A Proposal for Monocular Shape Sensing for *Continuum Robot*

Thomas EnXu Li* tli@cs.toronto.edu
Jimmy Chengnan Shentu* cshentu@cs.toronto.edu
Vicky Chaojun Chen* chaojun.chen@mail.utoronto.ca

1. Introduction

Continuum Robot refers to the subcategory of robotic manipulators that do not contain rigid links or identifiable joints. Due to their narrow curvilinear shape, structural compliance, and miniaturization capability, they have been researched for applications involving cluttered environments, such as minimally invasive surgery [3], non-destructive inspection [5], and space/sea exploration [7, 10].

Performing precise motion control of continuum robots requires real-time and accurate shape sensing. Model-based shape-sensing methods are sensitive to unknown external loads, and sensor-based methods take up valuable space in the robots and pose challenges to miniaturization. Thus, we approach visual shape estimation with potential application scenarios in mind – image from a single viewpoint at a time (monocular input) is almost always achievable (e.g., X-ray), but a stereo setup or depth camera may not be available. The purpose of the project is to investigate the feasibility of monocular visual shape estimation for a continuum robot in terms of accuracy and computation time. If successful, the method would efficiently close the control loop thus improving performance for many continuum robot applications.

2. Related Work

2.1. Monocular Depth Estimation

Accurate depth estimation is a fundamental task in many applications including scene understanding and reconstruction. The state-of-the-art methods for depth estimation often involve an encoder-decoder style architecture that estimates the distance of each pixel relative to the camera [12]. Inspired by the development of the Convolution Neural Network (CNN), depth information can now be extracted from monocular images.

Various approaches have been developed for this task. One technique is to leverage transfer learning and use a high-performing pre-trained network as an encoder for feature extraction [1]. This allows the method to achieve a state-of-the-art performance even with a very simple decoder [1]. Another paper by Wong et al. proposed to use an RNN-based method that incorporates LSTM units with convolution layers [11]. This configuration takes advantage of previous images and depth information through recurrent units and thus achieves the best performance when running on a continuous video sequence [11].

2.2. 3D Reconstruction of Continuum Robot

For 3D shape reconstruction of continuum robots, vision-based shape-sensing techniques have been validated to provide more direct and accurate measurement than solely using kinematic modeling [9]. One way to do this is to first extract the robot/needle curve through segmentation and then estimate the 3D points for shape reconstruction using epipolar geometry analysis [2]. Burgner et al. achieved a mean error of 0.473 ± 0.353 mm on an anthropomorphic liver phantom with tumors and vessels [2]. Another method by Dalvand et al. uses a stereo vision system and a 3D reconstruction algorithm based on the closed-form analytical solution for quadratic curve reconstruction in 3D space [4]. This method achieves a maximum measurement error of 0.5 mm for the tip position and length and 0.5 degrees for the bending and orientation angles [4].

While some of the results from previous research are promising in terms of accuracy, their suitability for real-life application is limited by slow speed, the requirement of multiple cameras or input images, and the dependency on tip- or body-mounted markers [4]. Additionally, there is a lack of common hardware or software benchmarks in the field, and some approaches are application specific or use simplifying assumptions to achieve good performance.

3. Method

We start by presenting the problem setup and then explain the proposed approach in detail. Specifically, we propose a two-stage model and the overview of the network is depicted in Figure 1. The model consists of (A) *DepthNet* which predicts pixel-wise depth values from RGB images; (B) *Inverse Projection* module, which takes the predicted depth value, the binary robot mask, as well as the given camera extrinsic and intrinsic matrices and re-projects points back to 3D; (C) *ShapeNet*, which predicts the robot shape from the point cloud.

*These authors contribute equally.

