State estimation and object modeling for dexterous manipulation

Sudharshan Suresh
suddhu@cmu.edu

Michael Kaess
kaess@cmu.edu

1 Introduction and motivation

Dexterous manipulation considers multiple manipulators co-operating to achieve fine-grained operations on target objects [1]. While initially explored in a sensorless fashion [2], modern methods close the loop with rich sensing, and decompose the problem into manipulation primitives [3]. Since dexterous manipulation is object-centric, precise knowledge of object and contact state are valuable. Small errors in state information can result in incorrect palm configurations, and ultimately impact grasp stability. For unknown objects, sensory information can be fused to create global shape models.

A common, well-studied approach towards object localization is tracking with vision or depth [4]. This is unreliable—especially for interactions with small manipulands, where global sensors are frequently occluded by end-effectors. The synergy of touch and vision is natural due to their complementarity—while touch gives fine local information, vision provides noisy global constraints.

Simultaneous localization and mapping (SLAM) methods have found success across domains [5]—most recently in manipulation [6, 7, 8, 9, 10, 11]. Probabilistic methods are well-suited to fuse multi-modal sensing data in a unified manner, handling ambiguities and imperfect observations. In particular, smoothing-based frameworks can perform nonlinear least-square optimization efficiently [12], and even incorporate hard constraints [13]. Concurrently, there has been improvements in tactile sensing [14]—evolving from point contact sensors to high-resolution tactile arrays [15, 16]. **The advances in probabilistic inference, deep-learning, dexterity, and sensing open up opportunities for robust object state estimation and modeling with rich, multi-modal data.**

This proposal aims to develop both robust, efficient state estimation and mapping capabilities for object-centric dexterous manipulation. We promote these algorithms as unified probabilistic frameworks to fuse dense touch, vision, and physics. Specifically, our objectives are (Section 3):

1. **Robust estimation of object and contact state** (Section 3.1): We propose a factor graph-based incremental smoothing solution that incorporates constraints via dense touch, monocular vision, and quasi-static interactions. Dense local geometry and deep-learning based visual pose tracking is fused to additionally infer contact state. Specifically, running estimates of object pose, contact points, and forces are obtained. We further introduce multi-hypothesis tracking as a means to reason over ambiguous pose distributions.

2. **Efficient object modeling using Gaussian processes** (Section 3.2): Global shape perception overwhelmingly assumes fixed and stationary objects, and is computationally expensive. We aim to develop a general object modeling framework with Gaussian processes using dense touch and RGB-D data. Efficient approximations to kernel methods, and active exploration strategies are also of interest.

2 Background

2.1 State estimation for object manipulation

There exists a large body of prior work related to object localization, predominantly assuming known shape models. While purely touch-based localization has been demonstrated in limited domains, its combination with global pose constraints (e.g. vision) is a more practical solution. While initial work focused on point contact sensors, the emergence of high-resolution tactile arrays presents richer localization cues. Each modality can be fused with visual data for robust object tracking. Detailed accounts of tactile methods are highlighted by Luo et al. [17].

Sparse contact methods

Initial work focused on the simplified problem of predicting the pose of an object grasped by a hand through joint angle and torque information [18]. Moll et al. [19] used three point contacts and dynamics equations to recover the motion and shape of a smooth, convex object. Bayesian [20, 21] and particle filter (PF) [22, 23] methods draw analogies to the SLAM problem. These problems highlight the sufficiency of sparse touch information for pose estimation of known objects.

If vision is used together with sparse touch, robust localization can be achieved. Hebert et al. [24] fuse vision with touch for in-hand localization, but without considering frictional mechanics. Alternatively, vision can bootstrap the initial object pose, and touch can further refine it [25, 26]. Zhang and Trinkle [27] incorporate both tactile and camera measurements into a PF to track an object during grasping. They highlight the problem of particle depletion, which results from fusing sensors of largely different accuracies. Subsequent work by Koval et al. [28, 29, 30] addresses this by constructing a lower-dimensional manifold of contact states the particles must lie on.

Yu et al. [7] formulate shape and pose estimation of a planar object as a nonlinear least-squares problem. They represent contact measurements and a quasi-static motion model as constraints in a factor graph, and performs an offline optimization. Incremental, graph-based approaches with iSAM [31] are later considered, but assume known object model and incorporate vision [8, 10].They additionally show empirical evidence that contact measurement noise is normally distributed in planar pushing, justifying the Gaussian noise assumption [8]. Lambert et al. [10] also estimate contact points and the force vectors—beneficial to contact-rich planning.

