Blind Robotic Grasp Stability Estimation Based on Tactile Measurements and Natural Language Prompts

Jan-Malte Giannikos
Faculty of Technology
University of Bielefeld
Bielefeld, NRW 33615, Germany
jan-malte.giannikos@uni-bielefeld.de

David Leins
Faculty of Technology
University of Bielefeld
Bielefeld, NRW 33615, Germany
dleins@techfak.de

Alexandra Moringen
University of Greifswald
Greifswald, MV 17489, Germany
alexandra.moringen@uni-greifswald.de

Oliver Kroemer
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
okroemer@andrew.cmu.edu

Abstract

We design and train a composition of neural network modules that predicts robotic grasp success based on tactile sensor measurements and natural language prompts identifying the object. We use a Franka Emika Panda robot arm equipped with two DIGIT sensors for grasping and language descriptions generated by chatGPT. Our short-term goal is to utilize this approach to improve the accuracy of a grasp stability estimator. The longer-term goal of this work is to enhance haptically driven robot control with language-based context, i.e. task-relevant information which might not be robustly inferred from vision.

1 Introduction

Multiple researchers have proposed successful approaches to solve the problem of robotic grasping [13]. Previous work has been dedicated mainly to vision [6], haptics [12] or a combinations of both modalities [3]. However, in interactive scenarios, relevant information is often encoded in a third modality, language. Since natural language commands often implicitly contain context cues, they can be a valuable source of knowledge, especially in situations where other sensors may be unreliable or unavailable. Thanks to large language models, the use of natural language in robotic manipulation and navigation has become more accessible. Various projects [16, 1, 17] have focused on processing and implementing high-level natural language commands [11, 19]. In contrast to projects like [1] or [17], we do not aim to structure our task through natural language commands but instead seek to utilize context that such language commands might contain. Other projects mainly rely on vision for the robot's perception. In situations where factors like poor lighting, occlusion, and self-occlusion hinder vision, these models still struggle. Humans can rely on their sense of touch to navigate such situations, an approach several papers have applied successfully [9, 12, 18, 5, 7, 8]. Building upon previous work, we explore the combination of natural language as a source of context information, provided by a user, and haptics as a robust perceptual modality controlled by the robot. We evaluate our modeling approach on the exemplary task of grasp stability estimation [2, 5].

Our preliminary experiments have been conducted on a small set of six household objects. We train a neural network model to predict the stability of a grasp based on tactile sensor measurements and a basic natural language prompt. To this end, we first generate a dataset in simulation, which contains tactile sensor readings from roughly 53,000 grasp attempts on the six household objects. We then

NeurIPS 2023 Workshop on Touch Processing: a new Sensing Modality for AI .

used chatGPT to generate a total of 420 short object descriptions, i.e. multiple descriptions for each one of the six objects. During training and testing, for each tactile sensor measurement, we randomly select an appropriate description from the object-specific set.

Our evaluations have shown that the use of randomly selected natural language context yields similar performance as a baseline model utilizing only tactile measurements. This result implies that not all language descriptions are useful and that the tactile data already contains significant object information. We plan on extending our dataset to contain more objects that vary in weight, surface friction, and rigidity, allowing us to capture task-relevant features that have so far been ignored. Our long-term goal is to achieve more efficient and flexible robot control with both tactile and language information.

2 Dataset

Our dataset consists of two components generated for six household objects: 1) A set of robot grasps D_g and 2) A set of corresponding language descriptions D_l of each object generated with chatGPT. We used the PyBullet [4] simulation environment to generate a dataset of grasps for six different household objects. To record our tactile sensor readings we chose the Tacto simulation interface [20] to simulate two DIGIT [10] sensors. The sensors are mounted on the fingers of a Franka robot model taken from [14].

The robot grasp dataset $D_g = \{D_{g_1}, D_{g_2}, D_{g_3}, D_{g_4}, D_{g_5}, D_{g_6}\}$ contains six data subsets $D_{g_i} : i \in \{1, \ldots, 6\}$ each containing interactions with a single object. Each data subset is a collection of data points $D_{g_i} = \{(x_{g_{i1}}, y_{g_{i1}}), \ldots, (x_{g_{in}}, y_{g_{in}})\} : i \in \{1, \ldots, 6\}$ with $n \approx 8900$. A single datapoint contains two tactile images $x_{g_{ij}} : i \in \{1, \ldots, 6\}, j \in \{1, \ldots, n\}$ and a corresponding binary classification label $y_{g_{ij}} : i \in \{1, \ldots, 6\}, j \in \{1, \ldots n\}$ describing if the grasp was successful $y_{g_{ij}} = 1$ or unsuccessful $y_{g_{ij}} = 0$. We do not include any object-relative hand pose information as the object pose and overall geometry is assumed to be unknown.