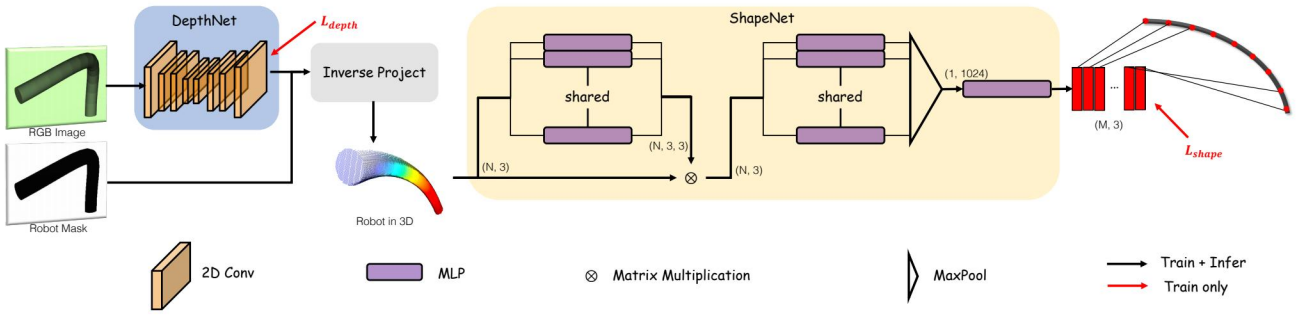


Figure 1. Overview of the Framework

3.1. Problem Formulation

Assume we are given an RGB image of the robot, $I_{RGB} \in \mathbb{R}^{H\times W\times 3}$ taken by a camera with extrinsic parameter $T \in SE(3)$, intrinsic parameter $K \in \mathbb{R}^{3\times 3}$, and resolution H, W. Assume the binary occupancy mask of the robot in the image is known as $O \in \mathbb{B}^{H\times W}$. The goal is to find the position of the robot in 3D, parameterized by the 3D coordinates of M evenly-spaced points on the centerline of the robot, denoted as $P_r \in \mathbb{R}^{M\times 3}$.

3.2. DepthNet for Depth Sensing

We adopt an encoder-decoder style network [1] based on 2D CNNs as DepthNet for predicting the depth at each pixel. Specifically, it takes as input the RGB image of the robot I_{RGB} and outputs the predicted depth map, denoted as $\hat{D} \in \mathbb{R}^{H\times W}$. We supervise the depth learning using L2 regression loss. The per-pixel depth loss is computed as follows, and we take the average of the loss that belongs to the robot (using the given occupancy mask) as the total loss.

$$L_{depth,\ u,v} = \|\hat{d}_{u,v} - d_{u,v}\|_2 \tag{1}$$

where $\hat{d}_{u,v}$ and $d_{u,v}$ are the predicted and ground truth depth value at index (u, v) on the image.

3.3. Inverse Projection

Assume accurate camera parameters are obtained from calibration where $K = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}$ is the intrinsic matrix, and $T \in SE(3)$ is the extrinsic. We transform the points to 3D using the predicted depth map $\hat{d}$ as the following.

$$p_{u,v} = T \begin{bmatrix} x'_{u,v} \\ y'_{u,v} \\ \hat{d}_{u,v} \\ 1 \end{bmatrix} \tag{2}$$

where $x'_{u,v}$ and$y'_{u,v}$ are calculated as follows,

$$x'_{u,v} = \frac{(u_{u,v} - u_0)\cdot \hat{d}_{u,v}}{f_x},\ y'_{u,v} = \frac{(v_{u,v} - v_0)\cdot \hat{d}_{u,v}}{f_y} \tag{3}$$

We then use the provided binary mask of the robot to filter the points, denoted as $P_{in} \in \mathbb{R}^{N\times 3}$ and N is the number of points belong to the robot.

3.4. ShapeNet for Shape Estimation

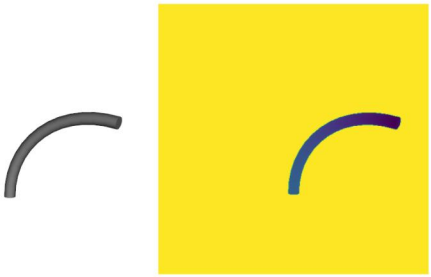
We design a network similar to PointNet [6] for estimating the shape of the robot given the 3D point cloud. It takes as input the 3D coordinates of the observed robot in the form of a point cloud, $P_{in} \in \mathbb{R}^{N\times 3}$. The first part of the network tries to learn a per-point transformation, and the second part is expected to learn point-wise features followed by a max-pooling layer to obtain the global feature of the point cloud. The output of *ShapeNet* is a predicted set of M points on the centerline of the robot that are evenly-spaced, denoted as $P_{out} \in \mathbb{R}^{M\times 3}$. Similar to *DepthNet*, we supervise the shape learning using L2 regression loss.

