IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Probabilistic Closed-Loop Active Grasping

Henry Schaub[1], Christian Wolff[2], Maximilian Hoh[1] and Alfred Schöttl[1]

Abstract— **Picking a specific object is an essential task of assistive robotics. While the majority of grasp detection approaches focus on grasp synthesis from a single depth image or point cloud, this approach is often not viable in an unstructured, uncontrolled environment. Due to occlusion, heavy influence of noise or simply because no collision-free grasp is visible from some perspectives, it is beneficial to collect additional information from other views before opting for grasp execution. We present a closed-loop approach that selects and navigates towards the next-best-view by minimizing the entropy of the volume under consideration. We use a local measure of estimation uncertainty of the surface reconstruction, to sample grasps and estimate their success probabilities in an online fashion. Our experiments show that our algorithm achieves better grasp success rates than comparable approaches, when presented with challenging household objects.**

I. Introduction

Picking objects using a representation of the scene is an essential skill for robot manipulators. While enormous progress has been made for industrially motivated setups, e.g. top-down setups or single view approaches, assistive robotics has received comparatively little attention [10]. Compared to the traditional industrial setup, the assistive field introduces a variety of challenges. The environment is unstructured and often cluttered, objects may be transparent or reflective and lighting conditions may be less than perfect. Under these non-ideal conditions sensor noise can be quite serious, especially for consumer-grade sensors (e.g. MS Kinect, Intel Realsense), and is sometimes significantly higher than specified by the manufacturer.

Others have already shown that the fusion of sensor data has a benign effect on the success rate and reduces the influence of sensor noise e.g. [3], [4]. However, the precision of depth sensors varies with the measured depth and the angle of incidence [1], surface properties like color, texture or material [12] and illumination [7]. This noise behavior is difficult to reproduce in a simulated environment.Closing the optical domain gap in robotic simulators remains a challenging problem, especially in the context of assistive robotics where the environmental conditions vary greatly and cannot be controlled beforehand. Instead the uncertainty of surface estimations need to be evaluated locally and during runtime in order to avoid infeasible actions.

The household context also requires that multiple perspectives are considered. Due to partial occlusion, a cluttered environment or simply a poorly chosen initial perspective the first view might not even show a single collision-free grasp. An intelligent grasping algorithm must therefore not only be able to select the candidate with the highest probability of success, but also dynamically generate a trajectory in order to maximize the available information.

Our approach combines next-best-view (NBV) path planning and probabilistic evaluation of grasping possibilities in a closed-loop manner. New measurements are continuously inserted into a probabilistic voxel-grid, grasp candidates are distributed over the entire surface reconstruction, their probability of success is approximated, and based on entropy considerations, the reachable NBV is selected and steered towards. Our approach is real-time capable. We assume that a rough bounding box of the target object is given and investigate the corresponding volume until a reliable, collision free grasp possibility is found. Compared to other state-of-the-art algorithms our approach handles unknown objects and environments and does not need prior knowledge either through available CAD-models or implicitly through datasets tailored to the use case. We validate our approach through extensive tests in a simulated setup and demonstrate that the presented method can also significantly improve the performance of grasp evaluators other than the proposed probabilistic one. We show further in real-world experiments with challenging household objects that our approach is able to handle difficult settings where state-of-the-art algorithms might fail. The contributions of this work are

- a probabilistic depth fusion method that allows to determine grasp success probabilities based on local estimation variances,
- an entropy based closed-loop grasping pipeline that improves the reliability of surface estimates,
- and a computationally efficient grasp sampling and evaluation method.

II. Related work

Many previous approaches focus on the estimation of grasp poses in a planar setup [16], [15]. However, these approaches inevitably restrict the solution space for grasps and are therefore prone to fail in a non top-down setting.

In order to provide more flexibility and address more complex scenarios, recent works favor the use of depth maps or point clouds to directly predict 6-dof grasps [22],

Manuscript received: July, 17, 2023; Revised January, 31, 2024; Accepted February, 13, 2024.

