Vision-Based Force Estimation for Minimally Invasive Telesurgery Through Contact Detection and Local Stiffness Models

My H. Le[1,2], Shuyuan Yang[2], Kyle R. Golobish[3], Juan C. Beaver[2], Zonghe Chua[1]
Departments of Electrical Engineering, Computer, and Electrical Engineering[1], Computer and Data Sciences[2], and Mechanical and Aerospace Engineering[3], Case Western Reserve University

Introduction

Force sensing in surgical robots is challenging and often lacks cost-effective solutions, resulting in minimal haptic feedback, making safe tissue manipulation challenging. While some solutions use robot and visual data to estimate force, they have limitations in their adaptability to unseen visual scenarios[a]. Here, we present a versatile approach that combines a learning-based module for tissue contact detection with a local stiffness model for force estimation. This approach offers scalability and adaptability to various surgical scenarios, by requiring minimal fine-tuning using crowd-source human labels instead of sensor measurements and works without the need for access to the robot state and camera parameters.

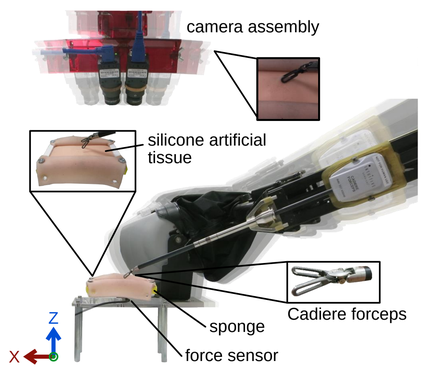

Figure 1. Hardware setup to detect contact and estimate force at the surgical end-effector

Methods

Dataset

- 46 demonstrations of a da Vinci Research Kit (dVRK) patient-side manipulator (PSM) performing various actions on simulated tissue from nine viewpoints and manipulator configurations (see Figure 1).
- PSM joint encoder, joint torques, stereo images, ground truth force data were recorded at 30Hz.
- Camera parameters are unknown.
- Training, validation, and test split sets had 16, 8, and 22 demonstrations, with each demonstration including over 3,000 video frames from the left and right cameras each.
- Visual contact labels were crowdsourced with Amazon Mechanical Turk.
- Ground truth contact labels were generated from force sensor data by classifying forces above a 0.2 N threshold as "in contact".

System Design

Vision-based Contact Detector:

- Classifies tissue contact.
- Uses EfficientNetB3 as a feature encoder to predict contact from images.
- Trained with crowd-sourced contact labels (VisualContact) or ground truth contact (GTContact)
- Image augmentation was performed using randomized image transformations.

Different approaches to force estimation based on availability of robot state information:

- *Contact-conditional Local Force and Stiffness Estimation* (with Robot State Information)
  - Combines vision-based contact signals with robot end-effector force and position measurements to estimate material stiffness.
  - A stiffness $\hat{k}_i$ is estimated from the noisy torque-based end-effector force estimates using linear regression (Stiffness Estimator in Figure 2).
  - Benchmark force estimation models use visual contact signals and $\hat{k}_i$ estimated from ground truth force (VisualContactGTStiff), ground truth force and ground truth stiffness (GTContact) and position difference from the end-effector combined with stiffness constants estimated using F_{PSM} (PosDiff).
- *Contact-conditional Local Force Estimation* (with No Robot State Information)
  - *Vision-based Normalized Position Estimator* (see Figure 2) is a fully connected neural network that estimates 3D end-effector positions from 8 keypoints extracted by DeepLabCut[b] from the video frames.
    - Trained two variants (stereo and monocular image streams).
  - Vision-based normalized positions are used to estimate contact-conditional forces, eliminating the need for robot state data.
  - Uses arbitrary scaling coefficients k_i, which are benchmarked against constants fit using ground truth force measurements.

Training Parameters

- *Vision-based Contact* Detection:
  - EfficientNetB3 was trained for 150 epochs with a batch size of 32, and hyperparameters were determined using a pseudo-randomized grid search.
- *Vision-based Normalized Position* Estimator:
  - 2 neural network variants - a stereo network and a mono network. Using a hyperparameter grid search, learning rate and L2 regularization were chosen to be 0.0001, and training was carried out over 200 epochs using a batch size of 32.