High-resolution tactile methods

Rich local information from tactile-arrays enable robust pose constraints and powerful visio-tactile synergy. With greater spatial resolution, tactile sensors can localize and perceive similar to conventional imaging. Recently, there has been work that estimates object motion through a unified representation of vision/depth and dense touch [33, 34, 35, 36, 37]. These rely heavily on vision, and do not explicitly incorporate physics-based motion models.

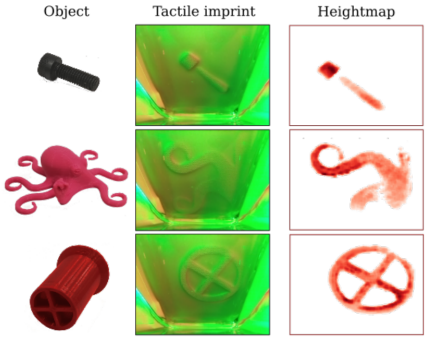


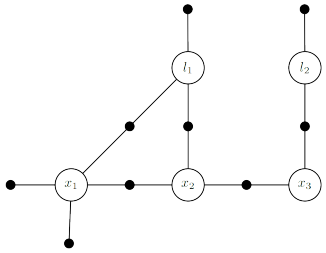
Figure 1: Tactile imprints and CNN-generated heightmaps of complex, novel objects with the GelSlim sensor [32]

GelSight [15] generates dense tactile imprints of local surfaces via visual data. The sensor is equipped with a camera viewing an opaque elastomer gel. The gel is illuminated, and the camera captures deformations when in contact with an object. This gives a heightmap that we can correlate to a local tactile shape. Initially bulky, the GelSlim [16] was later developed as a compact sensor for manipulation. While force distribution is complex and hard to predict, it can be done via inverse FEM [38] or analyzing force equilibrium [3].

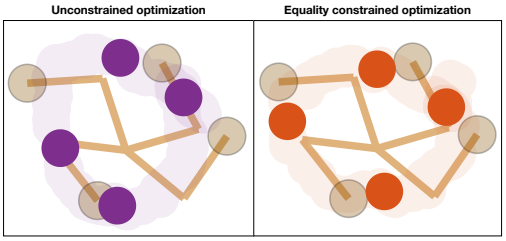
Li et al. [39] perform feature-based matching of local height maps to accurately recover the pose of small parts. Izatt et al. [40] fuse RGB-D and GelSight with iterative closest point (ICP) and a Kalman filter framework. Bauza et al. [32] uses the GelSlim to precompute a global tactile map, and localize through a combination of tactile imprints and ICP.

2.2 SLAM and factor graphs

The state-of-the-art of SLAM can be split into *filtering* and *smoothing* algorithms. As a primer, Cadena et al. [5] provides a detailed overview of the progress and challenges in SLAM. Initial probabilistic methods for state estimation used filtering—such as the extended Kalman filter (EKF) [41], and particle filter (PF) [42]. While the EKF is popular for real-time, online estimation [40, 24], it is prone to linearization errors. At each timestep, it linearizes about the potentially incorrect current estimate, leading to inaccurate tracking. The PF is computationally expensive and susceptible to particle depletion around the true state of the variable [27].



(a) A sample factor graph representing poses and observations [43]. In this example, the variable nodes are the poses (x_i) and landmarks (l_i), while measurement factors are denoted by smaller dots.



(b) Enforcing equality constraints restrict the possible pose configurations of the object [13]. Equality constrained nonlinear least squares optimization ensures non-penetration of physical systems.

Smoothing is more accurate as it preserves the temporal history of cost functions, and solves a nonlinear least-squares problem. Maintaining this history keeps the optimization robust to noisy measurements, and the linearization point is periodically updated. The nonlinear least-squares problem is commonly represent as a *factor graph*, as shown in Fig. 2a. It is a bipartite graph with *variables* in the optimization and *factors* that constrain the system. In this representative problem—taken from [43]—the state $\mathcal{X}$ comprises of robot poses x_i and observed landmarks l_j. The edges in the graph show the dependencies and constraints in the system. The measurements in the system—$\mathcal{Z}$—are a mix of binary factors (e.g. odometry, camera measurements), and unary factors (e.g. pose priors). The *maximum a posteriori* (MAP) estimate gives us variables that maximally agree with the sensor measurements. A common assumption is the Gaussian noise model, which reduces this inference to a nonlinear least-squares optimization problem. For further details and derivations, we direct the reader towards Dellaert et al. [43].