This paper was recommended for publication by Editor Pascal Vasseur upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the BMBF, Ref.: 13FH755IX6

[1]The authors are with the Institute for Applications of Machine Learning and Intelligent Systems, University of Applied Sciences, 80335 Munich, Germany {henry.schaub, alfred.schoettl, maximilian.hoh}@hm.edu

[2]The author is with the University of Regensburg, Faculty for Informatics and Data Science, 93053, Germany christian.wolff@computer.org

Digital Object Identifier (DOI): see top of this page.

[17]. Although these approaches perform well under ideal conditions, the household context poses a challenge to them. The first perspective the target object is seen from might not be benign. Their performance could be severely impacted by the lack of information and distinct sensor noise may strongly influence predictions made on the basis of a single measurement. In addition to the depth images, Gou et al. [9] consider the less noisy color images to predict grasps, while Ma et al. [14] propose to partially augment the synthetic training data with real noisy measurements to reduce the influence of sensor noise. Both approaches rely on a benign perspective and may fail under high levels of occlusion. Another line of work tackles these problems using an active perception approach [11][5]. Common among both approaches is that they rely on grasp hypotheses to determine the motion goal. In contrast, Song et al. [21] propose a reinforcement learning approach that utilizes visual data from manual grasping demonstrations in order to decide between different exploration actions or opt for grasp execution. However, adaption to novel objects or environments requires fine tuning. Perhaps the work most closely related is the work of Breyer et al. [2]. They fuse depth data into a truncated signed distance function (TSDF) and use a raycasting algorithm to evaluate the information gain of different view candidates. Their exploration phase continues until a grasp candidate is considered stable. The employed volumetric grasping network (VGN) [3] is trained using simulated trial-and-error experiments. It is notoriously difficult to incorporate real sensor behavior into a simulation and thus the results of real-world experiments are often less satisfactory. In an effort to close this optical sensing domain gap [23] proposes a pipeline which simulates real noise behavior. They precisely align simulated and real world and estimate setup-specific material and lighting parameters. Their grasping experiments show clear improvements if the policy is trained using the realistically simulated data. However, this approach is unsuitable for the field of service robotics where the environment is seldom precisely known beforehand.

III. Approach

We employ a classical eye-in-hand setup, where a depth sensor is mounted on the end effector of a manipulator arm. The transformation between the end effector and the sensor's optical center, as well as its intrinsic parameters are known beforehand. We consider the problem of finding a parallel-yaw grasp on a target object and assume that a corresponding 3D bounding box is provided by a preceding algorithm. This bounding box could either be supplied by a secondary, statically mounted sensor or can be inferred from the initial view (full or partial visibility).

Since initial and partial knowledge about the object is often insufficient to find reliable, collision-free grasp candidates, we propose a pipeline that actively explores the volume until a reliable grasp configuration is found. An overview of the framework is seen in figure (1). At every time-step t_i we integrate the depth image d_i, taken from the current sensor pose T_i, into a volumetric representation of

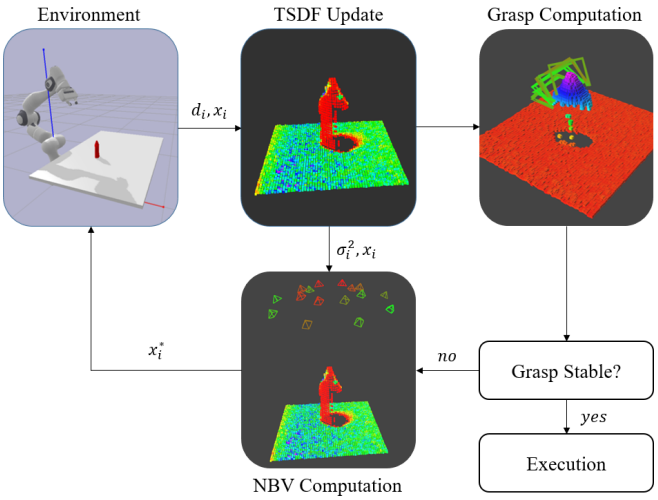

Fig. 1: Schematic overview of the framework. New depth images d_i, taken from sensor poses T_i are integrated continuously. Until a stable grasp is found, the next-best-view with greatest utility T_i^* is calculated and approached.

