Force/Torque Sensing for Soft Grippers using an External Camera

Jeremy A. Collins¹, Patrick Grady¹, Charles C. Kemp¹

Abstract—Robotic manipulation can benefit from wrist-mounted force/torque (F/T) sensors, but conventional F/T sensors can be expensive, difficult to install, and damaged by high loads. We present Visual Force/Torque Sensing (VFTS), a method that visually estimates the 6-axis F/T measurement that would be reported by a conventional F/T sensor. In contrast to approaches that sense loads using internal cameras placed behind soft exterior surfaces, our approach uses an external camera with a fisheye lens that observes a soft gripper. VFTS includes a deep learning model that takes a single RGB image as input and outputs a 6-axis F/T estimate. We trained the model with sensor data collected while teleoperating a robot (Stretch RE1 from Hello Robot Inc.) to perform manipulation tasks. VFTS outperformed F/T estimates based on motor currents, generalized to a novel home environment, and supported three autonomous tasks relevant to healthcare: grasping a blanket, pulling a blanket over a manikin, and cleaning a manikin’s limbs. VFTS also performed well with a manually operated pneumatic gripper. Overall, our results suggest that an external camera observing a soft gripper can perform useful visual force/torque sensing for a variety of manipulation tasks.

I. INTRODUCTION

During robotic manipulation, grippers often apply forces and torques to the environment. Sensing the force and torque applied by the gripper has been useful for autonomous manipulation, but sensors that provide this information have limitations. Notably, conventional F/T sensors can be expensive, difficult to mount, and damaged by high loads.

For example, a common approach to measuring the load applied to a gripper is to mount an F/T sensor between the gripper and the robot’s wrist. F/T sensors often use strain gauges to sense tiny deformations in an elastic element of the sensor. This approach requires that the strain gauges be resilient to the external load applied to the gripper as well as gravitational and inertial forces from the gripper itself. For many applications, the strain gauges need to be both stiff and sensitive, and protective coverings could reduce performance by interfering with the load on the strain gauges. Together, these design objectives are difficult to achieve.

We present an alternative to conventional F/T sensors. Instead of relying on the deformation of internal components, VFTS directly observes the deformation of a soft gripper using an external camera. The high compliance of soft grippers results in deformations that can be visually observed using a commodity camera. By observing this load-dependent phenomenon, the causative forces and torques can be estimated. We rigidly mount a camera with a fisheye lens to the gripper (i.e., an eye-in-hand camera). We then train a convolutional neural network, VFTS-Net, to estimate the applied force and torque based on a single RGB image from this camera (Figure 1).

In contrast with conventional F/T sensors, our approach relies on a low-cost USB camera (~$60). Our method eases installation by allowing the camera to be mounted to the exterior of the gripper rather than between the gripper and the wrist. Since the camera visually senses the loads from a distance, it is also less likely to be damaged by high loads.

Researchers have investigated related methods that involve placing a camera inside a gripper behind a compliant surface. Loads applied can be estimated by observing deformation in the surface. In contrast, our approach uses an external camera to observe a soft gripper. Our approach does not require modification of the gripper’s contact surfaces or interior. The global view of the external camera facilitates estimation of

---

¹Jeremy A. Collins, Patrick Grady, and Charles C. Kemp are with the Institute for Robotics and Intelligent Machines at the Georgia Institute of Technology (GT). This work was supported by NSF Award #2024444. Code, data, and models are available at https://github.com/Healthcare-Robotics/visual-force-torque. Charles C. Kemp is an associate professor at GT. He also owns equity in and works part-time for Hello Robot Inc., which sells the Stretch RE1. He receives royalties from GT for sales of the Stretch RE1.
We provide evidence for the feasibility of our approach by collecting a dataset of in-the-wild robotic manipulation in multiple environments and testing on held-out data from novel environments with manipulation of unseen objects. We also provide an analysis indicating that lower performance corresponds with types of gripper deformation that are more difficult to visually observe.