An information smoothing approach was first proposed for efficient, sparse factorization [44]. This was followed-up by work on online, incremental inference—incremental smoothing and mapping (iSAM) [31]. Rather than re-calculating the entire system at every timestep, iSAM updates previous matrix factorization with the new measurements. iSAM2 [12] further connects this to graphical model inference—for accurate, efficient, incremental nonlinear optimization. Recent efforts by Sodhi et al. [13] allow for hard constraints by leveraging primal-dual methods. This retains the structure of the optimization problem, while ensuring it does not violate physics or dynamics constraints. Fig. 2b illustrates its application to ensure hard equality on rigid-body contact interactions.

The state-of-the-art in SLAM considers unimodal state estimates, which make it prone to data association and perceptual aliasing errors. These include errors from occlusion, incorrect registration, and featureless scenes. This has a severe impact on online inference, as it will lose track of the true mode. Strategies adopted by the community include particle filtering [45], robust estimators [46], switchable constraints [47], dynamic covariance scaling [48], and max-mixtures [49]. Recently, Huang et al. [50] address this by tracking each mode through parallel Gaussian inference threads. This results in exponential growth in hypotheses and computational intractability, but Hsaio et al. [51] instead use a hypo-tree with iSAM2 for efficient multi-hypothesis inference.

2.3 Object modeling and Gaussian processes

Most prior methods from Section 2.1 track the pose of known object models. *Global shape perception* concerns the problem of building and refining an object model, with a representation that can accommodate different sensing modalities. Various parametric models have been explored, such as superquadrics [52, 53, 54], shape primitives [55, 56], polyhedrons [57], voxel maps [58, 59], point-clouds [40], and standard meshes. For example, Yu et al. [7] recovers both shape and pose in the planar case, using a piecewise-linear discrete representation.

Implicit surface models represent objects with a signed distance function (SDF)—zero on the surface, positive outside, and negative inside. Specifically, Gaussian process implicit surfaces (GPIS) are non-parametric representations that fuse uncertain measurements probabilistically [60]. They have

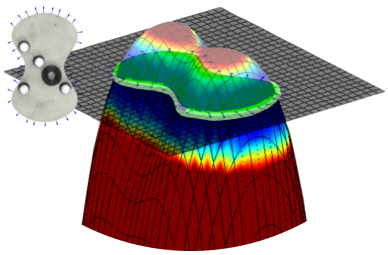


Figure 3: Gaussian process and its implicit surface (green) for noisy contact measurements on the `butter` shape [8]. Measurements are added sequentially and colormap shows uncertainty (video).

gained popularity over parameteric methods as they **(i)** faithfully approximate arbitrary geometries [61], **(ii)** fuse noisy measurements from multiple sensors [33], **(iii)** provide surface estimate uncertainty to close the loop [62, 63]. The SDF is also directly useful for a grasp controller [61].

A GP regressor learns a continuous, nonlinear function from sparse, noisy datapoints [64]. While contact points are zero-value observations, contact normals are gradient observations. For Fig. 3, we learn a mapping from contact points $X_{1\cdots t}$ to signed-distance $d_{1\cdots t}$ and normals $N_{1\cdots t}$:

$$\mathcal{F} : \{X_i\}^{i=1\cdots t} \mapsto \{d_i, N_i\}^{i=1\cdots t} \tag{1}$$

The covariance or kernel function can be chosen, and describes the relatedness of measurement quantities [64]. The GP gives an output mean function as its MAP estimate, and a variance. The implicit surface $\mathcal{S}$ is the zero-level set contour of $\mathcal{F}$, such that:

$$\mathcal{S} \triangleq \{X \mid \mathcal{F}_d(X) = 0\} \tag{2}$$

One disadvantage of GPs is the computational complexity—$O(N^3)$—where N is the size of training data. This is a bottleneck especially when dealing with high-resolution data, like RGB-D and GelSight [40]. Efficient solutions have been developed which admit limited samples [65], split the problem into manageable subsets [66, 67, 68], or use sparse covariance functions [69].

3 Project objectives

3.1 Robust estimation of object and contact state

Accurate state information can greatly benefit planning and control for dexterous manipulation [3], as well as intricate tasks like insertion and packing [9, 70]. This includes 6-DoF object pose, as well as running estimates of auxilary variables—contact points, and forces. These parameters must be inferred and refined from a combination of noisy sensors—coarse global visual measurements, and fine local tactile data.