the object and compute a set of grasp candidates G_i as well as corresponding probabilities of success. If no probability exceeds a predefined threshold, we compute the view T_i^* with the greatest utility from the set of NBV candidates. A Cartesian velocity controller is then used to continuously guide the sensor towards the most promising viewpoint in a closed loop manner. The implemented subsystems data fusion, exploration, grasp sampling and evaluation are presented in the following sections IV - VII

IV. Data Fusion

A common strategy for sensor fusion algorithms to fuse new sensor data into a truncated signed distance function (TSDF) is to use an inverse variance weighting scheme, where the variance of the measurement is estimated by an empirically determined sensor noise model. Although this method takes into account the varying strength of sensor noise due to geometric factors such as depth and angle, additional interferences such as illumination [7] or material [12] cannot be taken into account.

Parallel-jaw grasping requires a very accurate surface estimation since surface normals, i.e. the local gradients, strongly influence the outcome of the grasp, which implies a great sensitivity against sensor noise. Taking the local estimation variance of contact areas into account could therefore prevent many failure cases.

We opt for a probabilistic TSDF approach of previous work [20], which is is briefly summarized here. Each voxel contains a tuple $(\hat{\mu}, W, \hat{\tau}, v)$, where $\hat{\mu}$ is estimated distance of voxel's center towards the closest surface and W is the accumulated weight aka the inverse estimation variance. $\hat{\tau}$ represents an estimate for the measurement scatter caused by the surface properties of the closest surface element and v its corresponding estimation variance. We model the

measurement x_i at time-step i of one surface patch as

$$x_i \sim \mu + N(0, \sigma_{s,i}^2) + N(0, \tau^2) \quad , \tag{1}$$

where μ is the real distance from the camera's focal point to the surface patch. The standard deviation of the sensor $\sigma_{s,i}$ can be empirically determined beforehand e.g. [1] and τ describes the standard deviation introduced by the local surface properties.

Compared to [18] where only the empirical sensor variance is used to update voxel estimations and to [19] where the reliability of an estimate is determined solely by past measurements, the advantage of this modeling is that both existing empirical knowledge about the precision of the sensor as well as an estimate of the unknown fraction of sensor noise can be included and a more accurate estimation of voxel's reliability given. Given the set of measurements X and corresponding known sensor noise variances σ_i^2, estimations for $\hat{\mu}$ and τ are often obtained via a maximum likelihood estimation. We opt for a recursive scheme instead, in order to keep memory requirements as well as update performance constant. We propose to recursively refine initial $\hat{\mu}_0$ and $\hat{\tau}_0$ in a Bayesian manner. The least squares update equation is given by

$$\hat{\mu}_i = \alpha_i x_i + (1 - \alpha_i)\, \hat{\mu}_{i-1} \quad , \tag{2}$$

with

$$\alpha_i = \frac{\rho_i}{\sum_{j=1}^{i} \rho_j} = \frac{\rho_i}{W_i} \quad ,$$

where $\hat{\mu}_i$ is optimal for measurement weight $\rho_i = 1/(\sigma_i^2 + \tau^2)$ also known as inverse variance weighting. Since τ is unknown as well, we propose to estimate τ^2 by a similar linear update scheme. Let $x_i' = x_i - \mu_i$, then

$$\hat{\tau}_i^2 = \beta_i(x_i'^2 - \sigma_{s,k}^2) + (1 - \beta_i)\hat{\tau}_{i-1}^2 \quad . \tag{3}$$