VFTS outperformed a baseline method that uses motor currents to estimate 6-axis F/T measurements. We also provide evidence that the estimates from VFTS can support autonomous manipulation by enabling a mobile manipulator to perform three autonomous tasks. We additionally show that VFTS-Net can be trained to work with two distinct soft grippers: a tendon-actuated gripper and a pneumatic gripper.

In summary, our paper includes the following contributions:

- We present Visual Force/Torque Sensing (VFTS), a method that uses a convolutional neural network to estimate the forces and torques exerted on a soft gripper given an RGB image from an eye-in-hand camera.
- We demonstrate the utility of VFTS for real-world applications with a series of robotic tasks conducted with a mobile manipulator.
- We will release our code, data, and trained models.

II. RELATED WORKS

A. Tactile Sensing using Physical Sensors

A wide range of methods have been proposed for robotic tactile sensing. Sensors have been developed to measure vibration [1], temperature [2], or pressure inside a fluid-filled cavity [3], [4]. To enable safe operation around humans, some collaborative robots (cobots) have been constructed with actuator torque sensors [5], [6]. Other research has used the compliant joints of a robot for force estimation [7].

Six-axis force/torque sensors are ubiquitous in both research and industrial applications. These sensors function by measuring elastic deformation in the structure of the sensor [8]. Force/torque sensors are often placed between the robot arm and end-effector, and they have been used for a wide array of applications such as robotic surgery [9], [10], humanoid robot locomotion [11], cloth manipulation [12], and robot-assisted dressing and feeding [13], [14]. Force/torque sensors are also widely used in industrial applications, including sanding [15], deburring [16], part alignment [17], and robot teaching [18].

B. Tactile Sensing using Vision

We make the distinction between two types of vision-based approaches to understand tactile signals, those using internal versus external sensors.

Researchers have developed tactile sensors that use a camera mounted inside a robot to perceive a soft exterior. Prior techniques have used dot patterns [19], [20], [21], photometric stereo [22], or fiducial markers [23] to track the motion of the exterior and infer deflection at a high resolution. A variety of other optical techniques have been proposed [24] to sense deformation from internal sensors.

Other work uses cameras mounted external to the robot to estimate forces. A common technique is to leverage the deformation caused by a rigid gripper making contact with soft objects. This approach has seen success for estimating forces on soft tissue during surgery [25], [26] and manipulation of soft household objects [27]. Other work has used the trajectory of grasped objects to infer the net forces and moments necessary to cause this motion [28], [29].

The deformation caused by a soft body interacting with a rigid environment can also be used to perceive force for human hands and bodies [30], [31]. Urban et al. [32] use a camera mounted to a human fingernail to observe changes in appearance to estimate force and torque.

This paper builds upon work for estimating the contact pressure applied by a soft gripper to a planar surface [33]. They use external static cameras in a controlled environment to observe deformations in the part of the gripper contacting the surface. Our work estimates forces and torques measured at the base of the gripper in uncontrolled environments and only uses sensors mounted to the robot. This allows
collection of a more diverse training dataset and contributes to VFTS-Net’s ability to generalize to novel environments.

III. METHODS

In this section, we describe the grippers used in this study, the data collection hardware and protocol, and the network architecture and training process for VFTS-Net.

A. Selected Grippers

In order for the forces and torques exerted by a gripper to be visually detected, the gripper must visibly change under the application of force. As a result, we focus on soft grippers. These grippers are compliant by design, and this compliance enables the visual estimation of forces. We demonstrate our method with two models of gripper:

1) Tendon-Actuated Gripper: The first gripper used is a tendon-actuated gripper included with the Stretch RE1 robot (Figure 1). The gripper features rubber suction-cup-like fingertips which are supported by spring steel flexures. The robot closes the gripper by retracting the inner flexures, causing the fingertips to be pulled together.

During the application of forces and torques, the flexures deform slightly, causing a noticeable displacement in the fingertips of the gripper. Figure 2b shows the response of the gripper to a load on one of the fingertips. The rubber fingertips are soft, and parts of the gripper may conform to the object being manipulated, especially under high force.