Smoothing over vision, touch, and physics

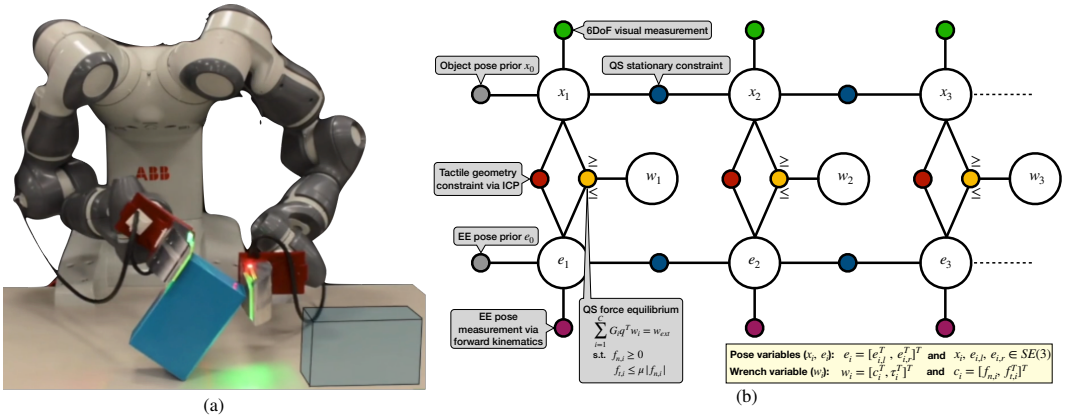


Figure 4: **(a)** Manipulating a known object on a table top to a target pose [3]. This dexterous platform has a pair of high-resolution tactile sensors. Accurate knowledge of object and contact state is vital for successful interactions. **(b)** Our proposed graphical model formulation, where circles are optimization variables and colored dots are the constraining factors. A combination of learned visual tracking, dense tactile registration, and force equilibrium constraints can robustly recover our augmented state (x_i, e_i, w_i)

We propose a unified probabilistic approach towards online object and contact state estimation for dexterous manipulation tasks (Fig. 4a). Recent efforts combine touch and RGB-D in an EKF framework [40], or solely through tactile information [32]. While neither reasons about physics, Hogan et al. [3] use tactile measurements and quasi-static force equilibrium [71] to predict object and contact state. We propose a factor graph-based incremental smoothing solution that incorporates constraints via dense touch, monocular vision, and quasi-static interactions. The approach is summarized as a graphical model in Fig. 4b, and briefly explained below.

We propose using RGB data—with recent advances in deep learning pose regression [72, 73, 74]—for global constraints on the object. At each timestep we obtain dense local geometry from the sensor pair, and we register these measurements with respect to the object model for an additional constraint. This can be achieved via conventional frame-to-model ICP [32], or robust point-to-implicit methods [75, 76, 40]. Assuming quasi-static interactions, we can use force equilibrium constraints (Fig. 4b) alongside Coulomb friction inequality constraints. The graph can be optimized incrementally using GTSAM [77] with iSAM2 [12]. **As output, this module maintains a running online estimate of object pose (x_t), refined end-effector poses (e_t), contact locations, and contact wrenches (w_t).**

Robust multi-hypothesis tracking

One disadvantage of using local tactile geometry is that the small sensor coverage may not completely disambiguate object pose. Izatt et al. [40] points to both geometric similarity and visual occlusion as reasons for the tracker falling into local minima. This is a pitfall in our scenario too, given frequent camera occlusion and occasional uninformative tactile data. As we illustrate in Fig. 5, dense touch registration can fail to disambiguate the true pose of a regular object. This bears resemblance to the contact manifold highlighted by Koval et al. [29], except with richer tactile information.