The variance $v_i = \mathrm{Var}(\hat{\tau}_i^2)$ of this estimator is given by

$$v_i = \beta_i^2 \mathrm{Var}(x_i'^2 - \sigma_i^2) + (1 - \beta_i)^2 v_{i-1} \quad , \tag{4}$$

where

$$\mathrm{Var}\left(x_i'^2 - \sigma_i^2\right) = 2\left(\sigma_{s,i}^2 - \tau^2\right)^2 \quad . \tag{5}$$

The optimal estimator yields minimal variance. Hence, the optimal update weight β_i is found by differentiating (5) and given by

$$\beta_i = \frac{v_{i-1}}{2\left(\tau^2 + \sigma_{s,i}^2\right)^2 + v_{i-1}} \quad , \tag{6}$$

where τ^2 is replaced by the current best estimate $\hat{\tau}_i^2$. The state of one voxel is defined by $(\hat{\mu}_i, \hat{\tau}_i, v_i, W_i)$ and updated using equations (2 - 6). By estimating τ parallel to μ, the variance $1/W_i$ of the current estimate can be specified voxel-wise and converge at a different rate, depending on the local noise behavior of past measurements. An example for this behaviour is illustrated in figure (2) where the sensor was aimed head-on at the two cups shown in the top image and remained static during the recording.

V. Computation of Next-Best-View

For reasons of readability, the time index i is omitted in the following chapters. We distribute a set of view candidates $\mathbf{T}^* = \{T_1^*, ..., T_{16}^*\}$, where $\mathbf{T}^* \subset SE3$ on the hemisphere placed around target's bounding box. The radius of the sphere was chosen such that the distance between the object's bounding box and the spheres surface is greater than the minimum measurable distance of the camera. This approach was partially adopted by Breyer et al. [2]. See figure (3) for a graphical representation of the view candidate placement. For each frame, the utility of all view candidates is computed in order to determine the next motion direction of the robot arm.

Compared to traditional exploration approaches, the objective of our reconstruction algorithm is not tailored towards maximum completeness but rather towards reducing the uncertainty of the representation. For example, in case of a grasp pose with high uncertainty it might be beneficial to view one of the corresponding contact areas again in order to improve the chance of success and prevent potentially infeasible actions. This is impossible if only a binary distinction between known and unknown volume as proposed by [2] is used. To compute the information gain of a view candidate T_n^* we choose the Average Entropy, introduced by [13] which is defined as

$$\bar{H}(T_n^*) = \frac{1}{\|V(T_n^*)\|} \sum_{v \in V(T_n^*)} H(v) \quad , \tag{7}$$

where $V(T_n^*)$ is the set of visible voxels from T_n^* within the objects bounding box that store distance values smaller than the truncation threshold. Visible voxels are found via a raycasting algorithm. Under the assumption of a Gaussian distribution, the entropy $H(v)$ of a voxel v is given by $H(v) = 0.5 \log\left(2\pi e\left(\sigma^2 + \epsilon\right)\right)$, where $\sigma^2 = 1/W_i$ is the voxel's current estimation variance and $\epsilon = 1/(2\pi e)$ is

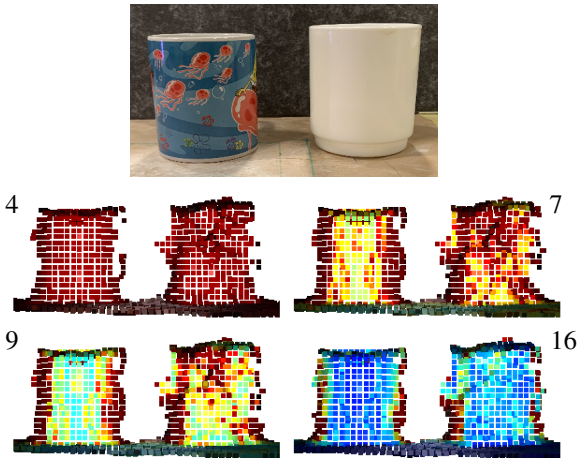

Fig. 2: The convergence of $1/W_i$ for noisy surface segments (right cup) vs. less noisy surfaces (left cup) after 4, 7, 9 and 16 views. The points color indicates the corresponding estimation variance $1/W_i$, where red signals a high variance and dark blue represents values close to zero.

