A Belief representation for Active Uncertain Pose Estimation via Simulation-based Inference

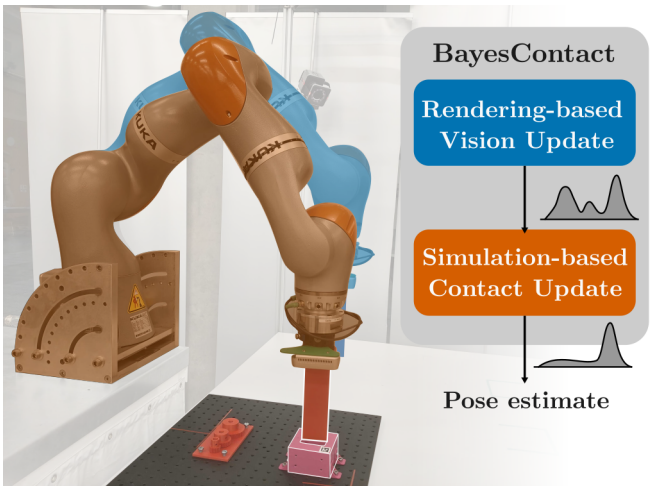

Fig. 1: A Simulation-based Inference framework to approximate posterior beliefs using Depth and Force–Torque data for **Forceful** and **Contact-rich** Manipulation.

***Abstract*—Contact-rich manipulation tasks require more accurate scene understanding than is normally possible solely using visual sensing. For peg-insertion, the pose of the hole must be estimated; however, camera noise can easily exceed insertion part tolerance. By leveraging force and torque measurements during part interaction, contact sensing can provide valuable information that refines estimates based on noisy visual sensing and enable contact-rich peg insertion. Here we propose a Simulation-based Inference framework to fuse visual information and contact information via probabilistic program proposals utilising rendering and physics as generative functions.**

***Index Terms*—Sampling based Inference, Simulation based Inference, Perception for Dexterous Manipulation**

I. Introduction

In this paper, we revisit the contact-rich peg-in-hole insertion problem, by focusing on relative pose estimation between robot and the *hole* object. We present *BayesContact*, an online visuo-tactile pose estimation based on Simulation-based Inference [1]. Our method retains the Bayesian Filtering structure by using a physics and rendering engine within sensor models. BayesContact estimates the pose of a known target object for an insertion task, from RGB-D and force/torque (F/T) feeds without prior data or learning. By doing this, physics and graphics can seemlessly be integrated into the generative modelling, obtaining more expressive likelihoods for robotics tasks. Furthermore, we posit that using non-parametric representations allows for tractable Approximate Information Gain computation for the purposes of applications involving Active Sensing.

II. Method

A. Problem Formulation

We consider a partially observable, contact-rich inference problem in which the objective is to estimate the unknown configuration of a rigid *hole* through a sequence of perception and interaction measurements. Let $q \in \mathcal{Q}$ denote the latent configuration of the hole, typically an element of $\mathrm{SE}(3)$. Measurements are indexed by k, with $a_k \in \mathcal{A}$ denoting the probing or control action executed for the k-th measurement, and $o_k \in \mathcal{O}$ the resulting observation. The configuration q is assumed to be static, while observations are acquired sequentially. Each observation is multimodal, consisting of a depth image o_k^d and a force–torque measurement sequence o_k^f, such that $o_k = (o_k^d, o_k^f)$. The inference task is to estimate the posterior belief over configurations conditioned on the history of actions and observations, $p(q_k \mid o_{1:k}, a_{1:k})$.

B. Approximate Bayesian Computation

Given a sequence of measurements, inference over the static hole configuration q admits a recursive Bayesian update. Let $p(q \mid o_{1:k-1}, a_{1:k-1})$ denote the posterior belief after $k-1$ measurements. Upon executing action a_k and observing o_k, the posterior is updated as

$$p(q \mid o_{1:k}, a_{1:k}) \propto p(o_k \mid q, a_k)\, p(q \mid o_{1:k-1}, a_{1:k-1}),$$

where $p(o_k \mid q, a_k)$ denotes the observation likelihood induced by the k-th measurement. As this posterior is generally intractable to represent analytically, we approximate the belief using Sequential Monte Carlo (SMC). At measurement k, the belief is represented by a weighted particle set $b_k = \{(q_k^{(i)}, w_k^{(i)})\}_{i=1}^N$, yielding the empirical approximation $p(q \mid o_{1:k}, a_{1:k}) \approx \sum_i w_k^{(i)} \delta(q - q_k^{(i)})$. Particles are proposed according to a distribution $\pi(q \mid q_{k-1}^{(i)})$, and their weights are updated via importance sampling,