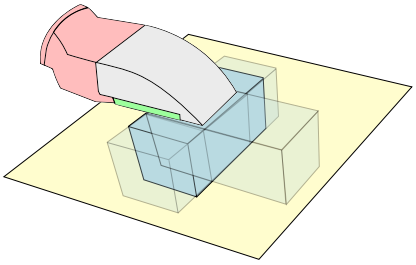
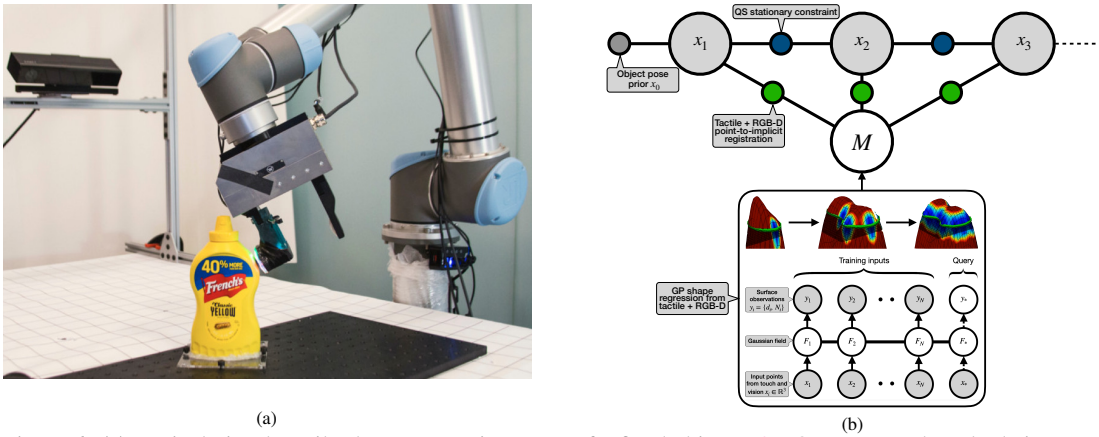


Figure 5: Possible pose hypothesis that represent the "manifold" that satisfies the contact data. With multi-hypothesis tracking [51], we can efficiently branch our inference problem and prune away incorrect candidates as they fail to agree with future measurements.

While particle filters address this, they are computationally expensive and prone to particle depletion (Section 2.1). We wish to simultaneously reason over possible object pose distributions such that **(i)** it is computationally tractable, **(ii)** they can be disambiguated with future measurements. Here, we aim to build on recent work in multi-hypothesis smoothing [51], except in the context of object pose estimation.

3.2 Efficient object modeling using Gaussian processes



(a) (b)

Figure 6: **(a)** Typical visual-tactile shape perception setup of a fixed object [73]. Our proposed method aims to build accurate models despite motion induced by tactile exploration. **(b)** The factor graph comprises of poses x_i optimized incrementally, and an implicit surface M constructed in tandem by the GP. The model is first bootstrapped with an RGB-D prior, and at every timestep a hybrid point-cloud from tactile and depth measurements is formed. The object pose is then correct by point-to-implicit alignment of the cloud, and the GP is finally updated.

This work develops shape perception algorithms for apriori unknown objects on a tabletop, with a dexterous manipulator. Through visual-tactile shape exploration we can build dense models for manipulation tasks, or perform object identification. Importantly, we wish to move away from the

contrived *fixed pose* setup of shape perception [32, 73, 78, 79], and accommodate object motion. To our knowledge, no methods use 3D implicit surfaces with online pose estimation for mapping.

As previously demonstrated by Suresh et al. [80], we are interested in combining a GPIS representation with a graphical model for robust mapping in-the-wild. Unknown model information precludes us from incorporating interaction physics, but we can fuse RGB-D and dense tactile measurements for frame-to-model pose corrections. Active tactile exploration is driven by GP uncertainty, with policies to ensure measurement overlap and minimal RGB-D occlusion. High-resolution sensing necessitates efficient GP regression through hierarchical methods or sparse approximations (Section 2.3).

References

[1] Allison M Okamura, Niels Smaby, and Mark R Cutkosky. An overview of dexterous manipulation. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, volume 1, pages 255–262. IEEE, 2000.

[2] Michael Erdmann. An exploration of nonprehensile two-palm manipulation. *Intl. J. of Robotics Research (IJRR)*, 17(5):485–503, 1998.

[3] Francois R Hogan, Jose Ballester, Siyuan Dong, and Alberto Rodriguez. Tactile dexterity: Manipulation primitives with tactile feedback. *arXiv preprint arXiv:2002.03236*, 2020.

[4] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1386–1383. IEEE, 2017.

[5] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Trans. on Robotics (TRO)*, 32(6):1309–1332, 2016.

[6] Matthew Klingensmith, Siddartha S Sirinivasa, and Michael Kaess. Articulated robot motion for simultaneous localization and mapping (ARM-SLAM). *IEEE Robotics and Automation Letters (RA-L)*, 1(2):1156–1163, 2016.

[7] Kuan-Ting Yu, John Leonard, and Alberto Rodriguez. Shape and pose recovery from planar pushing. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1208–1215. IEEE, 2015.

[8] Kuan-Ting Yu and Alberto Rodriguez. Realtime state estimation with tactile and visual sensing. application to planar manipulation. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 7778–7785. IEEE, 2018.