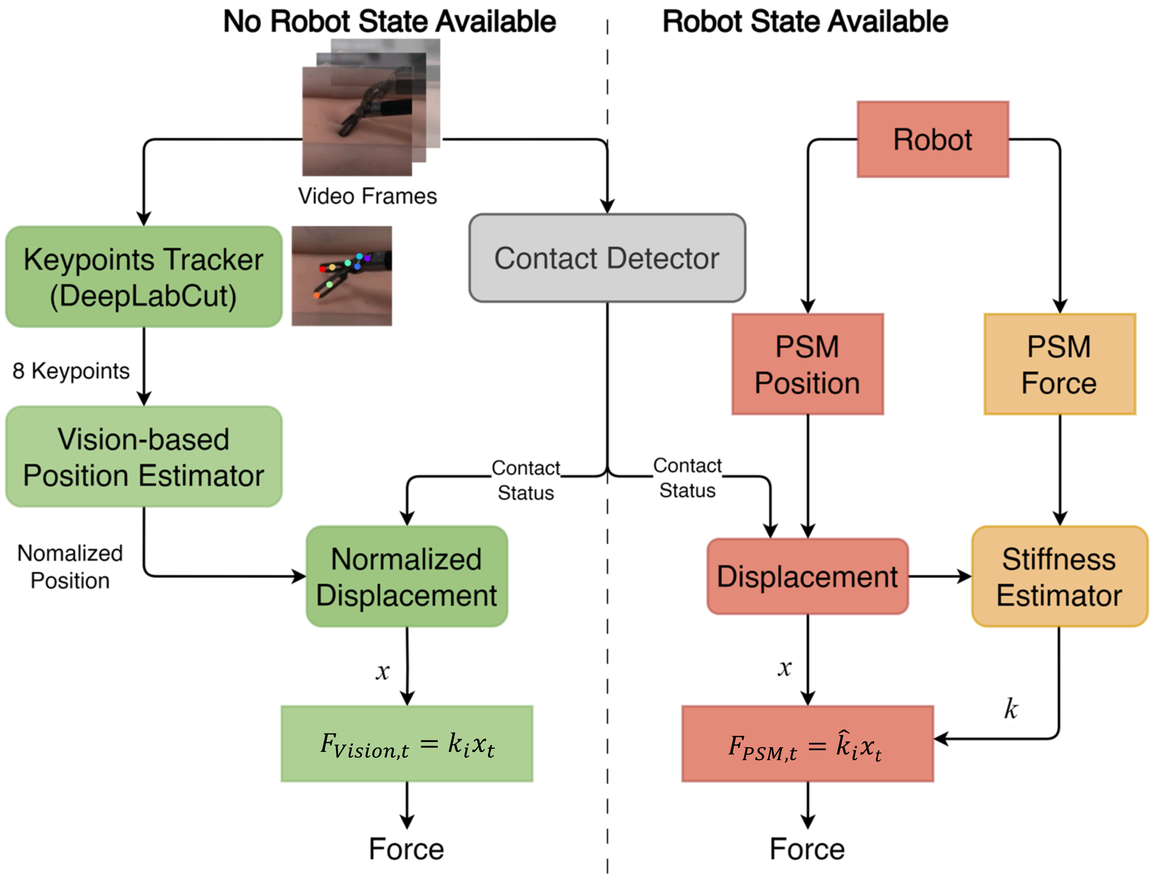

Figure 2. Methods for contact detection and force estimation: (Left) When there is no robot state available for force estimation. (Right) When robot position and joint torque estimates are available.

