International Symposium on Experimental Robotics (ISER 2020)

Tactile-based Self-supervised Pose Estimation for Robust Grasping

Padmaja Kulkarni, Jens Kober, and Robert Babuska

Abstract

We consider the problem of estimating an object's pose in the absence of visual feedback after contact with robotic fingers during grasping has been made. Information about the object's pose facilitates precise placement of the object after a successful grasp. If the grasp fails, then knowing the pose of the object after the grasping attempt is made can also help re-grasp the object. We develop a data-driven approach using tactile data that computes the object pose in a self-supervised manner after the object-finger contact is established. Additionally, we evaluate the effects of various feature representations, machine learning algorithms, and object properties on the pose estimation accuracy. Unlike other existing approaches, our method does not require any prior knowledge about the object and does not make any assumptions about grasp stability. In experiments, we show that our approach can estimate object poses with at least 2 cm translational and 20 degrees rotational accuracy despite changed object properties and unsuccessful grasps.

1 Introduction

Touch is a prime human sense, which enables humans to grasp in dynamic environments with low lighting, occlusions, and shadows. Similarly, for robot hands, the use of tactile information has proven to be essential for grasping [1–4]. In prior works, tactile information was used to estimate object properties, object-classes, and grasp stability [1–4]. This interpretation aids in the determination of, for example, grasping strategy or contact-force computation. However, it adds little insight into the pose of an object after robot-object interaction happens.

Not being able to have reliable object pose detection after the robot-fingers are in contact with the object is one of the major reasons for the still moderate grasping

Department of Cognitive Robotics, TU Delft, Delft, The Netherlands
e-mail: (P.V.Kulkarni,J.Kober,R.Babuska)@tudelft.nl

performance of robot hands [5–7]. When fingers make contact with an object, the force applied by them can cause the object's location to change. Furthermore, it may no longer be possible to use a camera to monitor the object's pose in hand due to occlusion. Knowing this pose helps to place the object precisely after picking it up or to re-grasp if the grasp fails. Thus having a method for reliable in-hand object pose estimate is essential.

In this paper, we address the problem of estimating object pose, especially when visual feedback is not possible *during or after the grasp*. We develop a data-driven approach based on tactile data to estimate an object's pose (position and orientation) relative to the robotic gripper.

Tactile sensors provide only local information about the object in contact, whereas vision sensors provide extensive environmental information. Hence, to simplify the problem of object pose estimation with tactile sensors, either assumptions like grasp stability or known object properties (e.g., mesh-model, geometry) are made, or the object is fixed to restrict its motion under gripper forces. For example, Bimbo et al. [8] estimated the object's pose in a robotic gripper using tactile and force data assuming the object's model or geometrical information is accurately known. They minimized the difference of angles between the normals on the object's surface obtained from the geometric model and the observed force-normals to estimate the object's pose using evolutionary algorithms. Corcoran et al. [9] used a particle filter to estimate object pose in the robot gripper. The authors' model computed the likelihood of true contact measurement over possible contact positions, assuming a known object model. The approaches in [10, 11] used *haptic exploration* for object pose prediction with tactile data for a fixed object and assumed the object model to be known. The authors of [5] computed the object's pose with data-driven techniques, using an under-actuated robotic gripper and assuming a stable grasp.

In contrast to these approaches, our paper focuses on directly estimating the object pose based on tactile data without making any assumptions about the object model or grasp stability, and without fixing the object. Furthermore, in the training phase, we use a camera to estimate the ground truth object pose, and thereby generate the target object pose in a self-supervised manner. Note that, for the trained model, the camera is only used to estimate the initial object-pose and *it is not used thereafter during or after the grasp attempt.* The initial object pose facilitates the estimation of changes in the object pose, especially for the symmetric objects, as tactile sensing would not discriminate if the object were rotated by, for example, 180° degrees.

This paper's contribution is two-fold: i) a novel approach for computing Tactile-based Self-supervised Pose Estimation (T-SPoE), and ii) evaluation of the effect of various feature representations, machine learning algorithms, and changing object properties on the accuracy of the pose estimation.

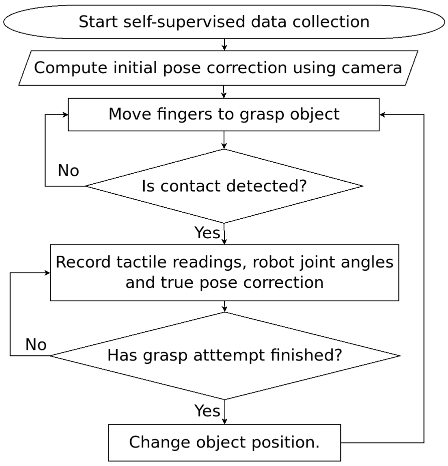
Start self-supervised data collection
Compute initial pose correction using camera
Move fingers to grasp object
Is contact detected?
No
Yes
Record tactile readings, robot joint angles and true pose correction
Has grasp atttempt finished?
No
Yes
Change object position.

