentation. Needle grasping has also been studied using visual servoing policies in D’Ettorre et al. [9], where the needle is painted with green markers to track its position during closed loop visual servoing. Varier et al. [35] study tabular RL policies for needle regrasping in a discretized space in a fixed setup with known needle pose and experiments without the needle on the dVRK. The experiments in [35] suggest that value iteration-trained policies can mimic expert trajectories used for inverse reinforcement learning. In contrast, we present an algorithm compatible with significantly varying initial needle and gripper poses using only image observations. We additionally present many physical experiments with a needle, evaluating the success rate and speed of the algorithm.

B. Visual Servoing, and Active Perception

Visual servoing (VS) is a popular technique in robotics [11, 18], and has recently been applied to compensate for surgical robot imprecision in the surgical peg transfer task [24]. While classical VS approaches typically make use of hand-tuned visual features and known system dynamics [5, 6], recent work proposes learning end-to-end visual servoing policies from examples [14, 20]. To reduce the need for tuned features and dynamics models and also reduce the number of training samples required to create a robust VS policy for bimanual needle regrasping, we present a hybrid approach that combines coarse motion planning with fine control, where a learned VS policy is only used in parts of the task where high precision is required. This framework, called intermittent visual servoing (IVS), was studied in detail by [24], where the system switches between a classical trajectory optimizer and imitation learning VS policy based on the precision required at the time. Inspired by this technique, we present an IVS approach to bimanual needle regrasping, that combines coarse perception and planning with fine VS control. Because this task requires reasoning about depth across several directions, we present a multi-view VS policy that learns to precisely hand over the needle based on several camera views.

Active perception is a popular technique with many variations to localize objects prior to manipulation by maximizing information gain about their poses [1, 2, 22, 29, 36]. This has been studied in the context of robot-assisted surgery, where the endoscope position is automatically adjusted via a policy learned from demonstrations to center the camera focus on inclusions during surgeon teleoperation. In this work, we actively servo the needle to highly visible poses to maximize the accuracy of its pose estimate. This is most similar to [1], where the authors propose an algorithm to actively select views of a grasping workspace to uncover enough information about unknown objects to plan grasps.

III. PROBLEM FORMULATION

The HOUSTON problem extends and generalizes the previous problem definition from Chiu et al. [7] to include unmodified needles, occlusion, and multiple handoffs.

A. Overview

In the HOUSTON problem, the surgical robot starts with a curved surgical needle with known curvature and radius grasped by one gripper and must accurately pass it to the other gripper and back. This is challenging due to the needle’s reflective surface, thin profile, and pathological configurations [7, 32] as depicted in Figure 2. Once the needle is successfully passed to the other end effector, it is passed back to the first end effector. This process is repeated $N_{\text{max}}$ times, or until the needle is dropped. We also consider a special case of this problem, the single-handover version, in which $N_{\text{max}} = 1$.

B. Notation

Let $p_L(t)$ and $p_R(t)$ denote the poses of the left and right grippers, respectively, at discrete timestep $t$ with respect to a world coordinate frame. The needle has pose $p_N(t)$ with respect to the world frame. Observations of the workspace are available via RGB images from a stereo camera, $I_L(t)$.
In order to deterministically evaluate HOUSTON policies in a wide variety of needle configurations, we discretize the needle-in-gripper pose possibilities by choosing a number of categories across three degrees of freedom:

1) The needle’s curve can face either towards or away from the camera, providing 2 possibilities.
2) The gripper may hold the needle either at the tip, or 30° inwards following the curvature of the needle. This degree of freedom has 2 possibilities.
3) The rotation of the needle about the tangent line \( \omega \) to the point the gripper intersects ranges from 0° to 180°, and is discretized into 7 bins as in Figure 3.

This gives a total of 28 possible needle configurations and we perform a grid search over these possibilities. We chose these configurations to be representative of those seen in suturing tasks post-needle extraction.

While the robot encoders provide an estimate of the gripper poses \( p_L(0) \) and \( p_R(0) \), the precise needle pose is unknown due to cabling effects of the arms. We assume access to a stereo RGB pair in the workspace, an overhead RGB camera, and the transforms between the coordinate frames of these cameras and the robot arms.

C. Assumptions

In the first phase, the algorithm repositions the needle to a pose where the other arm can easily grasp it without collisions and the cameras can clearly view it. Throughout execution of the coarse policy, we parameterize the needle as a circle of known radius, and measure its state in world frame as a center point \( c_p \), normal vector \( c_n \), and a needle tip point \( c_t \) as shown in Figure 6. Active Needle Presentation consists of two stages: needle acquisition and handover positioning.