[9] Kuan-Ting Yu and Alberto Rodriguez. Realtime state estimation with tactile and visual sensing for inserting a suction-held object. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1628–1635. IEEE, 2018.

[10] Alexander Sasha Lambert, Mustafa Mukadam, Balakumar Sundaralingam, Nathan Ratliff, Byron Boots, and Dieter Fox. Joint inference of kinematic and force trajectories with visuo-tactile sensing. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 3165–3171. IEEE, 2019.

[11] Mandy Xie and Frank Dellaert. A unified method for solving inverse, forward, and hybrid manipulator dynamics using factor graphs. *arXiv preprint arXiv:1911.10065*, 2019.

[12] Michael Kaess, Hordur Johannsson, Richard Roberts, Viorela Ila, John Leonard, and Frank Dellaert. iSAM2: Incremental smoothing and mapping using the Bayes tree. *Intl. J. of Robotics Research (IJRR)*, 31(2):216–235, 2012.

[13] P. Sodhi, S. Choudhury, J.G. Mangelson, and M. Kaess. ICS: Incremental constrained smoothing for state estimation. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Paris, France, May 2020.

[14] Rui Li and Edward H Adelson. Sensing and recognizing surface textures using a GelSight sensor. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1241–1247, 2013.

[15] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson. Shape-independent hardness estimation using deep learning and a GelSight tactile sensor. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 951–958, 2017.

[16] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez. GelSlim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1927–1934, 2018.

[17] Shan Luo, Joao Bimbo, Ravinder Dahiya, and Hongbin Liu. Robotic tactile perception of object properties: A review. *Mechatronics*, 48:54–67, 2017.

[18] David M Siegel. Finding the pose of an object in a hand. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 406–411. IEEE, 1991.

[19] Mark Moll and Michael A Erdmann. Reconstructing the shape and motion of unknown objects with active tactile sensors. In *Algorithmic Foundations of Robotics V*, pages 293–309. Springer, 2004.

[20] Klaas Gadeyne, Tine Lefebvre, and Herman Bruyninckx. Bayesian hybrid model-state estimation applied to simultaneous contact formation recognition and geometrical parameter estimation. *Intl. J. of Robotics Research (IJRR)*, 24(8):615–630, 2005.

[21] Anna Petrovskaya, Oussama Khatib, Sebastian Thrun, and Andrew Y Ng. Bayesian estimation for autonomous object manipulation based on tactile sensors. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 707–714. IEEE, 2006.

[22] Craig Corcoran and Robert Platt. A measurement model for tracking hand-object state during dexterous manipulation. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 4302–4308. IEEE, 2010.

[23] Anna Petrovskaya and Oussama Khatib. Global localization of objects via touch. *IEEE Trans. on Robotics (TRO)*, 27(3):569–585, 2011.

[24] Paul Hebert, Nicolas Hudson, Jeremy Ma, and Joel Burdick. Fusion of stereo vision, force-torque, and joint sensors for estimation of in-hand object location. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 5935–5941. IEEE, 2011.

[25] Kyuhei Honda, Tsutomu Hasegawa, Toshihiro Kiriki, and Takeshi Matsuoka. Real-time pose estimation of an object manipulated by multi-fingered hand using 3D stereo vision and tactile sensing. In *Proceedings. 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems. Innovations in Theory, Practice and Applications (Cat. No. 98CH36190)*, volume 3, pages 1814–1819. IEEE, 1998.

[26] Joao Bimbo, Lakmal D Seneviratne, Kaspar Althoefer, and Hongbin Liu. Combining touch and vision for the estimation of an object's pose during manipulation. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 4021–4026. IEEE, 2013.

[27] Li Zhang and Jeffrey C Trinkle. The application of particle filtering to grasping acquisition with visual occlusion and tactile sensing. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 3805–3812. IEEE, 2012.

[28] Michael C Koval, Mehmet R Dogar, Nancy S Pollard, and Siddhartha S Srinivasa. Pose estimation for contact manipulation with manifold particle filters. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 4541–4548. IEEE, 2013.

[29] Michael C Koval, Nancy S Pollard, and Siddhartha S Srinivasa. Pose estimation for planar contact manipulation with manifold particle filters. *Intl. J. of Robotics Research (IJRR)*, 34(7):922–945, 2015.

[30] Michael C Koval, Nancy S Pollard, and Siddhartha S Srinivasa. Manifold representations for state estimation in contact manipulation. In *Robotics Research*, pages 375–391. Springer, 2016.