Fig. 1 Data collection with T-SPoE method. Data is recorded continuously during each grasp attempt, and the labels are generated in a self-supervised manner.

2 Tactile-based Self-supervised Pose Estimation

T-SPoE estimates the object pose based on tactile data. It consists of the following three phases:

1. **Data Collection:**
   Figure 1 shows our data collection approach. We use a gripper equipped with tactile sensors and design an experimental setup allowing object grasping. We use a camera to estimate the target object pose in a self-supervised manner. The gripper is in the open position and tries to grasp the object until it is successful (shown in Fig. 2) and then places the object at the same place. During this phase, we record true object pose, tactile data, and finger joint angles at each time step. For recording the true object pose, we mount Fiducial markers on the object and use the ArUco library [12] to compute the marker pose.

2. **Model Training:**
   In this phase, based on the collected data, we train models to estimate the pose using machine learning. A different model is trained for every object. The tactile data is preprocessed to obtain data-representations explained in Section 3. The input to the model is the preprocessed tactile data, the finger joint angles, and the object's pose before grasping (*initial object pose*). The output of the model is the 6D object pose relative to the gripper.

3. **Pose Estimate:**
   In this phase, we process the current tactile data frame to obtain data-representations, as explained in Section 3. We then use the processed tactile data frame as input to the trained model to predict the object pose.

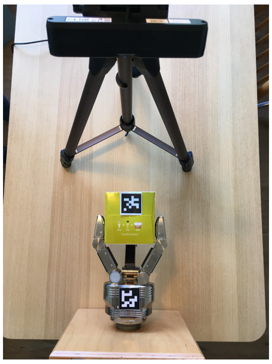

Fig. 2 Setup for collecting data. Two fingers of the SDH hand are used to grasp an object. The Asus Xtion PRO camera tracks the marker positions continuously during grasp attempts.

3 Feature Representations

To train the model using machine learning, we use tactile data, finger joint angles, and initial object pose as input features. The initial object pose is a 6D vector representing the initial object pose with respect to the hand, and it is estimated using a camera. The object pose consists of the 3D Cartesian position and three roll, pitch and yaw Euler angles with XYZ version. We use two fingers of the Schunk Dexterous Hand (SDH),[1] which have in total four degrees of freedom. Along with this 10D vector, we use the tactile data from two fingertips of the SDH hand. Each fingertip is equipped with a Weiss Tactile pad with 13×6, i.e., 78 taxels. These total of 156 taxels output *pressure* at the fingertips, which we use as our tactile input.

The tactile feature length of 156 is large and could be potentially redundant. To reduce the dimensionality, along with raw tactile-values, we use the following feature representations and train three different machine learning models per technique:

1. **Tactile Data as a List:**
   Here, we use tactile data as a list, where one data point has a dimension of 156×1, and the input feature vector has a total dimension of 166.

2. **Max-pooled Tactile Data:**
   We use this operation to reduce the dimensionality of tactile data, with a rectangular filter of size 1×6 and stride 1. This entails that we take the maximum tactile value from each taxel row. As there are 13 rows, we reduce the dimensions to 13×1 for one finger. The input tactile feature vector, in this case, has a total dimension of $13 \times 2 = 26$.

3. **Principal Component Analysis (PCA):**
   Here, we take co-variance (C) of N tactile data vectors (T) [13]: $C = \sigma(T_{1\ldots N})$.

[1] `https://schunk.com/nl_en/gripping-systems/series/sdh/`

The most significant feature vectors, which explain 80% of data, are computed by computing the largest eigenvalues (λ) as: $v^T Cv = \lambda$, with v as the eigenvector of matrix C. In our case, PCA was able to reduce the dimensions to 20×1.

4 Experiments

A setup for data collection is shown in Figure 2. The Asus Xtion PRO camera is used to track the object's relative position with respect to the SDH hand..

1. **Self-supervised Data Collection:**
   We used two objects, a Nescafe box and a ceramic cup for data collection. The object is kept between the hand's open fingers such that object-finger interaction or grasping is possible. The orientation of the object is allowed to change by 360° about the vertical axis. The SDH grasps the object by closing the fingers until either the desired force value is reached or the finger joint-limits are reached. *A grasp is successful when the gripper grasps an object and fails when the object slips out of the gripper.* The objects used for the experimentation are listed in Table 2. For training, the Nescafe box and cup is used. We record the data for 40 successful and 40 unsuccessful grasps for each object. After the training data-collection is complete, the Nescafe box and the cup are enveloped in 1 cm thick foam. Enveloping an object with foam changes the object's dimensions and surface properties, e.g., friction coefficient. The aim of experimenting with the objects covered with foam is to evaluate if the method is robust to such property changes. *The foam-covered objects are only used for evaluation purposes and not during training.*

2. **Model Training:**
   For model training, we use four machine learning methods; namely, i) Neural Networks (NN), ii) Support Vector Regression (SVR), iii) K-Nearest Neighbours (K-NN), and iv) Random Forest (RF). The parameters used for these learning algorithms are listed in Table 1. These parameters are decided by trial and error using the highest model score as a performance indicator across 5-fold cross-validation. The parameter details are available in Scikit-learn library[2].

