Setting Realistic Expectations for Data-Driven Grasp Stability Prediction

Qian Wan and Robert D. Howe

***Abstract*—In data-driven, tactile-based grasp stability prediction algorithms, the quality of the tactile sensors has a direct impact on the performance of the algorithm. We present a modeling process that suggests there exists a performance ceiling for machine-learned prediction algorithms that are trained using sensors with limited spatial resolution. We also build a framework that allows for systematic examination of the accuracy and complexity of stability prediction neural networks as a function of the amount of noise and gap in the sensor signals.**

Highly reliable grasping is essential in many real-world robotics applications. A household robot that achieved even 99.9% success in grasping would still drop many objects each week, making it unacceptable for most potential users. A crucial method for attaining reliable grasping is stability prediction: assessing whether the grasp is stable after the fingers have made contact but *before* the object is lifted.

In theory, predicting grasp stability can make use of well-developed physics-based grasp analysis [1]. This involves determining whether the forces exerted on the object by the fingers and the environment (e.g., gravity) are in equilibrium. However, the sensors available today are incapable of delivering the necessary parameters accurately and quickly, and the assumptions behind the interaction models, such as point-with-friction contacts or uniformly deformed soft contacts, are idealistic and do not reflect real-world scenarios. As a result, analytical methods are rarely used in real-time and real life.

Many researchers have turned to empirical data-driven machine learning methods. However, despite a decade worth of effort, results are generally limited to 80% accuracy in novel objects (e.g. [2],[3]).

The observed limitations in grasp stability prediction using machine learning may be due to two issues: (1) the dimensionality in grasping is too high for the number of training samples we can collect, and (2) there is not enough information embedded in the input sensory data. The solution to the former issue is collecting a greater volume and variety of training samples, or building algorithms that are tolerant to high-dimension and low-quantity datasets; the solution to the latter issue is improving sensors. The most sophisticated solutions for the former issue will not resolve problems that originated from the latter, and limitations inherent to the dataset can set a ceiling in the performance of any empirical data driven algorithms. Many current approaches use the sensory signals with a blackbox learning algorithm and do not consider limitation within the signals themselves. As a result, when an algorithm performs less than perfect, it is unclear what is the most efficient way to improve performance.

Both authors are with Harvard Paulson School of Engineering and Applied Sciences, Cambridge, MA 02138, USA. Email:qwan,howe@seas.harvard.edu

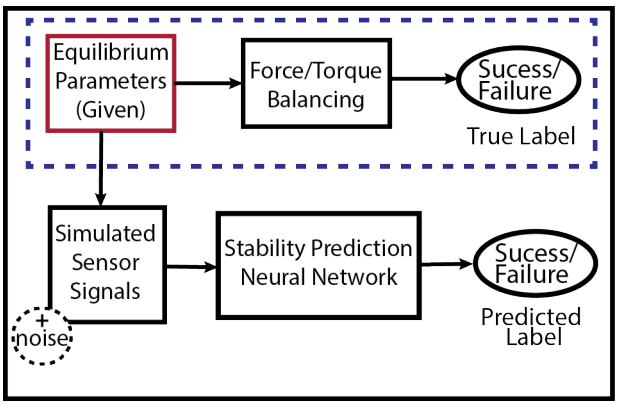

Fig. 1: Simulation framework for evaluating the complexity and accuracy of a grasp-stability-predicting neural network, as a function of noise and uncertainty of sensors. The reference outcomes are generated by evaluating force and moment equilibrium on the object (blue dash box).

Therefore, it is important to understand the relationship between sensors' information quality and the reliablity of grasp stability predictions, so that we can set reasonable expectations for an algorithm's performance given the available sensors, as well as develop quantitative ways of measuring the trade-offs when choosing between different sensors and algorithms.

In the work presented here, we first give an example of how sensors with limited spatial resolution may set limits for machine-learned stability prediction algorithms. We then build a simulation framework to quantitatively measure how the deterioration of input signals affect the performance of data-driven learning algorithms in a general sense.

To show how sensor limitations can pose a ceiling in data-driven algorithm performance, we analyze the effect of tactile sensors' limited spatial resolution on the accuracy of stability prediction algorithms. We show that when tactile sensors have limited resolution — a statement that is applicable to almost all of the existing integrated sensor suites, there will be regions in the task space where successful grasps and failed grasps evoke the same sensory response (Fig. 2). In other words, the same sensory signals in these regions will have seemingly random grasp outcome labels, and algorithm that is trained using these randomly labeled data would make equally random predictions for grasps inside the same region. Therefore the accuracy of the prediction algorithm will be limited by the size of this ambiguous region, which is inversely correlated with the tactile sensors' resolution (Fig. 3).