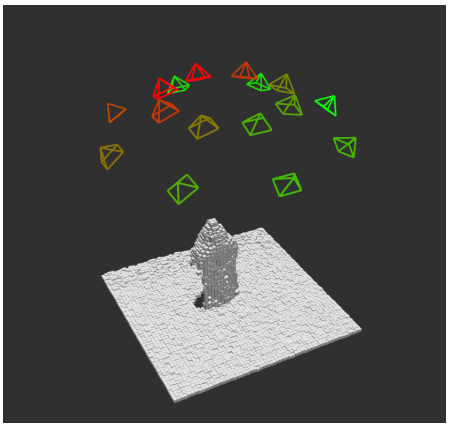

Fig. 3: We create the set of views on the semi-sphere above the object. The utility value of a view (8) is displayed in the color channel. The greener the frustum is displayed the higher the corresponding utility value of the view.

used to close the gap between differential and Shannon entropy.Compared to other metrics for information gain, i.e. the number of the visible, previously unobserved voxels or frontier voxels, the average entropy permits the examination of known estimation variances and therefore aligns with our goal of computing local grasping success probabilities. It can yield high information gain even for known surfaces where the rays traverse fewer unknown voxels and it prevents the controller from being stuck in case all potentially visible voxels are considered known. It continuously refines the reconstruction until a valid grasp is found.

In order to prevent the controller from oscillating between different promising views, we employ the utility function proposed by [8]. Let T be the current sensor pose, the utility U of one view candidate $T_n^* \in \mathbf{T}^*$ is then computed as

$$U(T_n^*) = (1-\gamma)\frac{\overline{H}(T_n^*)}{\sum_{\mathbf{T}^*}\overline{H}} - \gamma\frac{C(T,T_n^*)}{\sum_{\mathbf{T}^*}C} \quad , \tag{8}$$

where $\sum_{\mathbf{T}^*}\overline{H}$ and $\sum_{\mathbf{T}^*}C$ are the summed average entropy and cost of all view candidates respectively and $\gamma \in [0,1]$ is a user defined cost weight. We model the cost $C(T,T_n^*)$ to be proportional to the central angle between the current angular position of the camera on the sphere (λ,ϕ) and the view candidates position (λ_n,ϕ_n)

$$C(T,T_n^*) = \arccos\left(\sin\phi\sin\phi_n + \cos\phi\cos\phi_n\cos\Delta\lambda\right) \quad , \tag{9}$$

where $\Delta\lambda$ is the absolute distance between both longitude angles λ. The next best view T^* is then found by maximizing (8),

$$T^* = \underset{\mathbf{T}^*}{\operatorname{argmax}}\, U \quad . \tag{10}$$

In our setup, the robot's starting pose is outside the sphere. In order to avoid potential collisions, the end effector does not cross the semi-sphere surface on which the views are located.We found that modelling the costs as the central angle often leads to a more efficient approach phase, since neighboring views are assigned a relatively small cost value. In case of a cluttered setup this often causes a left-right alternation to look past the obstacle instead of directly steering towards the opposing side. A simulated example of this behaviour can be seen on the right in figure (4) where the target (red) object is initially partially occluded. Our algorithm initially steers to the right in order to look past the obstacle before the utility values of the left side surpass those of the right. The left image shows a common use case where only one side of target object is initially visible and the algorithm opts for movement towards the opposing side. Once sufficiently reliable, antipodal contact points are determined, valid grasp poses (shown in green) are usually found and executed.

