Tactile-visuo Search in Cluttered Spaces

Michael Lin
Mechanical Engineering
Stanford University
Stanford, USA
mlinyang@stanford.edu

***Abstract*—Robots executing tasks in home environments face the challenge of large uncertainty over the space they need to interact with. In this work we look into the problem of reaching into a cupboard-like environment to fetch a spice jar that could be occluded by other objects in the scene. We are interested in integrating both vision and tactile measurements in order to create a map of the environment, which can be used to inform where the target object is and also used to create motion plans to execute the task. We investigate using Gaussian Process Implicit Surfaces (GPIS) for integrating measurements from both sensing modalities into one common representation of the scene. In particular, we use vision to construct a prior and tactile measurements subsequently to update this GPIS map.**

***Index Terms*—Manipulation under uncertainty, tactile exploration, perception for grasping and manipulation.**

I. Introduction

There is still a large gap between the the ability of robots to execute tasks in unstructured environments in contrast to that of humans. For instance, we are able to seamlessly open a cupboard in the kitchen and take a "quick peek" at what is inside, then reach in with our hands and navigate through a clutter of objects only using our sense of touch until finding a desired object. For a robot, this is a daunting tasks as there is a lot of uncertainty to reason over, sensing needs to be multi-modal as both vision and tactile sensing might be needed to execute efficiently, and plans to reach the object might be complex sequential actions. Although there are multiple parts to this problem, the scope of this work focuses on perception and state estimation of a scene from visual and tactile measurements.

Visual and tactile sensing are very different in nature, so finding a common ground for representing these measurements is important. A spatial representation is a good option, but there are many flavors of representations such as voxel [1], superquadrics [2], point clouds [3], but Dragiev et al. shows that Gaussian Process Implicit Surface (GPIS) has many benefits for this application as it can be used to incorporate point clouds from depth camera measurements and it is also favorable for tactile measurements as it can use both contact points and contact normals for updating the representation [4]. In addition, GPIS inherently encodes variance in the spatial representation, so we can use this confidence map to determine regions that have been less explored to determine next robot motion, or use this map to guess where hidden objects may be if we cannot identify them at first sight. Finally, GPIS can be approximately used as a signed-distance function to obstacles which is useful for finding collision-free paths for motion planning.

In this work we will develop a system that will construct a prior belief from an initial depth view of a cupboard with multiple objects (illustrated in Fig. 1), and we will update this GPIS from subsequent tactile measurements. A difference from previous work is that we will be building off from our custom made robot gripper [5] that can perform non-intrusive contacts with objects. The advantage of non-intrusive contacts is that as the robot reaches into the cupboard, it can gather contact measurements of objects without disturbing their state, thus, preserving the initial depth camera measurements. In general, allowing for contact sensing that do not disturb the state of the environment makes measurements more coherent and estimates more accurate.

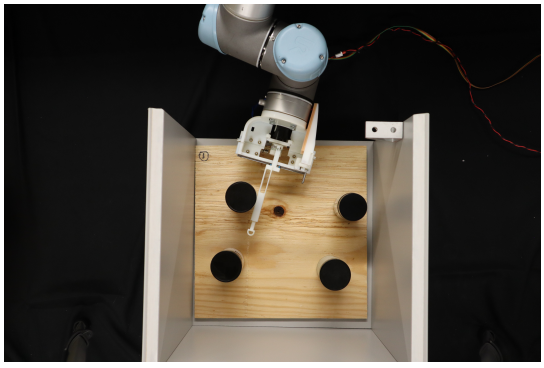

Fig. 1. Exploratory Hand reaching into a cupboard and touching free-standing objects.

II. Previous Work

The problem of reaching into a cabinet of free-standing objects to retrieve an object among clutter (also known as Mechanical Search) has been done recently by Huang et al. [6]. Part of their contribution was to develop a perception model to predict an occupancy distribution map of a target object given a depth image of the scene [7]. The drawback of this approach is that as the robot arm reaching in it blocks line-of-sight to the scene, so after every action on the scene the robot arm has to retract to take new sensor measurements. We are interested in improving this work by enabling a robot arm to use both vision and touch such that it will only need to reach into the cabinet once without retracting, and search for a target object based on touch.