To conduct a more systematic examination of the effect of sensor quality on data-driven algorithm performances, we

build a neural network to predict grasp outcomes using simulated sensory data. In this framework, the true stability of a simulated grasp is calculated by balancing forces on the object, and the sensory data input is carefully manipulated in terms of the amount of useful information it carries and its complexity (Fig. 1). We start with assuming a perfect sensing scenario, in which all the necessary force equilibrium parameters are accurately measured. For simple lifting tasks using precision grasps, these parameters are contact locations, contact forces, surface normals, coefficients of friction, masses, and objects' centers of mass. The outcome of the grasp when sensing is perfect can be calculated by balancing forces and torques on the object. Sensor quality is then manipulated by mapping the key parameters onto sensors that resembles the ones commonly used in research, such as tactile sensor arrays and joint-angle and joint-force sensors. The result is sensory signals that are compressed to a similar degree as the real world setups. Additional noise may also be included to simulate noise inherent to the sensor devices.

The simulated sensor signals and their corresponding true labels are used to train a neural network. When using the ideal sensors as input, the neural network simply has to learn how to add up the correct nodes. As more noise and signal compression are imposed onto the sensors, the network becomes more complex, and the size of the training data and the training time evolve accordingly. The utility of this network is evaluated by the accuracy of its predictions on novel grasps, the complexity of the network, and the amount of training data and time needed to achieve a desired accuracy.

By systematically increasing the complexity and noise in input sensor signals, we can observe the change in the neural networks in terms of their best-case accuracy, the amount and variety of training data needed, the length of training time, and their tolerance to specific uncertainties. The results here can help to build intuition for the expected performance of empirical data-driven learning algorithms when sensors' quality is known. The tunable-noise feature of the framework allows researcher to understand the system's sensitivity to different kinds of uncertainties, so that we can quantitatively evaluate the trade-offs between algorithmic and hardware improvements.

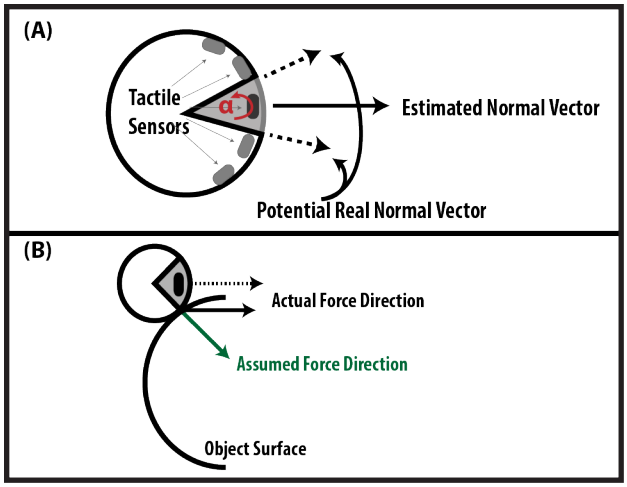

Fig. 2: An example of how limited spatial resolution in tactile sensors confounds stability prediction. A stable grasp is defined in classical grasp analysis as achieving force and moment equilibrium on the object without slipping at any of the contacts. The corresponding calculations require the knowledge of surface-normal vectors at contact, contact locations in relation to the object's center of mass, and the direction of forces at contact. When tactile sensors have limited spatial resolution, the actual contact surface normal can differ from the estimated normal vector by as much as $\alpha/2$ (A), as well as assumed and actual force directions (B).

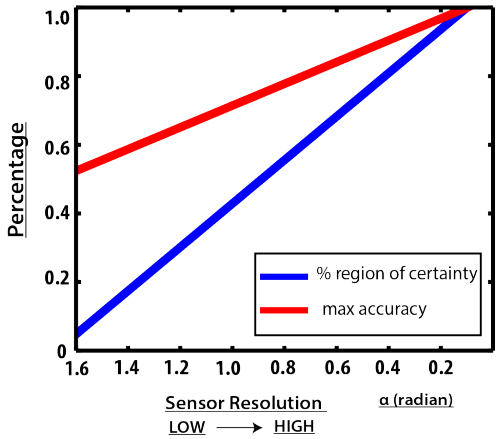

Fig. 3: Tactile spatial resolution limit sets a ceiling for machine-learned prediction performance. Spatial resolution limits lead to a percentage of the training data to having ambiguous labels (blue line), which sets the ceiling for maximum accuracy if signals from those ambiguous regions are used in a grasp prediction algorithm. Example here is for spherical objects.

References

[1] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *Robotics and Automation (ICRA), 2000. IEEE International Conference on*, volume 1, pages 348–353. IEEE, 2000.

[2] Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *arXiv preprint arXiv:1603.02199*, 2016.

[3] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 3406–3413. IEEE, 2016.