VI. Point Estimates and Surface Gradients

Let v be an arbitrary voxel, (x,y,z) its indices and $\hat{\mu}$ be the value of the corresponding TSDF estimation. The gradient $g \in \mathbb{R}^3$ is found by linearizing the N_6 neighbourhood of v. Since the estimator $\hat{\mu}$ in (2) is a linear combination of normal and independent random variables, $\hat{\mu}$ follows a normal distribution as well, and consequently $g \sim N(\mu_g, \Sigma_g)$ with

$$\hat{\mu}_g = \begin{bmatrix} \hat{\mu}(x_p,y,z) - \hat{\mu}(x_n,y,z) \\ \hat{\mu}(x,y_p,z) - \hat{\mu}(x,y_n,z) \\ \hat{\mu}(x,y,z_p) - \hat{\mu}(x,y,z_n) \end{bmatrix} \quad , \tag{11}$$

and

$$\hat{\Sigma}_g = I \cdot \begin{bmatrix} 1/W(x_p,y,z) + 1/W(x_n,y,z) \\ 1/W(x,y_p,z) + 1/W(x,y_n,z) \\ 1/W(x,y,z_p) + 1/W(x,y,z_n) \end{bmatrix} \quad , \tag{12}$$

where W is sum of weights the voxel received and I represents the identity matrix. The subscripts p and n denote the next neighbours in the respective dimension. If for a pair (v_1,v_2) of N_6 adjacent voxels the condition $\hat{\mu_1}\cdot\hat{\mu_2} < 0$ is met, a surface point between them is computed via linear interpolation of corresponding voxel positions. The gradient g_s of the surface point is then determined in a similar manner,

$$g \sim N\left(\beta\mu_{g,1} + (1-\beta)\mu_{g,2},\ \beta\,\Sigma_{g,1} + (1-\beta)\,\Sigma_{g,2}\right) \quad , \tag{13}$$

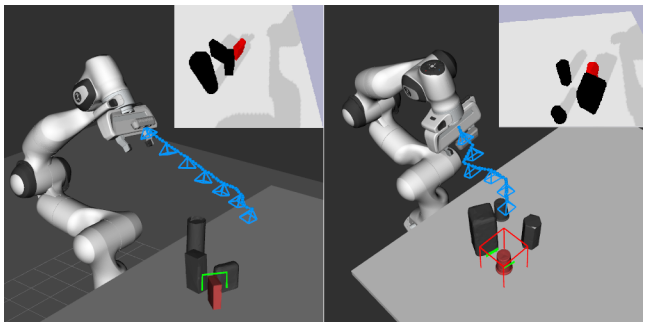

Fig. 4: Two exemplary trajectories of the simulated *cluttered* setup and corresponding initial views. The target object is depicted in red and obstacle objects in black. The right image additionally shows the bounding box box within the entropy is calculated.

with $\beta = -\hat{\mu}_1 / (\hat{\mu}_1 + \hat{\mu}_2)$. The gradient g is then the unnormalized vector perpendicular to the local surface. The evaluation of all possible pairs of contact points has a complexity of $\mathcal{O}(n^2)$ where n is the considered number of surface points within the bounding box. We apply two heuristics to keep n small and ensure the real-time capability of the system. First, we reject all points where $\angle(\hat{z}, g) > \theta_{up}$, where $\hat{z}$ is the unit vector in upwards direction. This effectively removes the ground plane as well as upper points of the object where finding an antipodal, collision-free match is unlikely. Second, we reject points where gradient significantly differs from its neighbours

$$\operatorname{argmax}(\angle(g, g_c)) > \theta_n \quad g_c \in G_c \ , \tag{14}$$

where the set of connected surface gradients G_c is determined by the classical Marching Cubes algorithm. This heuristic leads to a higher robustness of the system as it removes points in high curvature segments of the TSDF. Although grasps at such surface segments, such as the diagonal of a box-like object, are technically correct, they require highly accurate forward kinematics which in case of the *Franka Panda* can have positioning errors > 1 cm [24]. We evaluate the probability of enabling a safe grasp for all remaining possible combinations of surface estimates,