Table 1 Parameters used for the evaluation of various Machine Learning Approaches.

Algorithms	Parameters
NN	Hidden layer sizes: (100, 75, 50, 25, 20, 10)
SVR	C: 1.0, epsilon: 0.2
KNN	Nearest neighbours: 3
RF	Max. depth: 6

[2] https://scikit-learn.org/stable/

Table 2 Objects and their sizes used for the experimentation with SDH hand.

Objects	Sizes in cm
Box	13 × 13 × 13 (h × w × l)
Cup	8 × 10 (d × h)
Box with Foam Layer	13 × 15 × 15 (h × w × l)
Cup with Foam Layer	10 × 10 (d × h)

Table 3 RMSE in pose prediction. Translation (T-error) is in centimeters, and Rotation (R-error) is in degrees.

Features	Type	NN		SVR		KNN		RF	
		T-error	R-error	T-error	R-error	T-error	R-error	T-error	R-error
Tactile data as a list	Box grasped	0.46	1.01	0.82	0.27	0.15	0.8	0.39	1.4
	Cup grasped	1.14	9.4	0.76	0.42	0.8	5.1	0.65	8.7
	Box grasp failed	1.3	6.89	1.09	0.36	0.55	8.9	0.94	4.4
	Cup grasp Failed	2.2	9.7	3.03	1.01	2.07	11.06	1.91	8.76
Max pooled tactile data	Box grasped	0.79	1.9	0.8	0.26	0.66	2.6	0.59	1.60
	Cup grasped	1.1	9.6	0.85	0.28	0.46	5.2	0.36	3.5
	Box grasp Failed	1.3	6.2	0.97	0.32	0.6	9.5	0.8	4.2
	Cup grasp failed	2.1	9.4	2.7	0.92	1.9	9.6	1.7	8.2
PCA compressed tactile data	Box grasped	2.2	1.8	0.62	0.21	0.34	2.1	0.9	2.7
	Cup grasped	1.1	8.2	0.7 1	2.5	0.79	5.6	0.70	8.5
	Box grasp failed	1.7	3.1	0.74	2.4	0.6	13.34	0.69	1.69
	Cup grasp Failed	2.2	8.7	2.3	0.78	1.9	10.2	1.07	8.16

5 Results and Discussion

The pose is predicted using the data collected in Phase 1 and models from Phase 2 of T-SPoE. These results are divided into two grasp modes, depending upon whether the SDH hand grasped the object. According to feature representations and algorithms, the results are listed in Table 3. We use Root Mean Square Error (RMSE) to measure the prediction accuracy. For translational RMSE, RMS (Root Mean Square) distance between the predicted and the target position is computed. For rotational error, RMS error between the predicted and the target Euler angles is calculated. Here, we could use RMS distance between angles in Euler space because we vary the object orientation only about the vertical axis.

The average RMSE of each feature representation does not significantly vary. Hence, we use raw tactile-values in the list form as features. The lowest average RMSE for each feature representation is obtained by using Random Forest Regression. Therefore we use RF for the training phase for T-SPoE. The results obtained by changing the object surface properties and deformability are listed in Table 5. *Note that when the size and deformability of an object are varied, no additional training is performed.*

We calculated the error between the initial (pre-grasp) object pose and the actual object pose after gripper-object interaction to justify our method's usability. This error indicates how much the object has shifted due to the robot-object interaction. In the case of the box in the three scenarios — box grasped, box grasp failed, and box with foam — this average RMSE was found to be 13.64 cm in translation and 21.67° in rotation. For the cup, this error was 12.80 cm and 39.43° in translation and rotation, respectively. Thus this pose-shift is significant, and estimating the object pose after contact with robotic fingers during grasping has been made is necessary for successful manipulation.

In summary, T-SPoE can estimate pose relative to the gripper with RMSE less than 0.5 cm in translation and 3.5° in the rotation when the object is successfully grasped. When the gripper is unable to grasp an object, RMSE increases to 2 cm in translation and 8° in rotation. Moreover, for object size and property changes, T-SPoE was able to maintain the translational RMSE up to 1.9 cm. However, the rotational RMSE increases to 20°. This can be attributed to the fact that the finger-object contact dynamics change significantly for deformable objects. However, this error in pose prediction is less than the error that would occur without using T-SPoE if the object is assumed to be at the initial pre-grasp pose. We conclude that training with varying object-properties is required to make the method robust against the object-property variations and to reduce the error further.