We use Pybullet [4] as our simulation environment together with the simulation interface Tacto [20], which Meta provides to simulate its DIGIT [10] sensors. We use the Franka robot model from [14] with a custom finger design to record grasp interactions with a Bottle, Rubber Duck, Hair Dryer, Soda Can, Microphone, and Digital Camera, generating the six data subsets $\{D_{g_1}, D_{g_2}, D_{g_3}, D_{g_4}, D_{g_5}, D_{g_6}\}$. Example images of the data collection setup and some of the tactile sensor readings can be found here: `https://anonymous.4open.science/r/Additional-Resources-B8ED/README.md`

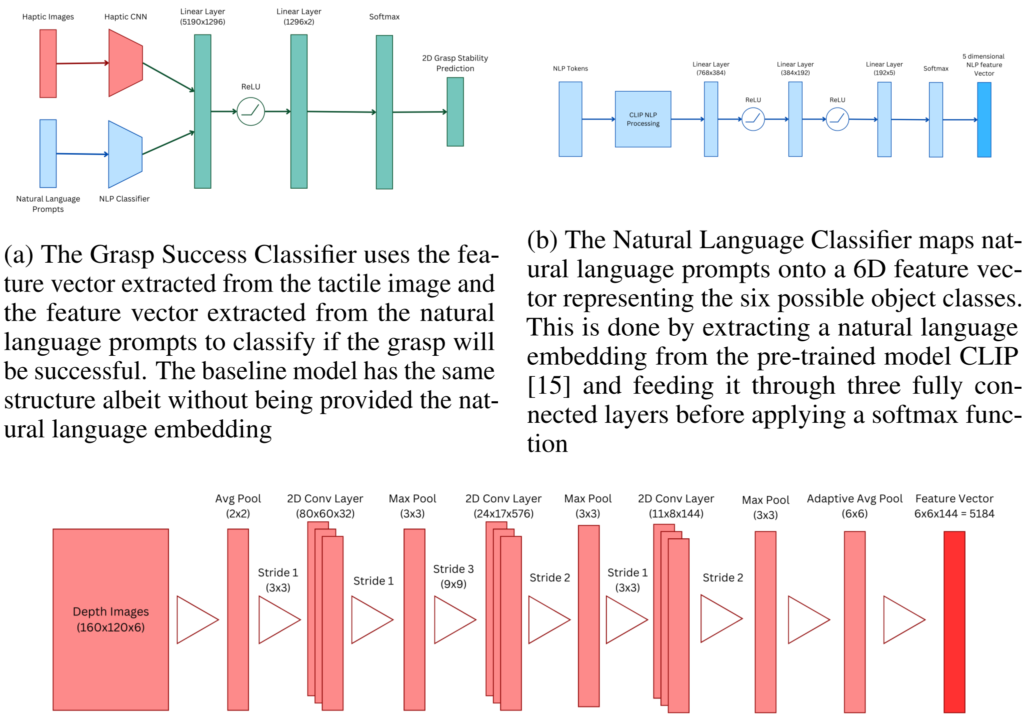
Each grasp is performed from the top down, by moving to a pre-grasp location above our desired grasp point and then moving downwards along the global z-axis. Grasp points are uniformly sampled within the object's bounding box. The gripper's rotation r around the global z-axis is also uniformly sampled with $r \in (0°, 180°)$.

To generate ground truth data, we save the tactile profile $x_{g_{ij}}$ after grasping and evaluate whether or not the grasp was successful by attempting to lift the object. We do this by comparing the location of the grasped object before and after the lift movement. If the difference in position along the global z-axis is sufficiently similar to the lift distance $y_{g_{ij}} = 1$ otherwise $y_{g_{ij}} = 0$. After each full grasp interaction, the object and robot poses are reset.

The language dataset D_l contains approximately 70 descriptions of each of the six objects. The descriptions are usually one to two-word names and synonyms for the object in question including both elaborate descriptions such as "picture-taking device" as well as short terms such as "cam".

We combine the language and tactile modalities during training and testing by randomly selecting a description from the object-specific subset of language descriptions for each grasp sample.