[a] D.-K. Ko, K.-W. Lee, D. H. Lee, and S.-C. Lim, "Vision-based Interaction Force estimation for robot grip motion without tactile/force sensor," *Expert Systems with Applications*, vol. 211, p. 118441, 2023. doi:10.1016/j.eswa.2022.118441

Results & Discussion

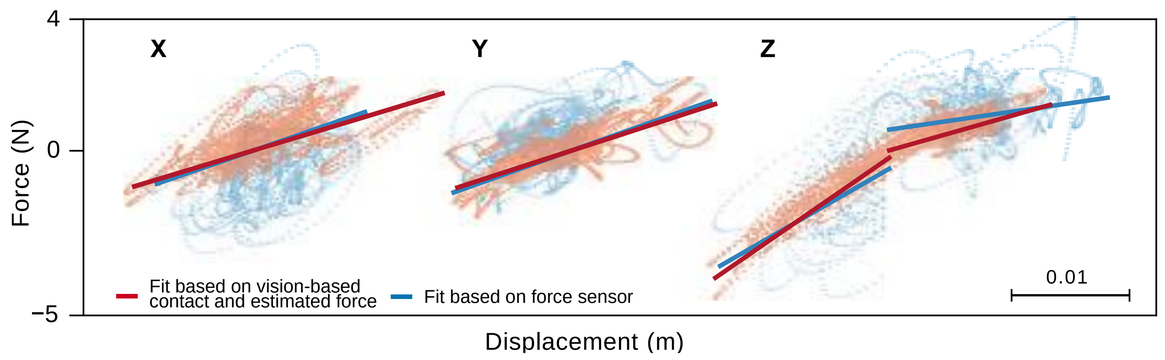

Figure 3. Best fit stiffness models based on either force sensor readings, or the estimated end effector forces using joint torques, with contact conditional displacement reading from joint encoders.

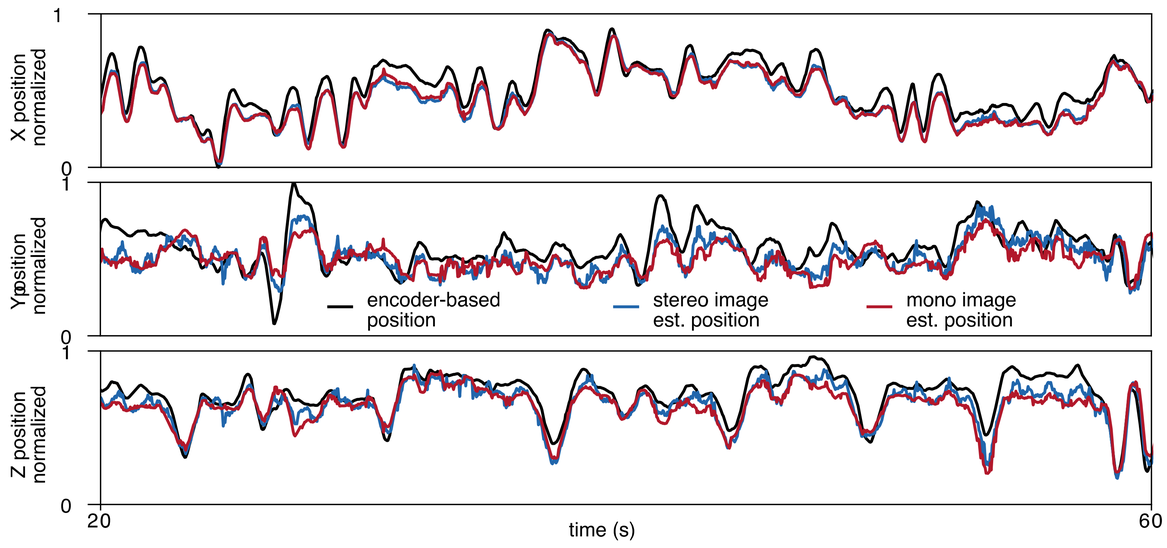

Figure 4. Normalized position estimates from the mono image and stereo image variants of the vision-based position estimator neural networks compared to that of the ground truth from the joint encoders. The position estimator networks take as input the pixel coordinates of eight instrument keypoints identified from Deep Lab Cut.