The needle acquisition stage moves the needle to maximize visibility, and the positioning stage uses visual servoing to move the needle into a graspable state.

1) Needle State Estimation: The state estimator passes stereo images into a fully convolutional neural network that is trained to output segmentation masks for the needle in each image. See the project website for architecture details. It computes a distance transform of the segmentation mask to label each pixel with its distance to the nearest unactivated pixel. Next, it finds peaks in the distance transform along horizontal lines in each image, which correspond to points near the center of activated patches. It then triangulates all pairs of peaks along each horizontal line in the images to obtain a point-cloud as in Figure 4. Because this may triangulate outlier points from the gripper or incorrectly match points on different parts of the needle, RANSAC is applied to filter out incorrect point correspondences and returns the final predicted needle state. At each iteration, it samples a set of 3 points, to which a plane is fit. Each subset of 3 points generates 2 candidate circles in the plane, corresponding to the two circles which pass through one of the pairs. RANSAC uses an inlier radius of 1mm and runs for 300 iterations.

The network is first trained on a dataset of 2000 simulated stereo images of randomly-placed, textured and scaled floating needles and random objects [4] above a surface plane generated in Blender 2.92. Lighting intensity, size, and position, and stereo camera position are also randomized. The segmentation network is fine-tuned on a dataset of 200 manually-labeled images of the surgical needle in the end effector. Training a network on a PC with an NVIDIA V100 GPU takes 3 hours, and fine tuning takes 10 minutes.

2) Visual servoing: HOUSTON uses Algorithm 1 to compute updates for visual servoing in both the needle presentation and the handover positioning phases. Algorithm 1 is a fixed point iteration method that uses a state estimator.
to iteratively visually servo to a target state. Similar to first and second order optimization algorithms, it computes a global update based on its current state and iterates until the computed update is zero. In each iteration, it queries the current 3D state estimate of the needle and then computes an update step in the direction of the target state. To compensate for estimation errors due to challenging needle poses, this process is repeated at each iteration until the algorithm converges to a local optimum within a pose error tolerance.

During needle acquisition, the arm moves to a home pose, then rotates around the world $z$ and $x$ axes, stopping when the state estimator observes at threshold number of inlier points $n_{\text{thresh}} = 20$ from RANSAC circle fitting. During trials, at most 2 consecutive rotations sufficed to resolve the state to this degree. After this initial iteration, we apply Algorithm 1 to align $c_n$ towards the left stereo camera position $p_{\text{left}}^{\text{pos}}$, with $d$ defined as a rotation about the axis $c_n \times (p_{\text{left}}^{\text{pos}} - c_p)$. Once clearly presented to the camera, we measure $c_i$ by choosing the inlier point from the circle fitting step which is furthest from the gripper in 3D space.

Subsequently, during handover positioning, we compute inverse kinematics to move $c_p$ towards the center of the workspace with $c_i$ pointing towards the other gripper and $c_n$ orthogonal to the table plane. This flat configuration is critical for the grasping step, since the dVRK arm is primarily designed for top-down grasps near the center of its workspace. Because the arm only has 5 rotational degrees of freedom, we use a numerical IK solver from [3] and attempt to find a workspace with a small needle is still challenging due to the positioning errors of the robot, which are a result of its cable-driven arms [12, 13, 21, 24, 25, 30]. The policy splits corrective actions along the $x$- and $y$-axes and learn two corresponding policies $\pi_{\text{grasp}, x}$ and $\pi_{\text{grasp}, y}$. Each policy begins by performing an ego-centric crop by projecting the gripper’s kinematically calculated approximate location into the input image and cropping around it. Then, it feeds the cropped image through a neural network to predict a correction direction.

During policy execution, we iteratively sample actions first from $\pi_{\text{grasp}, x}$, then $\pi_{\text{grasp}, y}$, waiting for each to converge
ill before continuing. We multiply the action magnitude by a scalar \( \beta_{\text{decay}} = 0.5 \) every time the network outputs the opposite action of the previous timestep. Servoing terminates when action magnitude decays to under 0.2mm. This enables implicit convergence to the goal without explicitly training the policy to stop. After \( x \) and \( y \) convergence, we execute a simple downward motion of 1cm to grasp the needle.