Occupancy distribution maps are useful for inferring where a hidden target object could be but they are not good for tracking the state of the scene (i.e. the shape of the environment). For this we can use GPIS which have been shown to be good candidates for incorporating tactile [4] and visual measurements [8]. Most recently, Suresh et al. demonstrated using GPIS in a Simultaneous Localization and Mapping (SLAM) problem to estimate the location and shape of a block being pushed on a surface from tactile measurements [9]. Although the results shown are impressive, it is a complex system that runs a heavy SLAM algorithm (iSAM2) and Bayesian regression, and results of the physical experiment rely on good models of sliding object dynamics on a surface which is difficult to predict accurately. The common problem of most of these previous work is that contacts are an intrusive way of sensing; the act of sensing perturbs the state of the object. This results in needing to do localization and shape sensing simultaneously. However, non-intrusive contacts are also achievable but it is necessary to approach the problem from low-level design as shown by Lin et al. [5].

Using Exploratory Hand [5] to sense through contacts, we will be able to achieve more accurate estimates of a scene since the state of the environment will not be changing. Also, we will be able to scale to an entire scene as opposed to just a single object as most of previous work have done.

III. Problem Formulation

A. Gaussian Process Implicit Shape with Vision and Touch

We need the GPIS update equations for both vision and tactile measurements. Depth image provide us with 3D point measurements of object surfaces. Since for the task of interest we only have line-of-sight initially, we will process these depth measurements into an initial GPIS (our prior). These point measurements can be turned into a GPIS similar to how previous work have done [4], [10] using the following regression equations

$$\bar{f}_* = \mathrm{k}_*^T (K + \sigma_{n,d}^2 \mathrm{I})^{-1} \mathrm{y} \tag{1}$$

$$\mathcal{V}[f_*] = k(\mathrm{x}_*, \mathrm{x}_*) - \mathrm{k}_*^T (K + \sigma_{n,d}^2 \mathrm{I})^{-1} \mathrm{k}_* \tag{2}$$

where k is the covariance function, k_* is the vector of covariances between the test point and all observation points, K is the covariance matrix, $\sigma_{n,d}^2$ is the depth measurement variance, y is the observed values and x_* are the test points (for which we use a grid map for plotting purposes).

Similar to [11], we use a thin plate model for the kernel function

$$k(x^i, x^j) = 2||x^i - x^j||^3 - 3\psi||x^i - x^j||^2 + \psi^3 \tag{3}$$

where ψ is a hyperparameter but can be easily specified as the farther distance among all the training inputs (i.e. $max(||x^i - x^j||), \forall x^i, x^j \in X$). This kernel function is advantageous because it provides for a good first-order continuity as opposed to a exponential kernel function. Illustratively, it makes the mean value outside of the cluster of training points taper off away from zero whereas an exponential function drives the mean to zero (which is a problem since our 2D shapes are determined using the 0-level set) [12].

To incorporate subsequent tactile measurements we can also use equations 1 and 2, but some modifications need to be done in computing the covariance between points x^i and normal measurements ω^j, and normal measurements ω^i and normal measurements ω^j. As shown by Dragiev et al. and Li et al., the contact point and contact normal measurements from tactile sensors provide a dual update of the mean of the GPIS and the gradient at those points, respectively [4], [11]. The covariance function for these two are:

$$\begin{aligned} cov(\omega_m^i, f(x^j)) &= \frac{\partial}{\partial x_m} cov(f(x^i), f(x^j)) \\ &= \frac{\partial}{\partial x_m} k(x^i, x^j) \end{aligned} \tag{4}$$

$$\begin{aligned} cov(\omega_m^i, \omega_n^j) &= \frac{\partial^2}{\partial x_m \partial x_n} cov(f(x^i), f(x^j)) \\ &= \frac{\partial^2}{\partial x_m \partial x_n} k(x^i, x^j) \end{aligned} \tag{5}$$

where $m, n \in \{1, 2\}$, k is the thin plate kernel function, w_m is the contact normal at x_m and w_n is the contact normal at x_n.