Model	RMSE			
	Overall	x	y	z
Stereo	0.083 ± 0.023	0.049 ± 0.011	0.116 ± 0.038	0.085 ± 0.020
Mono	0.094 ± 0.016	0.051 ± 0.010	0.136 ± 0.024	0.096 ± 0.014

Table 2. RMSE values of vision-based normalized position estimator

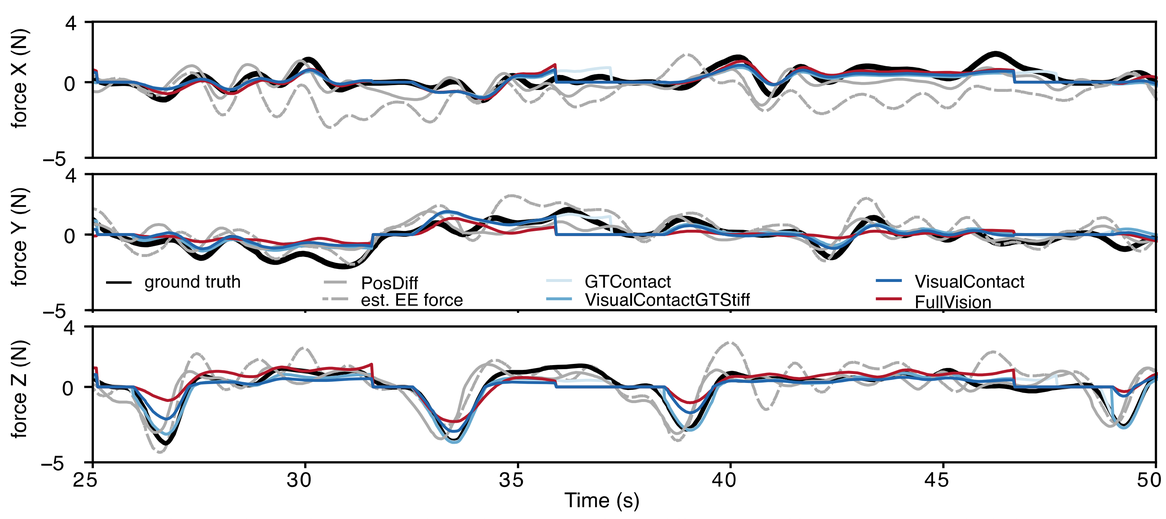

Figure 5. Example force predictions from a manipulation demonstration for the different type of force estimation methods that were tested.

Method	RMSE(Nm^{-1})			
	Norm	x	y	z
Joint torque-based	1.709 ± 0.172	1.096 ± 0.129	0.821 ± 0.092	1.007 ± 0.195
GTContact	0.948 ± 0.161	0.621 ± 0.129	0.563 ± 0.13	0.424 ± 0.101
VisualContactGTStiff	0.955 ± 0.156	0.619 ± 0.128	0.560 ± 0.126	0.445 ± 0.106
VisualContact (our approach)	1.114 ± 0.175	0.611 ± 0.125	0.542 ± 0.113	0.721 ± 0.24
PosDiff	1.477 ± 0.191	0.736 ± 0.107	0.716 ± 0.15	1.042 ± 0.213
FullVision	1.758 ± 0.316	0.834 ± 0.17	0.847 ± 0.232	1.264 ± 0.319

Table 1. RMSE values of force estimation methods

- F1 score of crowd-sourced labels against ground truth: **0.968**
- F1 score of vision-based contact detection method trained on crowd-sourced labels against ground truth: **0.974**
- The root mean squared error (RMSE) for stiffness estimates between GTContact and VisualContact were between 41 to 63 Nm^{-1} in each of the Cartesian force directions.
- VisualContact shows better performance than the position difference method PosDiff and the naïve joint-torque based approach.
- There was a smaller error increase from GTContact to VisualContactGTStiff than VisualContactGTStiff to VisualContact suggesting that error was mainly from stiffness estimation using F_{PSM}, especially in the Z direction.
- Full vision performs similarly to the naïve Joint torque-based method.

Acknowledgements

The authors would like to thank the Case Western High Performance Computing Cluster for providing the compute resources for training our neural networks.

[b] T. Nath, A. Mathis, A. C. Chen, A. Patel, M. Bethgeand M. W. Mathis, "Using DeepLabCut for 3D markerless pose estimation across species and behaviors", *Nature Protocols*, vol. 14, no. 7, pp. 2152–2176, 2019, doi: 10.1038/s41596-019-0176-0.