2) Pneumatic Gripper: The second gripper used is a commercially-available pneumatic gripper manufactured by SoftGripping [34]. This gripper is made of a flexible silicone rubber, and is constructed with a hollow cavity in each finger. The finger can be filled with compressed air, causing an asymmetric bending and closure of the gripper.

During the application of forces and torques, the entire silicone finger deforms relatively uniformly (Figure 2c). Generally, the gripper is less stiff than the tendon-actuated gripper and large deflections of the fingertips are possible under high forces.

B. Data Capture Hardware

During collection of the training and test datasets, we mount a 6-DoF ATI Mini45 force/torque sensor [35] between the gripper and the wrist of the robot (Figure 1). This allows the accurate sensing of forces and torques applied to the gripper. We do not perform any transforms on the force/torque data from the sensor. Since the sensor drifts over the course of tens of minutes, before each recording, we set the gripper to a level pose and zero the force/torque sensor readings. As a consequence, VFTS-Net estimates the forces and torques applied to the wrist of the gripper, minus the gripper’s weight when it is level.

To capture images of the gripper, we mount a Spinel UC20MPG-F185 fisheye camera above the gripper (Figure 2) using 3D-printed hardware. Both the camera and the mount are a part of the Stretch Teleop Kit from Hello Robot Inc. The camera captured images at 30 FPS and 640x480 resolution. Due to the wide-angle lens, we crop the images to focus on the area around the gripper (Figure 3). We also synchronize the camera and force/torque sensor readings in time.

During data collection with the tendon-actuated gripper, we mount the gripper to a Stretch RE1 mobile manipulator robot by Hello Robot Inc. [36]. The robot was equipped with the Dexterous Wrist add-on, allowing full 6-DoF control of the end-effector. We also record the robot state at each time step, allowing the creation of a baseline based on motor currents. During data collection with the pneumatic gripper, we mount the gripper to a long handle to enable manual operation.

C. Data Capture Protocol

We collect approximately 2 hours of data for each gripper. Our approach does not require external sensors, such as cameras mounted to the environment, which simplifies collecting data in a variety of settings. We collect data with each gripper in a lab setting and two distinct office settings. For the tendon-actuated gripper, we also collect data in a home environment. For each gripper, we hold out one complete environment for testing. For the tendon-actuated gripper we hold out the home environment, and for the pneumatic gripper we hold out an office environment.

To capture varied gripper actions, we collect three primitive interaction types: push, in which one or both fingertips of the gripper are pushed into a surface; slide, in which one or both fingertips of the gripper are dragged along a surface; and grasp, in which objects of varying size and mass were grasped at different orientations. In addition, we collect two high-level interaction types: manipulate object, which includes tool use, multi-object interactions, and using objects for their intended functions, and manipulate scene, which includes interactions with in-context items such as drawers, light switches, and faucets. We vary the wrist roll, wrist pitch, wrist yaw, and grip force throughout the recording process. We record at least 150 seconds of data for each combination of action and environment.

D. Network Architecture

We designed VFTS-Net to estimate forces and torques given a single RGB image of the gripper. We train one network for each gripper. The network has a ResNet-18 backbone initialized with weights from pretraining on ImageNet [37], and we modify the last linear layer of the network to output 6 scalar values (Figure 1) to directly represent forces and torques.