Algorithm 1 Presentation Visual Servoing Policy

Require: State estimator \( \Phi \), target needle state \( x_N^{\text{large}} \), arm grasping needle \( a \in \{L, R\} \), current time \( t \), number of iterations \( N_{\text{max}} \), tolerance \( \epsilon \), distance metric \( d \).

1: for \( i = 0; i < N_{\text{max}}; i = i + 1 \) do
2: \hspace{10pt} Predict needle state \( \hat{x}_N(t + i) = \Phi(o(t + i)) \)
3: \hspace{10pt} Compute gripper update \( p_\delta = d(\hat{x}_N(t + i), x_N^{\text{large}}) \)
4: \hspace{10pt} Update arm \( a \) pose: \( p_a(t + i + 1) = p_a(t + i) + p_\delta \)
5: \hspace{10pt} if \( |p_\delta| < \epsilon \) then
6: \hspace{20pt} break
7: \hspace{10pt} end if
8: \hspace{10pt} end if
9: end for

V. PHYSICAL EXPERIMENTS

The experiments aim to answer the following question: how efficient and reliable is HOUSTON compared to baseline approaches? We also perform several ablation studies of method components in this section.

A. Baselines

To evaluate the method for the task of active needle presentation, we compare to the following baselines:

- **Depth-based presentation**: Instead of using the stereo RGB network to detect needle pose, we use a depth image-based detection algorithm to detect the needle pose and servo it to the flat grasping pose. This method takes the depth image from the built in depth calculation from the stereo camera, \( I_D(t) \), as input, masks the gripper out of the depth image using the dVRK’s forward kinematics, then performs a volume crop around the end effector and fits a circle of known radius to the points using RANSAC to extract the state.

- **No-Sim-Data**: This is an ablation of the RGB stereo segmentation network that is only trained on the small dataset of real data.

To evaluate the design choices used in the fine-grained grasping policy, we compare to the following baselines:

- **Open Loop**: Executes an open loop grasping motion to grasp the needle based only on needle geometry and inverse kinematics.

- **Shared \((x, y)\) Grasp Policy**: Trains a single policy to output \( x \) and \( y \) displacements, and takes both \( I_O(t) \) and \( I_L(t) \) as input.

To evaluate whether the system can transfer to needles *unseen in training*, we evaluate HOUSTON on three additional needles as in Figure 7.

B. Experimental Setup

We perform experiments using the daVinci Research Kit (dVRK) surgical robot [15], a cable-driven surgical robot with two needle drivers, a gripper which can open 1cm.

For perception, the setup includes a Zed Mini stereo camera angled slightly downwards to face the arms, and an overhead Zivid One Plus M camera facing directly down. Stereo images are captured at 2K resolution, and overhead images are captured at 1080p. Locations of the arms relative to each of the cameras is statically calibrated.

1) **Single handover**: For single handover experiments, we manually vary the orientation of the gripper before each trial to the orientations described in III-C. A handoff is considered successful if the needle switches from one gripper to the other, and at the end is fully supported by the other gripper.

2) **Multiple handovers**: For multiple handover experiments, we start the needle in the left gripper in a visible configuration to the camera, so that all errors are a result of handoffs rather than initialization. We evaluate two configurations: one where the needle arc ends facing the stereo camera in the grasping configuration (Towards), and one where it faces the opposite direction (Away). This configuration is typically maintained throughout each multi-handover trial because of the consistency of the needle presentation step.

C. Single Handover Results

We evaluate HOUSTON and baselines on the single handover task in Table I, and perform multiple systematic passes over the 28 starting configurations described in Section III-C. We find that HOUSTON is able to more reliably perform the task than comparisons, which either experience many presentation errors or many grasp positioning errors.
We categorize all of the failure modes encountered in the workspace. We present results with HOUSTON using three needles (Figure 7) that were unseen in training samples. HOUSTON performs best on the larger two needles (2 and 3), and performs worse on the smaller needle 4, where occlusions with the gripper are more severe.