In combining the depth and tactile sensing data, we get a training set X with a mix of surface points without normal measurements x_p and surface points with their respective normal measurements x_{pn}. In order to combine both of these in the Bayesian regression we need to re-format the covariance matrix $K \in \mathbb{R}^{n+3m}$, where n is the number of depth sensor points and m is the number of tactile sensor measurements. K is a block diagonal matrix with the following (3x3), (3x1), (1x3) and (1x1) blocks:

$$K_{[x_{pn}^i, x_{pn}^j]} = \begin{bmatrix} \mathrm{cov}(f(x^i), f(x^j)) & \mathrm{cov}(f(x^i), \omega_1^j) & \mathrm{cov}(f(x^i), \omega_2^j) \\ \mathrm{cov}(\omega_1^i, f(x^j)) & \mathrm{cov}(\omega_1^i, \omega_1^j) & \mathrm{cov}(\omega_2^i, \omega_1^j) \\ \mathrm{cov}(\omega_2^i, f(x^j)) & \mathrm{cov}(\omega_1^i, \omega_2^j) & \mathrm{cov}(\omega_2^i, \omega_2^j) \end{bmatrix} \tag{6}$$

$$K_{[x_p^i, x_{pn}^j]} = \begin{bmatrix} \mathrm{cov}(f(x^i), f(x^j)) & \mathrm{cov}(f(x^i), \omega_1^j) & \mathrm{cov}(f(x^i), \omega_2^j) \end{bmatrix} \tag{7}$$

$$K_{[x_{pn}^i, x_p^j]} = K_{[x_p^j, x_{pn}^i]}^T \tag{8}$$

$$K_{[x_p^i, x_p^j]} = \begin{bmatrix} \mathrm{cov}(f(x^i), f(x^j)) \end{bmatrix} \tag{9}$$

In addition to these kernel matrix modifications, it is also possible to model different sensor noise for both sensing modalities. This can be done through using different noise

variance σ_n^2 in equation 2. In general, depth measurements have a much larger noise variance compared to tactile measurements since the latter relies on high accuracy robot proprioception as shown in [5].

IV. Methods

A. Simulated System for Depth and Tactile Sensor Acquisition

For this work we used PyBullet to simulate the environment for the task of reaching into a constrained space as shown in Fig. 2. For scope of this project we only worked with three spice jars that are near each other and we use round and square spice jars in different experiments.

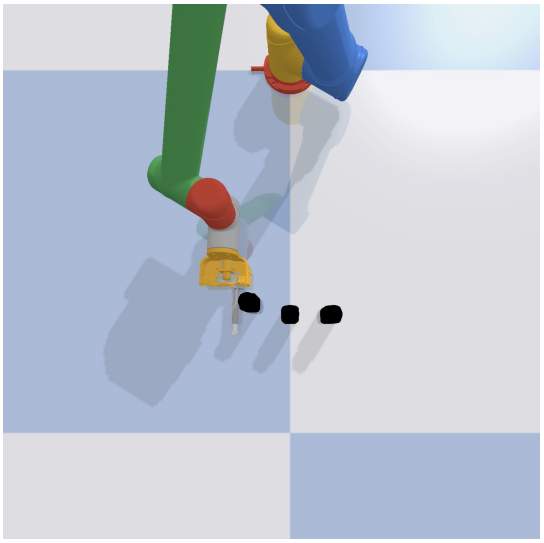

Fig. 2. PyBullet Simulation of a robot manipulator (Universal Robot UR5 with Exploratory Hand end-effector).

1) Camera Measurements: Before reaching towards the objects, we take an initial camera measurement of the group of objects from the robot perspective (Shown in Fig. 3). The 3D point cloud is obtained from the depth measurement by backprojecting the depth camera pixels through the extrinsic and intrinsic camera matrices (details can be found in https://tinyurl.com/ywv5k6w8). Since we are tackling the problem in 2D for now we took a cross-sectional slice from the 3D point cloud and flattened it to 2D point clouds as shown in Fig. 3 bottom left. Finally, we added different amounts of Gaussian White Noise to simulate a real sensor as shown in the figure bottom right.

2) Motion Planning and Tactile Measurements: Tactile measurements were gathered by controlling the robot to make contact with the object and follow different robot end-effector trajectories. The robot arm position was controlled in Cartesian space and the Exploratory finger of the gripper was controlled to render very low impedance such that it would exert small forces when making contact with the objects. To gather tactile measurements we move the gripper to make contact with the long edge of the finger and then control to re-orient the gripper such that the contact point moves on the surface of the object as shown in Fig. 4. It is easy from this figure to see that