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(o_k \mid q_k^{(i)}, a_k)\, p(q_k^{(i)})}{\pi(q_k^{(i)} \mid q_{k-1}^{(i)})},$$

The sequential Monte Carlo formulation described above requires evaluating, or approximating, the observation likelihood $p(o_k \mid q_k, a_k)$ in order to update particle weights. In contact-rich settings involving complex geometry, contact dynamics, and sensing processes, this likelihood is not available in closed form. We therefore define the observation model implicitly through a physics- and graphics-based simulator, which serves as a forward generative process for observations.

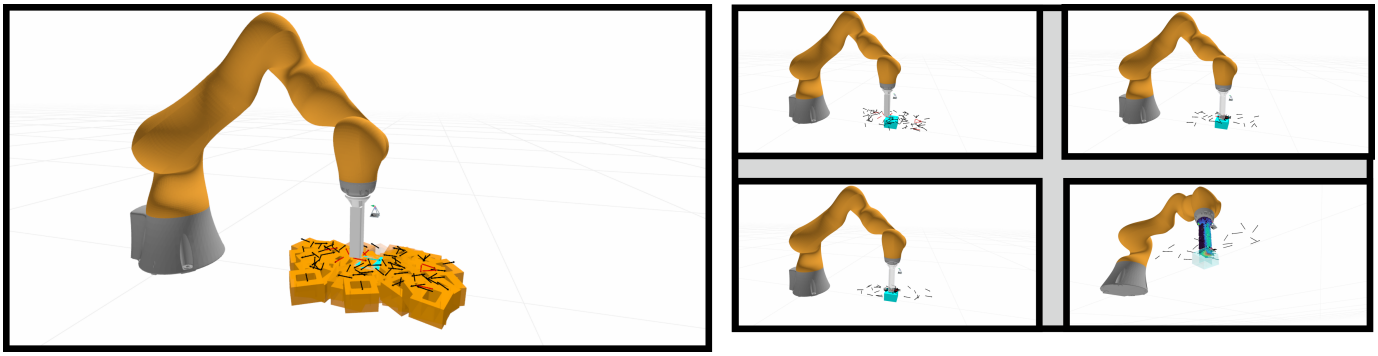

Fig. 2: Qualitative results of belief evolution. Pose of the hole is represented as weighted particles in SE(3) subspaces. A rendering engine is used to specify image likelihoods, and physics engine is used to specify contact likelihoods. Simulation-based Inference via Sequential Monte Carlo [2] updates using vision and contact observations are used infer pose of the hole. Image on the left shows particle initialisation, with a black arrow representing the orientation. An example evolution is shown within the gray box on the right. From top-left to bottom-left in the **clockwise** direction: particles converging towards the ground truth object (**cyan**). Red arrows represent the top 10 weighted particles. **Bottom-right** shows how contact information gets filtered.

C. Simulation based Generative Models.

The observation model is defined implicitly through a physics- and graphics-based simulator and follows a modality-indexed measurement structure. Measurements are indexed by $k = 1, \ldots, K$, with the first γ measurements corresponding to visual perception and the remaining measurements corresponding to physical interaction. Each measurement produces a single-modality observation

$$o_k = \begin{cases} o_k^d, & k \leq \gamma, \\ o_k^f, & k > \gamma, \end{cases}$$

where o_k^d denotes a depth observation and o_k^f a force–torque measurement.

Since the observation likelihood $p(o_k \mid q, a_k)$ is not available in closed form, we perform inference using Approximate Bayesian Computation with modality-specific surrogate likelihoods.