<table>
<thead>
<tr>
<th>TABLE I: Single Handover Physical Experiments</th>
<th>Success Rate</th>
<th>95% Conf. Int.</th>
<th>Completion Time ($)</th>
<th>Failures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Successes / Total</td>
<td>% Success</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>OPEN LOOP</td>
<td>28/56</td>
<td>50.0</td>
<td>36.3</td>
<td>63.7</td>
</tr>
<tr>
<td>SHARED (x, y) GRASP POLICY</td>
<td>2/28</td>
<td>7.1</td>
<td>0.9</td>
<td>23.5</td>
</tr>
<tr>
<td>NO SIM DATA</td>
<td>53/56</td>
<td>94.6</td>
<td>85.1</td>
<td>98.9</td>
</tr>
<tr>
<td>DEPTH-BASED PRESENTATION</td>
<td>6/28</td>
<td>21.4</td>
<td>8.3</td>
<td>41.0</td>
</tr>
<tr>
<td>HOUSTON (LEFT TO RIGHT)</td>
<td>108/112</td>
<td>96.3</td>
<td>91.1</td>
<td>99.0</td>
</tr>
<tr>
<td>HOUSTON (RIGHT TO LEFT)</td>
<td>108/112</td>
<td>96.3</td>
<td>91.1</td>
<td>99.0</td>
</tr>
</tbody>
</table>

| NEEDLES UNSEEN IN TRAINING                  |               |               |               |           |
| HOUSTON (RIGHT TO LEFT, NEEDLE 2)           | 26/28         | 92.9         | 76.5 | 99.1 | 23.79 ± 2.55     | 1 | 0 | 1  |
| HOUSTON (RIGHT TO LEFT, NEEDLE 3)           | 25/28         | 89.3         | 71.8 | 97.3 | 23.15 ± 3.46     | 3 | 0 | 0  |
| HOUSTON (RIGHT TO LEFT, NEEDLE 4)           | 21/28         | 75.0         | 55.1 | 89.3 | 23.44 ± 4.13     | 5 | 0 | 2  |

| TABLE II: Multiple Handover Physical Experiments |               |               |               |               |           |
| Time per handover (Time)                      |               |               |               |               |           |
| Failure mode (F)                              |               |               |               |               |           |
| NUM | Time | F | NUM | Time | F | NUM | Time | F | NUM | Time | F | NUM | T/H |
| Towards config.                               | 50 26.57 –    | 50 26.76 –    | 50 27.00 –    | 20 25.08 Y    | 23 26.48 P | 38.60 26.56 |

D. Multiple Handover Results

We evaluate HOUSTON on the multiple handover task with \( N_{\text{max}} = 50 \) with two different starting configurations (Table II). We observe that in the first configuration, the algorithm completes 26.20 successful handovers on average and 38.60 in the second. In three trials, no errors occur, and we manually terminate them after 50 successful handovers.

E. Failure Analysis

HOUSTON encounters three failure modes:

- **P**: Presentation error: the robot fails to present the needle in an orientation that is in the plane of the table with the needle tip pointing toward the grasping arm. This may lead to grasping angles that are unreachable or out of the training distribution for the grasping arm.

- **X**: Grasping positioning error (X): the \( x \)-axis grasping policy fails to line up with the needle prior to executing the \( y \)-axis grasping policy.

- **Y**: Grasping positioning error (Y): the \( y \)-axis grasping policy fails line up with the needle prior to grasping.

We categorize all of the failure modes encountered in Table I. We find that the open loop grasping policies are not able to consistently position well for grasping. HOUSTON has failures that are evenly distributed across the failure modes. Grasp policy servoing errors stem mainly from needle configurations that are far outside the distribution seen in training. Presentation phase failures stem primarily from misdetection of the needle true tip, either because of incomplete segmentation masks or from drift in robot kinematics causing the most distal needle point to not be the tip. This causes the servoing policy to rotate the needle away from the camera, after which sometimes it loses visibility and fails to bring the needle to the pre-handover pose. Multi-handoff failures most frequently arise because of subtle imperfections in grasp execution where the needle rotates to a difficult angle. During the subsequent handover the needle can become obstructed by the holding gripper, inhibiting the grasping policy.

VI. DISCUSSION

In this work we present HOUSTON, a problem and an algorithm for reliably completing the bimanual regrasping task on unpainted surgical needles. To our knowledge, this work is the first to study the unmodified variant of the regrasping task. The main limitations of this approach are its reliance on human demonstrations to learn the grasping policy, and sensitivity to needle and environment appearance. We hypothesize that the former could be mitigated via powerful network architectures leveraging stereo disparity such as [17], and designing more autonomous data collection techniques which can label real needle data without human input. In future work, we will also study how to reorient needles between handovers for precise control of needle-in-hand pose and attempt to make needle tracking more robust to occlusions from tissue phantoms.