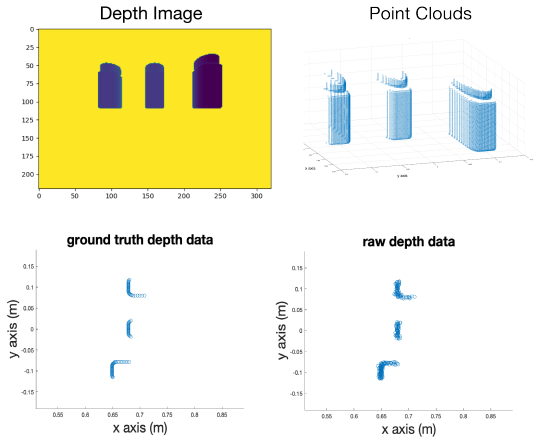

Fig. 3. Depth camera measurements from PyBullet. Top left shows the depth 2D image, top right the converted 3D point cloud by back-projecting, bottom left is a 2D cross section of the 3D point cloud and bottom right is that 2D point cloud with added white noise to simulate real sensor noise.

vision only provides a very limited view of the object shapes and tactile is able to complement these on object regions that cannot be observed. The point and normal measurements are given by the simulator.

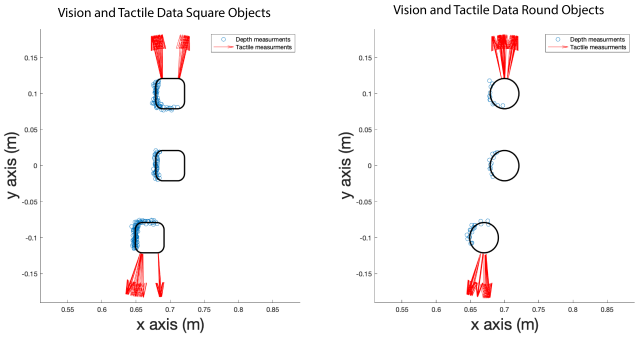

Fig. 4. Combined tactile and visual measurements. Left plot uses square objects and right plot uses round objects. A challenge with square objects is that it creates sparsity on the tactile measurements due to the two objects having flat surfaces.

3) More details on GPIS Implementation: In addition to the visual measurement prior we also use a circular prior (four points equally spaced on the circle perimeter with respective normal vectors). This is helpful as the normal measurements help in starting to shape the GPIS in the Z axis, otherwise the GPIS will just have zero mean as the depth measurements at only point measurements at the 0-level set. For visualization of the GPIS we use a set of evenly distributed points as the test set Y.

V. Results & Discussions

In Fig 5 we can see that only by incorporating the prior depth camera measurements the GPIS already start taking shape of the objects. In the plot to the right we can see

that where the blue colored regions coincide with where we have measurements, thus, lower variance. Note that this initial GPIS is very useful for the task of reaching into clutter as we can easily object regions in the map with high variance (red regions) and use this map to decide where to begin exploring with the robot gripper to gain a better estimate of the environment.

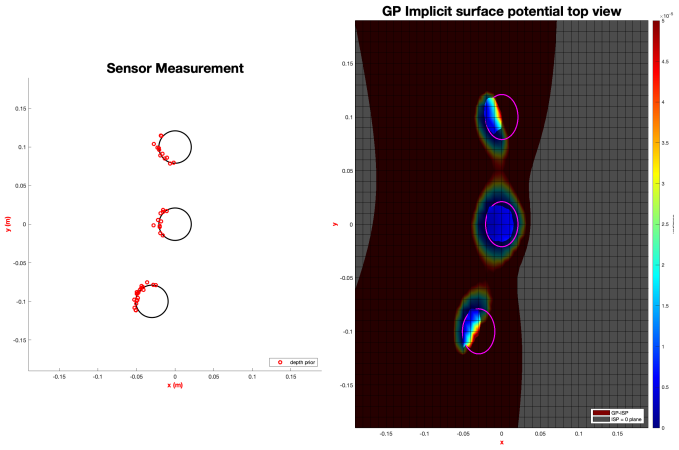

Fig. 5. GPIS prior with only depth data. The 2D shapes are derived from the GPIS such that negative mean values correspond to within the object (in the figure brighter colors correspond to inside the objects).

After incorporating the tactile measurements we get a more complete shape on the round spice jars at the top and bottom as can be seen in Fig. 7. They become much close to the ground truth shape shown in magenta.