Since most grasp attempts fail due to either missing the object completely, colliding with it before a grasp could be established, or slipping off of the object during the lift, the dataset resulting from this process is imbalanced. Only about 5% of all grasp attempts in the mixed dataset are successful, with success rates between 7.8% for D_{s_4} (Soda Can) and 2.5% for D_{s_3} (Hair Dryer) for the object-specific sub-datasets.



(a) The Grasp Success Classifier uses the feature vector extracted from the tactile image and the feature vector extracted from the natural language prompts to classify if the grasp will be successful. The baseline model has the same structure albeit without being provided the natural language embedding

(b) The Natural Language Classifier maps natural language prompts onto a 6D feature vector representing the six possible object classes. This is done by extracting a natural language embedding from the pre-trained model CLIP [15] and feeding it through three fully connected layers before applying a softmax function

(c) The Tactile CNN extracts tactile features from two tactile RGB images generated by the DIGIT [10] sensors that the robot is equipped with. The RGB channels of each image are appended, resulting in a 6-channel input image. An initial average pool reduces the dimensions of each image before it is processed by three 2D convolutional layers. After each convolutional layer, we apply max pooling. Finally, an adaptive average pool is used to map the feature vector to a universal size that can be fed into the Grasp Success Classifier.

Figure 1: The three model components. (a) an overview of the entire model, (b) the structure of the NLP Classifier, and (c) the architecture for the tactile CNN

3 Models

We compare a baseline model, which classifies grasp success based entirely on tactile data, and an NLP-supported model which is given a natural language prompt in addition to the tactile data. Both models have the same feature extraction network encoding DIGIT's tactile images into a feature vector (see Section (c) of Figure 1).

We use ReLU as non-linear activation function and batch normalization for regularization. The flattened feature vector generated by this model is then fed into the Grasp Success Classifier as seen in section (a) of figure 1.

While the baseline classifier works with only this feature vector, we append a six-dimensional natural language feature vector for our NLP-based classifier. To generate this vector we use CLIP [15], specifically the "ViT-L/14" model, as a pre-trained NLP unit to extract our initial prompt embeddings. We selected CLIP as the network is trained on vision and language data and may therefore be able to capture some geometric context within the language embedding. The embeddings produced by CLIP are then processed by the Natural Language Classifier shown in section (b) of figure 1. The result is a six-dimensional vector with each dimension corresponding to one of the six objects. We chose to further process the CLIP embeddings to ensure that the Natural Language Classifier encodes which object is currently being grasped.

Both the NLP-based model and our baseline then use a fully connected network to classify whether a grasp was successful or not based on their respective feature vector inputs. The structure of this

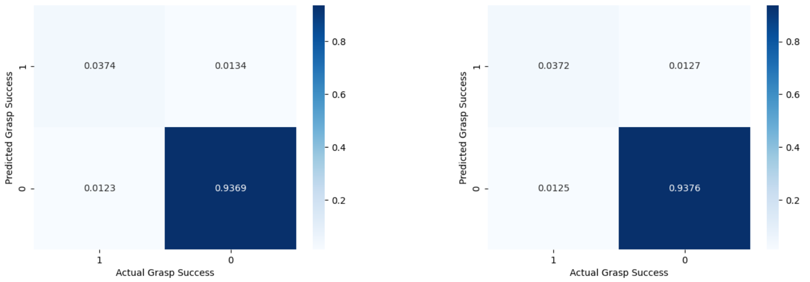


Figure 2: Confusion matrices for the experiments conducted on the full simulated dataset. On the left is the confusion matrix for our baseline model, while the right confusion matrix was generated by our NLP-supported classifier. The confusion matrices were calculated by averaging over all five folds of our cross-validation scheme. We can see that the baseline is slightly more biased toward positive predictions than the NLP-supported classifier but these differences are minimal.

network can be seen in section (a) of figure 1. Note that the "NLP Classifier" module shown in this image is missing for our baseline model.

3.1 Training

Both the baseline approach and the NLP-based grasp classifier are trained using weighted Negative Log Likelihood and Stochastic Gradient Descent. The initial learning rate is 0.002 to which we apply exponential decay with a gamma of $\gamma = 0.9$ every epoch. The model is trained for 50 epochs. We use a pre-trained Natural Language Classifier. During the training of the NLP-based model, we freeze the weights belonging to the underlying CLIP feature extractor but train the rest of the model.

For our NLP-based model, we pre-train the Natural Language Classifier to identify object classes based on input prompts. This pre-training is conducted with a cross-entropy loss and stochastic gradient descent as the optimizer. The Learning rate is set to an initial value of 0.0002 and decays exponentially every epoch with a gamma value of 0.99. Training is conducted on the entire prompt dataset over a total of 200 epochs. During training, we regularize the model by applying Dropout to the first two linear layers.