Depth Image Likelihood: For visual measurements $k \leq \gamma$, the observation o_k^d is a depth image and $\tilde{o}_k^{d,(i)}$ denotes the rendered depth image obtained by forward rendering the scene under particle configuration $q_k^{(i)}$. Let u index image pixels We define a Log-Laplace per-pixel likelihood given by

$$\log p(o_k^d \mid q_k^{(i)}) = -\frac{1}{b_d} \left| o_k^d(u) - \tilde{o}_k^{d,(i)}(u) \right| - \log(2b_d),$$

where b_d is the scale parameter controlling depth noise.

Contact Likelihood: For physical interaction measurements $k > \gamma$, the observation o_k^f is processed into a set of contact locations $P_{\text{obs}} \subset \mathbb{R}^3$ expressed in the peg reference frame [3]. Given a particle configuration $q_k^{(i)}$ and action a_k, the simulator produces a corresponding set of simulated contact locations $P_{\text{sim}}^{(i)} = \tilde{o}_k^{f,(i)}$ by simulating contact dynamics under the same action.

To compare observed and simulated contact geometry, we employ a bidirectional Chamfer distance $D_{\text{ch}}(\cdot, \cdot)$ between point sets. The resulting surrogate log-likelihood is defined as

$$\log p(o_k^f \mid q_k^{(i)}, a_k) = -\frac{1}{2\sigma_c^2} D_{\text{ch}}\left(P_{\text{obs}}, P_{\text{sim}}^{(i)}\right),$$

where σ_c is a scale parameter controlling tolerance to contact localization noise.

This likelihood assigns higher probability to particles whose predicted contact point

Algorithm 1 Myopic next-best probe

Require: belief $b_k = \{(q_i, w_i)\}_{i=1}^N$, candidate probes $\mathcal{A}$

1: **for** each $a \in \mathcal{A}$ **do**
2: simulate observations under b_k and a
3: estimate expected posterior entropy $\bar{H}(a)$
4: **end for**
5: select $a^\star = \arg\min_{a \in \mathcal{A}} \bar{H}(a)$
6: **return** $a^\star$

III. Results

To evaluate *BayesContact*, the robot is tasked to make a pre-computed probe and infer the pose of the hole. The contact event is timed at $t = 3$s. Fig. 3 reports the Average Distance of Model Points (ADD) [4] of $N = 75$ particles across time of the trajectory. An ADD=0 represents complete ground truth pose recovery. Our findings show that the proposed Simulation-based Inference framework recovers an ADD of 0.8 mm after the robot carries out a probing trajectory for $t = 6$s. Fig. 2 Shows the evolution of the particles across the trajectory of the probe. The particles enable a sample based representation of the posterior belief over the pose of the hole. Algorithm 1 proposes a one step look ahead approach utilising the particle based representation for active sensing across multiple hypothesis.

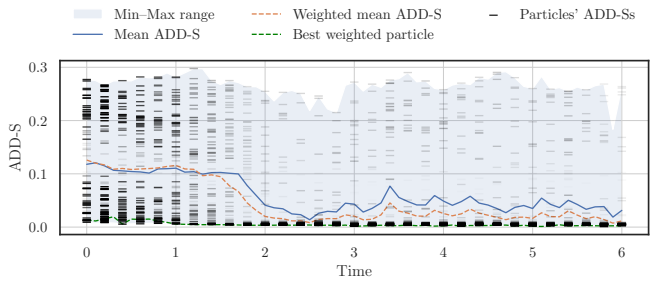

Fig. 3: Average ADD (m) after $t = 6$s of convergence. Gray dashes represent particles, with the intensity corresponding to particle weights.

REFERENCES

[1] K. Cranmer, J. Brehmer, and G. Louppe, "The frontier of simulation-based inference," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 055–30 062, Dec. 2020. DOI: 10.1073/pnas.1912789117.

[2] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later,"

[3] A. Kamireddypalli, J. Moura, R. Buchanan, S. Vijayakumar, and S. Ramamoorthy, *ContactFusion: Stochastic poisson surface maps from visual and contact sensing*, Mar. 20, 2025. DOI: 10.48550/arXiv.2503.16592. arXiv: 2503.16592[cs].

[4] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6d pose estimation and tracking of novel objects," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 17 868–17 879. DOI: 10.1109/CVPR52733.2024.01692.