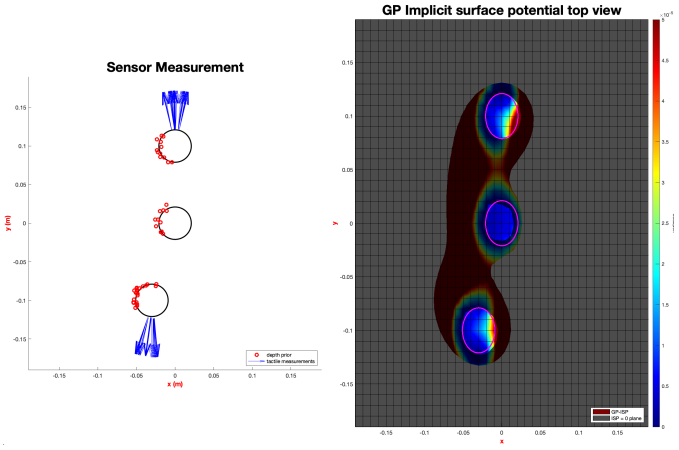

Fig. 6. GPIS with depth and tactile data on round spice jars.

As can be seen from Fig. 7, this method works well when performed on non-circulat objects such as squared spice jars.

Although the examples shown until now worked, there were some combination of parameters that yielded bad results as seen in Fig. 8. One example is changing circular prior mentioned in section B.3. If the circle is made too large compared to the shape of the objects than the shape estimate are very off. This is likely because adding this circular prior

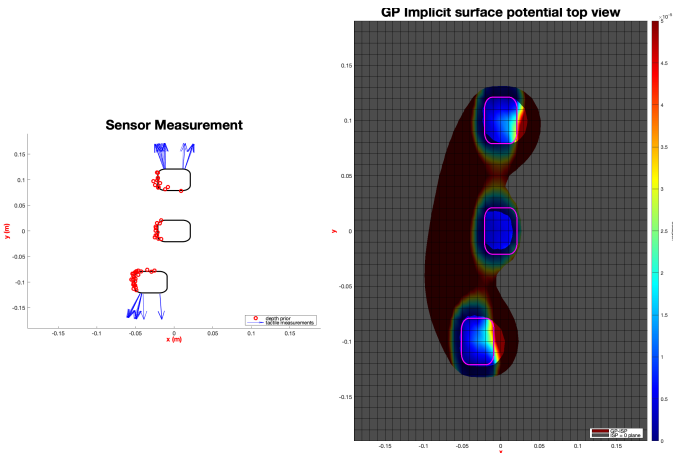

Fig. 7. GPIS with depth and tactile data on round square jars.

is basically faking data to make the GPIS take shape. One possible alternative to using this circular prior is to initialize the mean of the GPIS to some positive value.

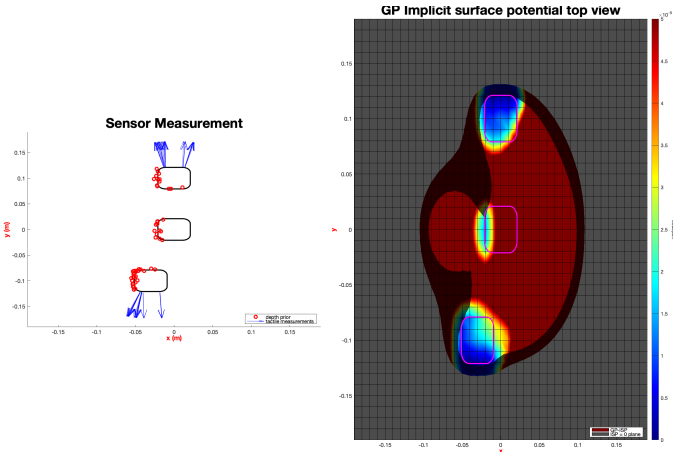

Fig. 8. GPIS with depth and tactile data on round square jars using circular prior of 0.2 m diameter as opposed to 0.03 m diameter for all other plots.

VI. Challenges & Future Work

One of the most noticeable challenges with GPIS is that as more measurements are added it very quickly becomes slow (computation time per iteration is shown in Fig. 9). Although GPIS are very expressive this run-time increase limits the ability to scale to more complex scenes with multiple objects.