[31] Michael Kaess, Ananth Ranganathan, and Frank Dellaert. iSAM: Incremental smoothing and mapping. *IEEE Trans. on Robotics (TRO)*, 24(6):1365–1378, 2008.

[32] Maria Bauza, Oleguer Canal, and Alberto Rodriguez. Tactile mapping and localization from high-resolution tactile imprints. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 3811–3817. IEEE, 2019.

[33] Marten Björkman, Yasemin Bekiroglu, Virgile Högman, and Danica Kragic. Enhancing visual perception of shape through tactile glances. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 3180–3186. IEEE, 2013.

[34] Jarmo Ilonen, Jeannette Bohg, and Ville Kyrki. Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing. *Intl. J. of Robotics Research (IJRR)*, 33(2):321–341, 2014.

[35] Shan Luo, Wenxuan Mou, Kaspar Althoefer, and Hongbin Liu. Localizing the object contact through matching tactile features with visual map. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 3903–3908. IEEE, 2015.

[36] Tanner Schmidt, Katharina Hertkorn, Richard Newcombe, Zoltan Marton, Michael Suppa, and Dieter Fox. Depth-based tracking with physical constraints for robot manipulation. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 119–126. IEEE, 2015.

[37] Pietro Falco, Shuang Lu, Andrea Cirillo, Ciro Natale, Salvatore Pirozzi, and Dongheui Lee. Cross-modal visuo-tactile object recognition using robotic active exploration. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 5273–5280. IEEE, 2017.

[38] Daolin Ma, Elliott Donlon, Siyuan Dong, and Alberto Rodriguez. Dense tactile force estimation using GelSlim and inverse FEM. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 5418–5424. IEEE, 2019.

[39] Rui Li, Robert Platt, Wenzhen Yuan, Andreas ten Pas, Nathan Roscup, Mandayam A Srinivasan, and Edward Adelson. Localization and manipulation of small parts using GelSight tactile sensing. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3988–3993. IEEE, 2014.

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

[41] Randall Smith, Matthew Self, and Peter Cheeseman. Estimating uncertain spatial relationships in robotics. In *Autonomous robot vehicles*, pages 167–193. Springer, 1990.

[42] Sebastian Thrun. Particle filters in robotics. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 511–518. Morgan Kaufmann Publishers Inc., 2002.

[43] Frank Dellaert and Michael Kaess. Factor graphs for robot perception. *Foundations and Trends in Robotics*, 6(1-2):1–139, 2017.

[44] Frank Dellaert and Michael Kaess. Square root SAM: Simultaneous localization and mapping via square root information smoothing. *Intl. J. of Robotics Research (IJRR)*, 25(12):1181–1203, 2006.

[45] Michael Montemerlo, Sebastian Thrun, Daphne Koller, Ben Wegbreit, et al. FastSLAM: A factored solution to the simultaneous localization and mapping problem. *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*, 593598, 2002.

[46] David M Rosen, Michael Kaess, and John J Leonard. An incremental trust-region method for robust online sparse least-squares estimation. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1262–1269. IEEE, 2012.

[47] Niko Sünderhauf and Peter Protzel. Switchable constraints vs. max-mixture models vs. RRR-a comparison of three approaches to robust pose graph SLAM. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 5198–5203. IEEE, 2013.

[48] Pratik Agarwal, Gian Diego Tipaldi, Luciano Spinello, Cyrill Stachniss, and Wolfram Burgard. Robust map optimization using dynamic covariance scaling. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 62–69. Ieee, 2013.

[49] Edwin Olson and Pratik Agarwal. Inference on networks of mixtures for robust robot mapping. *Intl. J. of Robotics Research (IJRR)*, 32(7):826–840, 2013.

[50] Guoquan Huang, Michael Kaess, and John J Leonard. Consistent unscented incremental smoothing for multi-robot cooperative target tracking. *J. of Robotics and Autonomous Systems (RAS)*, 69:52–67, 2015.

[51] Ming Hsiao and Michael Kaess. MH-iSAM2: Multi-hypothesis iSAM using Bayes tree and hypo-tree. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1274–1280. IEEE, 2019.

[52] Franc Solina and Ruzena Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Trans. Pattern Anal. Machine Intell.*, 12(2):131–147, 1990.

[53] Ales Leonardis, Ales Jaklic, and Franc Solina. Superquadrics for segmenting and modeling range data. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(11):1289–1295, 1997.