Table 4 RMSE in prediction of the pose with variations in object properties using T-SPoE. Random Forest Algorithm is chosen, as it has the lowest average RMSE with known objects.

Objects	RF	
	T-error (cm)	R-error (deg)
Box with Foam	1.90	20.1
Cup with Foam	1.74	10.11

6 Conclusion and Future Work

We proposed a self-supervised method T-SPoE for robust grasping tasks using tactile feedback. We evaluated self-supervised object pose prediction using data collected from a fixed SDH hand with two known objects. Our approach does not require any prior object knowledge and does not make any assumptions about the grasp stability. Additionally, we evaluate the effect of various feature representations, machine learning algorithms, and changing object properties on the proposed approach. With the known objects, our method predicted the object pose with RMSE less than 0.5 cm in translation and 3.5° in the rotation when the object is successfully grasped. With changed object properties, this error increases to 1.9 cm and 20° in translation and rotation, respectively. The next step to reduce this error would be to use more training data with changing object properties to make T-SPoE robust against object property variations.

Furthermore, in the future, we plan to evaluate the complete manipulation pick-and-place pipeline with more day-to-day objects of irregular natural shapes, like lettuces, tomatoes, and chicken pieces.

Acknowledgment

This research is funded by the Netherlands Organization for Scientific Research project Cognitive Robots for Flexible Agro-Food Technology, grant P17-01.

References

1. Calandra, R., Owens, A., Upadhyaya, M., Yuan, W., Lin, J., Adelson, E.H., Levine, S.: The Feeling of Success: Does Touch Sensing Help Predict Grasp Outcomes? IEEE Robotics and Automation Letters (RA-L) **3**(4) (2017) 3300–3307
2. Liarokapis, M.V., Calli, B., Spiers, A.J., Dollar, A.M.: Unplanned, model-free, single grasp object classification with underactuated hands and force sensors. In: IEEE International Conference on Intelligent Robots and Systems (IROS). Volume 2015-Decem., Institute of Electrical and Electronics Engineers Inc. (2015) 5073–5080
3. Kwiatkowski, J., Lavertu, J.S., Gourrat, C., Duchaine, V.: Determining object properties from tactile events during grasp failure. In: IEEE International Conference on Automation Science and Engineering (CASE). Volume 2019-Augus. (2019) 1692–1698
4. Chebotar, Y., Hausman, K., Su, Z., Sukhatme, G.S., Schaal, S.: Self-supervised regrasping using spatio-temporal tactile features and reinforcement learning. In: IEEE International Conference on Intelligent Robots and Systems (IROS). (2016) 1960–1966
5. Liarokapis, M., Dollar, A.M.: Learning the post-contact reconfiguration of the hand object system for adaptive grasping mechanisms. In: IEEE International Conference on Intelligent Robots and Systems (IROS). (2017) 293–299
6. Bütepage, J., Cruciani, S., Kokic, M., Welle, M., Kragic, D.: From Visual Understanding to Complex Object Manipulation. Annual Review of Control, Robotics, and Autonomous Systems **2**(1) (2019) 161–179
7. Okamura, A., Smaby, N., Cutkosky, M.: An overview of dexterous manipulation. In: IEEE International Conference on Robotics and Automation (ICRA). Volume 1., IEEE (2000) 255–262
8. Bimbo, J., Kormushev, P., Althoefer, K., Liu, H.: Global estimation of an objects pose using tactile sensing. Advanced Robotics **29**(5) (3 2015) 363–374
9. Corcoran, C., Platt, R.: A measurement model for tracking hand-object state during dexterous manipulation. In: Proceedings - IEEE International Conference on Robotics and Automation (ICRA). (2010) 4302–4308
10. Klaas, G., Bruyninckx, H.: Markov Techniques for Object Localization With Force-Controlled Robots. In: International Conference on Advanced Robotics (ICAR). (2001) 91–96
11. Schaeffer, M.A., Okamura, A.M.: Methods for intelligent localization and mapping during haptic exploration. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC). Volume 4. (2003) 3438–3445
12. Romero-Ramirez, F.J., Muñoz-Salinas, R., Medina-Carnicer, R.: Speeded up detection of squared fiducial markers. Image and Vision Computing **76** (2018) 38–47
13. Laaksonen, J., Kyrki, V., Kragic, D.: Evaluation of feature representation and machine learning methods in grasp stability learning. In: IEEE-RAS International Conference on Humanoid Robots, Humanoids. (2010) 112–117