We train the network with a mean-squared error loss to minimize the error between ground truth force and torque vectors $F, T$ and estimated force and torque vectors $\hat{F}, \hat{T}$.

$$L = ||F - \hat{F}||^2 + c||T - \hat{T}||^2$$  \hspace{1cm} (1)

To compensate for the intrinsic scale difference between force (Newtons) and torque (Newton-meters), we weight the loss for torque with a constant $c$. We set $c$ equal to the ratio of the force and torque standard deviations in the training data for each gripper.
Frames from the test set for both the tendon-actuated gripper and the pneumatic gripper are shown. Estimates from VFTS-Net are visualized as lightly shaded arrows, and ground truth measurements from the force/torque sensor are darkly shaded arrows. We collected data while the gripper was interacting with objects in a variety of environments, including a bathroom, kitchen, dining room, and office space. The bottom right figure for both gripper types shows a failure case. Both grippers estimate force in the Z-axis less accurately.

| Method        | Gripper   | RMSE$_F$ (N)  | RMSE$_T$ (Nm) |
|---------------|-----------|---------------|---------------|
| Mean-Guesser  | Tendon-Act.| 2.405         | 0.389         |
| Effort Baseline| Tendon-Act.| 1.936         | 0.246         |
| VFTS-Net      | Tendon-Act.| **1.688**     | **0.185**     |
| Mean-Guesser  | Pneumatic | 1.388         | 0.144         |
| VFTS-Net      | Pneumatic | **0.808**     | **0.051**     |

**TABLE I**

FORCE AND TORQUE ERROR ON TEST SET

During training, we horizontally flip images and the corresponding ground truth F/T sensor readings at random. We also augment the data by varying the brightness, contrast, saturation, and hue of the training images.

We train VFTS-Net for 300k iterations with a learning rate of $1e^{-4}$ and a batch size of 4. We use the PyTorch framework [38] and the Adam optimizer [39] to train the network. During evaluation, the network runs at 120 FPS on a desktop PC with an NVIDIA GeForce RTX 3090 GPU, or 25 FPS on the Stretch RE1’s Intel i5-8259U CPU.

IV. EVALUATION

We conducted a variety of experiments to assess the performance of force and torque estimation. We report results with the held-out test data. In addition, we present three evaluations with a mobile manipulator to demonstrate the relevance of VFTS to real-world tasks.

A. Evaluation on Held-Out Test Set

We evaluated the performance of VFTS-Net on the test set to assess its ability to infer the forces and torques reported by a conventional wrist-mounted force/torque sensor. In Table I we report the root-mean-squared error for estimated force and torque compared to the ground truth force and torque with force RMSE calculated as:

$$RMSE_F = \sqrt{\frac{1}{N} \sum ||\hat{F} - F||^2} \quad (2)$$

We also consider two baselines for comparison against VFTS-Net:

- **Mean-Guesser**: We report the error when always guessing the mean forces and torques from the training set.
- **Effort Baseline**: Since motor current relates to output torque, motor currents can be used to approximate the load on the gripper. The Stretch RE1 does not directly report motor current, but provides effort, a unitless quantity proportional to current. We train a multilayer perceptron (MLP) with 2 hidden layers to estimate force and torque using the same loss function, ground truth data, and testing set as VFTS-Net. The input to the
network is a vector of six efforts from the following actuators: wrist roll, wrist pitch, wrist yaw, gripper, telescoping arm, and arm lift. Since we do not use the robot to collect data with the pneumatic gripper, we only report results on the effort baseline for the tendon-actuated gripper.

As shown in Table I, VFTS-Net outperformed the Mean-Guesser and the Effort Baseline. Several frames from the test set are shown in Figure 3. Figure 5 illustrates that the estimated forces and torques have temporal alignment with ground truth measurements.

B. Grasping a Blanket

To evaluate the ability of VFTS to support autonomous manipulation, we implemented and tested three primitive fabric manipulation behaviors.

The first task evaluates the robot’s ability to stop when a force threshold has been exceeded when picking up a blanket off the surface of a bed. The robot starts with the gripper above the edge of a blanket and then moves downward until the vertical force estimated by VFTS-Net exceeds a 3N threshold. The robot then closes its gripper and picks the blanket up. If the robot successfully grasps and lifts the blanket, we consider the trial to be successful. Table II shows the results.