4. Evaluation

We train and evaluate the proposed model on a custom dataset collected from simulation. Common metrics used in continuum robot research are adopted to measure the accuracy of shape sensing and tip tracking. In the absence of common baseline methods or benchmarks, we plan to compare accuracy with results reported in the literature from different methods.

4.1. Dataset

We collected a custom dataset using an existing simulator. The simulated tendon-driven continuum robot is 200 mm in length and 10 mm in radius with a protective sleeve. Randomly sampled robot configurations are rendered with the Visualization Toolkit (VTK), where we save 512×512 RGB and depth images (Figure 2) along with



(a) RGB image (b) depth image

Figure 2. Sample Image Data from Custom Dataset

camera configuration and ground truth robot shape. Currently, the dataset contains 10000 robot configurations. We plan to make the dataset more realistic by adding texture to the robot and noise, and increase the size of the dataset.

4.2. Metrics

Shape sensing for continuum robots has typically been evaluated in terms of mean error of robot shape (MERS) and mean error of tip tracking (METE) [8]. Although they have been calculated differently across literature, we define MERS to be the average Euclidean distance between the predicted set of evenly-spaced points, $P_{out} \in \mathbb{R}^{M\times 3}$, and corresponding ground truth points across different configurations in the robot's workspace. We also constraint $M \geq 10$ so the points are representative of the robot's overall shape. METE is calculated in the same way but only accounting for the tip position. We will evaluate our method against these two metrics with and without external loading to better reflect application scenarios. The target accuracy is 1 mm for the method to be comparable with existing approaches while only requiring monocular input.

5. Milestones

- Nov 16: Finalize proposal
- Nov 23: Finalize on simulation dataset generation
- Nov 25: Train and Fine-tune Depth Sensing network on the simulated dataset
- Nov 30: Train and Fine-tune Shape Estimation network on the simulated dataset
- Dec 3: Train and Integrate the two-stage network
- Dec 8: Finalize poster and report for presentation

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. 2019. 1, 2

[2] Jessica Burgner, S. Duke Herrell, and Robert J. Webster. Toward fluoroscopic shape reconstruction for control of steerable medical devices. *ASME 2011 Dynamic Systems and Control Conference and Bath/ASME Symposium on Fluid Power and Motion Control, Volume 2*, 2011. 1

[3] Jessica Burgner-Kahrs, D. Caleb Rucker, and Howie Choset. Continuum robots for medical applications: A survey. *IEEE Transactions on Robotics*, 31(6):1261–1280, 2015. 1

[4] Mohsen Moradi Dalvand, Saeid Nahavandi, and Robert D. Howe. High speed vision-based 3d reconstruction of continuum robots. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016. 1

[5] Xin Dong, D. Axinte, David Palmer, S. Cobos, M. Raffles, Amir Rabani, and J. Kell. Development of a slender continuum robotic system for on-wing inspection/repair of gas turbine engines. *Robotics and Computer-Integrated Manufacturing*, 44:218–229, 2017. 1

[6] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[7] Michael Sfakiotakis, Asimina Kazakidi, Nikolaos Pateromichelakis, John A. Ekaterinaris, and Dimitris P. Tsakiris. Robotic underwater propulsion inspired by the octopus multi-arm swimming. In *IEEE International Conference on Robotics and Automation*, pages 3833–3839. IEEE, 2012. 1

[8] Chaoyang Shi, Xiongbiao Luo, Peng Qi, Tianliang Li, Shuang Song, Zoran Najdovski, Toshio Fukuda, and Hongliang Ren. Shape sensing techniques for continuum robots in minimally invasive surgery: A survey. 64(8):1665–1678. Conference Name: IEEE Transactions on Biomedical Engineering. 3

[9] Chaoyang Shi, Xiongbiao Luo, Peng Qi, Tianliang Li, Shuang Song, Zoran Najdovski, Toshio Fukuda, and Hongliang Ren. Shape sensing techniques for continuum robots in minimally invasive surgery: A survey. *IEEE Transactions on Biomedical Engineering*, 64(8):1665–1678, 2017. 1

[10] Ian D Walker. Continuum robot appendages for traversal of uneven terrain in in situ exploration. In *Aerospace Conference*, pages 1–8, 2011. 1

[11] Rui Wang, Stephen M. Pizer, and Jan-Michael Frahm. Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[12] ChaoQiang Zhao, QiYu Sun, ChongZhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, 63(9):1612–1627, 2020. 1