[54] Hongbin Zha, Tsuyoshi Hoshide, and Tsutomu Hasegawa. A recursive fitting-and-splitting algorithm for 3-D object modeling using superquadrics. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, volume 1, pages 658–662. IEEE, 1998.

[55] Zoltan-Csaba Marton, Lucian Goron, Radu Bogdan Rusu, and Michel Beetz. Reconstruction and verification of 3D object models for grasping. In *Robotics Research*, pages 315–328. Springer, 2011.

[56] Kai Huebner, Steffen Ruthotto, and Danica Kragic. Minimum volume bounding box decomposition for shape approximation in robot grasping. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1628–1633. IEEE, 2008.

[57] Stefano Caselli, Corrado Magnanini, Francesco Zanichelli, and Enrico Caraffi. Efficient exploration and recognition of convex objects based on haptic perception. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 3508–3513. IEEE, 1996.

[58] Donald Meagher. Geometric modeling using octree encoding. *Comput. Graph. Image Process.*, 19(2):129–147, 1982.

[59] Kester Duncan, Sudeep Sarkar, Redwan Alqasemi, and Rajiv Dubey. Multi-scale superquadric fitting for efficient shape and pose recovery of unknown objects. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 4238–4243. IEEE, 2013.

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

[61] Stanimir Dragiev, Marc Toussaint, and Michael Gienger. Gaussian process implicit surfaces for shape estimation and grasping. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 2845–2850. IEEE, 2011.

[62] Stanimir Dragiev, Marc Toussaint, and Michael Gienger. Uncertainty aware grasping and tactile exploration. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 113–119. IEEE, 2013.

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

[64] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[65] Nicolas Sommer, Miao Li, and Aude Billard. Bimanual compliant tactile exploration for grasping unknown objects. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 6400–6407. IEEE, 2014.

[66] Shrihari Vasudevan, Fabio Ramos, Eric Nettleton, and Hugh Durrant-Whyte. Gaussian process modeling of large-scale terrain. *Journal of Field Robotics*, 26(10):812–840, 2009.

[67] Simon T O'Callaghan and Fabio T Ramos. Gaussian process occupancy maps. *Intl. J. of Robotics Research (IJRR)*, 31(1):42–62, 2012.

[68] Yirong Shen, Matthias Seeger, and Andrew Y Ng. Fast Gaussian process regression using kd-trees. In *Advances in neural information processing systems*, pages 1225–1232, 2006.

[69] Soohwan Kim and Jonghyuk Kim. GPmap: A unified framework for robotic mapping based on sparse Gaussian processes. In *Proc. Conf. on Field and Service Robotics (FSR)*, pages 319–332. Springer, 2015.

[70] Siyuan Dong and Alberto Rodriguez. Tactile-based insertion for dense box-packing. *arXiv preprint arXiv:1909.05426*, 2019.

[71] Bruno Siciliano and Oussama Khatib. *Springer handbook of robotics*. Springer, 2016.

[72] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[73] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson. 3D shape perception from monocular vision, touch, and shape priors. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1606–1613, 2018.

[74] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6D object pose estimation under hybrid representations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 431–440, 2020.

[75] Erik Bylow, Jürgen Sturm, Christian Kerl, Fredrik Kahl, and Daniel Cremers. Real-time camera tracking and 3D reconstruction using signed distance functions. In *Proc. Robotics: Science and Systems (RSS)*, volume 2, page 2, 2013.

[76] Tanner Schmidt, Richard A Newcombe, and Dieter Fox. DART: Dense articulated real-time tracking. In *Proc. Robotics: Science and Systems (RSS)*, volume 2. Berkeley, CA, 2014.

[77] Frank Dellaert. Factor graphs and GTSAM: A hands-on introduction. Technical report, Georgia Institute of Technology, 2012.

[78] Zhengkun Yi, Roberto Calandra, Filipe Veiga, Herke van Hoof, Tucker Hermans, Yilei Zhang, and Jan Peters. Active tactile object exploration with Gaussian processes. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 4925–4930. IEEE, 2016.

[79] Danny Driess, Daniel Hennes, and Marc Toussaint. Active multi-contact continuous tactile exploration with Gaussian process differential entropy. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 7844–7850. IEEE, 2019.

[80] Sudharshan Suresh, Joshua G Mangelson, and Michael Kaess. Incremental shape and pose estimation from planar pushing using contact implicit surfaces. In *ICRA 2020 workshop - ViTac 2020: Closing the Perception-Action Loop with Vision and Tactile Sensing*, 2020.