Towards the future, in order to improve this computation time, we are interested in the idea of grouping the tactile measurements. In the tactile measurements we gathered it is easy to see that a lot of these points are redundant as we just used raw data and did not do any post processing to make sure we were only choose distinct measurements. However, instead of selecting these points it may be possible to group them into lines or splines and use these units as the measurements. The thin plate kernel function only takes as input the distance between points. If we chose lines as measurements then we

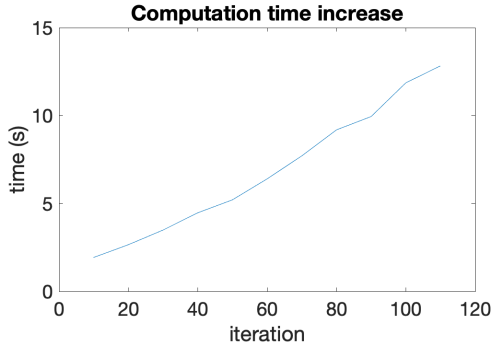

Fig. 9. Time to compute the posterior vs number of iterations or tactile measurements added.

could use distance from points to lines as as input to the kernel function. For example, if our line measurement L^i is $a^i x_1 + b^i x_2 + c^i = 0$, then the kernel function can be

$$k(L^i, x^j) = 2D(L^i, x^j)^3 - 3\psi D(L^i, x^j)^2 + \psi^3 \tag{10}$$

Where D(L, x) is the projection distance from a point to a line.

$$D(L^i, x^j) = \frac{|a^i x_1^j + b^i x_2^j + c^j|}{\sqrt{a^{i2} + b^{i2}}} \tag{11}$$

Although it is easy to use line representations they may challenging to use to represent curves such as round objects. In the near future we will be looking at better alternatives to line representations.

REFERENCES

[1] Jeannette Bohg, Matthew Johnson-Roberson, Mårten Björkman, and Danica Kragic. Strategies for multi-modal scene exploration. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4509–4515. IEEE, 2010.

[2] Pradeep Khosla and Richard Volpe. Superquadric artificial potentials for obstacle avoidance and approach. In *Proceedings. 1988 IEEE International Conference on Robotics and Automation*, pages 1778–1784. IEEE, 1988.

[3] Gregory Izatt, Geronimo Mirano, Edward Adelson, and Russ Tedrake. Tracking objects with point clouds from vision and touch. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4000–4007. IEEE, 2017.

[4] Stanimir Dragiev, Marc Toussaint, and Michael Gienger. Gaussian process implicit surfaces for shape estimation and grasping. In *2011 IEEE International Conference on Robotics and Automation*, pages 2845–2850. IEEE, 2011.

[5] Michael A Lin, Rachel Thomasson, Gabriela Uribe, Hojung Choi, and Mark Cutkosky. Exploratory hand: Leveraging safe contact to facilitate manipulation in cluttered spaces. *IEEE Robotics and Automation Letters*, 2021.

[6] Huang Huang, Marcus Dominguez-Kuhne, Jeffrey Ichnowski, Vishal Satish, Michael Danielczuk, Kate Sanders, Andrew Lee, Anelia Angelova, Vincent Vanhoucke, and Ken Goldberg. Mechanical search on shelves using lateral access x-ray. *arXiv preprint arXiv:2011.11696*, 2020.

[7] Michael Danielczuk, Anelia Angelova, Vincent Vanhoucke, and Ken Goldberg. X-ray: Mechanical search for an occluded object by minimizing support of learned occupancy distributions. *arXiv preprint arXiv:2004.09039*, 2020.

[8] Gabriela Zarzar Gandler, Carl Henrik Ek, Mårten Björkman, Rustam Stolkin, and Yasemin Bekiroglu. Object shape estimation and modeling, based on sparse gaussian process implicit surfaces, combining visual data and tactile exploration. *Robotics and Autonomous Systems*, 126:103433, 2020.

[9] Sudharshan Suresh, Maria Bauza, Kuan-Ting Yu, Joshua G Mangelson, Alberto Rodriguez, and Michael Kaess. Tactile slam: Real-time inference of shape and pose from planar pushing. *arXiv preprint arXiv:2011.07044*, 2020.

[10] Marcos P Gerardo-Castro, Thierry Peynot, Fabio Ramos, and Robert Fitch. Robust multiple-sensing-modality data fusion using gaussian process implicit surfaces. In *17th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2014.

[11] Miao Li, Kaiyu Hang, Danica Kragic, and Aude Billard. Dexterous grasping under shape uncertainty. *Robotics and Autonomous Systems*, 75:352–364, 2016.

[12] Oliver Williams and Andrew Fitzgibbon. Gaussian process implicit surfaces. 2006.