4 Results and Conclusions

We evaluate and compare our models based on their f1 score to offset the highly imbalanced dataset. The model is trained in a 5-fold cross-validation scheme. The results presented are calculated by taking the mean performance over all five cross-validation folds.

Performance on mixed dataset Since we aim to explore if and how integrating natural language context into a tactile grasping model is beneficial, we first tested how our NLP-supported classifier performs compared to our baseline. We find that the latter peaks at an f1 score of ≈ 0.745, while the former improves only marginally on this result with an f1 score of ≈ 0.747.

Performance on single object datasets Along with the bimodal inputs, we trained and evaluated our baseline model on the data subsets $\{D_{s1}, D_{s2}, D_{s3}, D_{s4}, D_{s5}, D_{s6}\}$. We then averaged the performance of the six resulting models and compared the mean performance of those models with the baseline performance on the mixed dataset D_s. The experiment shows that the mean f1 score of the six single-object models reaches a peak f1 score of 0.746, which is not a significant improvement over the baseline. This implies that the baseline is already as good as a collection of models that are each fitted to only one object shape. The results suggest that the tactile information is already capturing a large amount of object information, and the language does not help to provide more useful context. It may be useful to identify objects that have similar grasps in terms of tactile observations, but different grasp outcomes, e.g., grasping a small marble versus grasping the spherical tip of a drumstick.

In conclusion, the NLP-supported model currently does not significantly outperform our baseline on the simulated dataset. We believe that using more realistic objects varying in weight, surface friction, and/or rigidity will give us the opportunity to encode more relevant features in the language component of our dataset. The features we currently encode could also become more relevant if we change the task at hand from simply predicting grasp stability to generating full grasping solutions.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[2] Yasemin Bekiroglu, Janne Laaksonen, Jimmy Alison Jorgensen, Ville Kyrki, and Danica Kragic. Assessing grasp stability based on learning and haptic data. *IEEE Transactions on Robotics*, 27(3):616–629, 2011.

[3] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, 2018.

[4] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`, 2016–2021.

[5] Hao Dang and Peter K Allen. Stable grasping under pose uncertainty using tactile feedback. *Autonomous Robots*, 36:309–330, 2014.

[6] Guoguang Du, Kai Wang, Shiguo Lian, and Kaiyong Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 54(3):1677–1734, aug 2020.

[7] Javier Felip, Jose Bernabé, and Antonio Morales. Contact-based blind grasping of unknown objects. In *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pages 396–401. IEEE, 2012.

[8] Sascha Fleer, Alexandra Moringen, Roberta L. Klatzky, and Helge Ritter. Learning efficient haptic shape exploration with a rigid tactile sensor array. *PLOS ONE*, 15(1):1–22, 01 2020.

[9] Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. *arXiv preprint arXiv:2303.12076*, 2023.

[10] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[11] Jason Xinyu Liu, Ziyi Yang, Ifrah Idrees, Sam Liang, Benjamin Schornstein, Stefanie Tellex, and Ankit Shah. Lang2ltl: Translating natural language commands to temporal robot task specification. *arXiv preprint arXiv:2302.11649*, 2023.

[12] Adithyavairavan Murali, Yin Li, Dhiraj Gandhi, and Abhinav Gupta. Learning to grasp without seeing. In *International Symposium on Experimental Robotics*, pages 375–386. Springer, 2018.

[13] Rhys Newbury, Morris Gu, Lachlan Chumbley, Arsalan Mousavian, Clemens Eppner, Jürgen Leitner, Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, Dieter Fox, and Akansel Cosgun. Deep learning approaches to grasp synthesis: A review. *IEEE Transactions on Robotics*, page 1–22, 2023.

[14] PaulPauls. franka emika panda pybullet, 2021.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[16] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[17] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[18] Stephen Tian, Frederik Ebert, Dinesh Jayaraman, Mayur Mudigonda, Chelsea Finn, Roberto Calandra, and Sergey Levine. Manipulation by feel: Touch-based control with deep predictive models. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 818–824. IEEE, 2019.

[19] Matthew R Walter, Sachithra Madhaw Hemachandra, Bianca S Homberg, Stefanie Tellex, and Seth Teller. Learning semantic maps from natural language descriptions. Robotics: Science and Systems, 2013.

[20] Shaoxiong Wang, Mike Lambeta, Po-Wei Chou, and Roberto Calandra. TACTO: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters (RA-L)*, 7(2):3930–3937, 2022.