C. Pulling a Blanket Over a Manikin

In the second task, the robot pulls a blanket over a manikin. This task evaluates the robot’s ability to pull a blanket and stop when a horizontal force threshold is exceeded. The trial begins with the blanket in the gripper, then the mobile base is commanded to travel toward the head of the bed. When the horizontal force estimated by VFTS-Net exceeds a threshold of 2.5N, the robot lowers and releases the blanket. If the blanket was not unintentionally released from the gripper and the robot lowers the blanket onto the manikin, we consider the trial to be successful. Table II shows the results.

D. Cleaning a Manikin’s Forearm and Leg

Tasks such as cleaning, polishing, and sanding often use force/torque sensing to apply pressure to the surface. We evaluate VFTS’s ability in a similar task of cleaning a medical manikin’s limbs. This is related to bathing, which is an important activity of daily living (ADL) with which many people require assistance [40].

We design a simple controller to clean the limbs of a manikin using force/torque sensing performed by VFTS. The robot starts by holding a cloth above the manikin’s limb. It then lowers the gripper until the force magnitude exceeds a 5N threshold, at which point it stops lowering its gripper. The estimated force vector \( \hat{F} \) is assumed to be approximately normal to the surface.
The robot’s pose at time $t$ is denoted as $X_t$. The robot moves its gripper tangent to the estimated surface normal toward a desired wiping direction $\vec{d}$. Simultaneously, the robot moves along the estimated force vector $\hat{F}$ to maintain a desired contact force magnitude, $k_f$. Equation 3 represents the control algorithm used by the robot. A new force estimate from VFTS is used at each time step. The robot moves along the limb by 2 cm at each time step, and reverses the desired direction of motion $\vec{d}$ once the gripper has reached a minimum vertical position or has not detected contact with the surface for more than 3 seconds. This results in a back-and-forth motion. The robot does not have any prior knowledge of the cleaning surface geometry.

$$X_{t+1} = X_t + (||\hat{F}|| - k_f)\frac{\hat{F}}{||F||} + \vec{d} - \text{proj}_{\vec{d}}(\vec{d})$$  \hspace{1cm} (3)

Using this control algorithm, the robot cleans off dry-erase marker from the limb of a manikin in bed. We conduct 10 trials: 5 trials on the manikin’s left forearm and 5 trials on the manikin’s left thigh. No parameters were varied or tuned during the trials aside from the starting position. We define success as the percentage of the marker area erased as measured using a simple color filter applied to images taken by an overhead camera. In our 10 trials, the robot cleaned an average of 87% of the marker off of the manikin’s limbs (see Table III). The algorithm typically succeeded near the center of the arm and leg, but failed on the edges of the leg due to the high curvature and inaccuracies in the $Z$-axis force estimation.

**E. Discussion of Performance**

Figure 6 shows 2D histograms calculated by comparing the estimated and ground truth force and torque for each of the six axes. The figure shows the ranges of forces and torques in the testing dataset and visualizes per-axis performance.

We found that for both grippers, force estimation in the $Z$-axis performs more poorly than the X and Y axes. We suspect that this is because deformations in depth are less observable with a single RGB image, and more likely to be obscured by held objects and gripper forces. In contrast, forces in the X and Y axes are aligned with the image plane.

The pneumatic gripper estimates have lower error than the tendon-actuated gripper for all axes, which may be due to larger deformations of the pneumatic fingers under load.

**V. CONCLUSION**

We presented Visual Force/Torque Sensing (VFTS), a method that visually estimates the 6-axis force/torque measurement that would be reported by a conventional wrist-mounted force/torque sensor. Our method uses an external camera with a fisheye lens to observe a soft gripper as it deforms due to loads. VFTS uses a deep learning model that outputs a 6-axis F/T estimate when given a single RGB image as input. Our method outperforms motor-current-based F/T estimation when tested with data from teleoperated manipulation of novel objects in a novel environment. Our method also enabled successful autonomous mobile manipulation of fabrics in three tasks. Overall, our results suggest that an external camera observing a soft gripper can perform useful visual force/torque sensing for a variety of manipulation tasks.