VII. Grasp Evaluation

We rely on the force closure (FC) metric to score grasp candidates. Assuming the Coulomb friction model and a constant friction coefficient f, FC is a binary classifier that determines if any external forces and moments can be compensated by a positive linear combination of contact forces exerted by the fingers. Geometrically, $\theta_f = \arctan(f)$ can be interpreted as the maximum angle under which the object can be loaded without slipping, aka. critical ramp angle. Let (p_1, p_2) be a pair of potential contact points with corresponding normal vectors (n_1, n_2). This pair is in force closure if

$$n_1 \cdot \frac{p_1 - p_2}{\|p_1 - p_2\|} > c_f \quad \wedge \quad c_f < n_2 \cdot \frac{p_2 - p_1}{\|p_2 - p_1\|} \ , \tag{15}$$

where $c_f = \cos(\arctan(f))$ and f is the friction coefficient. We modify the FC-classifier to enable for a qualitative comparison of grasp candidates. Let p be the vector from p_1 to p_2. We are interested in the probability of the grasp defined by p, n_1 and n_2 being in force closure. The surface normal is calculated by normalizing the TSDF gradient g. Assuming independence of both sides, this probability can be defined using (15),

$$P_{fc} = P\left(\frac{p \cdot g_1}{\|p\|\|g_1\|} > c_f\right) \cdot P\left(\frac{p \cdot g_2}{\|p\|\|g_2\|} > c_f\right) \ . \tag{16}$$

The normalized Gaussian vector $g/\|g\|$ is equivalent to its projection on the unit sphere, which is said to have a projected normal distribution $PN_3(\mu, \Sigma)$. The polar representation of g is given by $u = \left(\sin(\theta)\cos(\varphi), \sin(\theta)\sin(\varphi), \cos(\theta)\right)$ and $r = \|g\|$. Its probability density function (pdf) is as follows,

$$\text{pdf}(r, \theta, \varphi) = \left(\prod_i^3 2\pi\sigma_i^2\right)^{-0.5} \exp\left(-\sum_i^3 \frac{(r\ u_i - \mu_i)^2}{2\sigma_i^2}\right) \ r^2 \ . \tag{17}$$

The shape of the distribution does not change with rotations of the spherical coordinate system. Hence, we are able to choose the coordinate system such that p aligns with the polar axis. The pdf of the angle $\theta = \angle(\frac{g}{\|g\|}, p)$ is then given by

$$\text{pdf}(\theta) = \int_0^\infty \int_0^{2\pi} g(r, \theta)\ dr\ d\varphi \ , \tag{18}$$

and the probability of the surface normal being within the friction cone is consequently given by

$$P\left(\frac{p \cdot g}{\|p\|\|g\|} > c_f\right) = \int_0^{\theta_f} \text{pdf}(\theta)\ d\theta \ . \tag{19}$$

The polar representation of (19) is shown schematically in figure (5). Although a parameterized representation for the Projected Normal Distribution exists for the three dimensional case and would allow for the evaluation of (19), we opt for a computationally more efficient look-up-table approach. To reduce the dimensionality of the approach we perform the simplification $\Sigma_g \approx \sigma_m^2\ I_3$ where I is the identity matrix and $\sigma_m = max(\sigma_1, \sigma_2, \sigma_3)$ in order to err on the side of caution. Additionally, the shape of the distribution does not change if the random variable g is scaled for any $a > 0$. The look-up-table can therefore be parameterized by the ratio of the expected length to the maximum standard deviation along all axes $E(\|g\|)/\sigma_m$ and the expected angle $E(\theta)$. Hence, we discretize all combinations of $\theta \in [0, \pi/3]$ and $\sigma_m/\|g\| \in]0, 50]$ into a grid. For each grid index we use a large number of simulated trials and consider the number of positive evaluations of $\frac{p \cdot g}{\|p\|\|g\|} > c_f$ as associated probability value. The grid is then stored and used as look-up table where (19) can be conveniently evaluated at run-time with high performance. A user defined threshold is used to determine whether two points are considered in force closure or not and a point pair is stored for further evaluations if

$$P_{fc} > P_{thr} \tag{20}$$

is satisfied. P_{thr} acts as a trade-off between grasp safety and search time, since the $W_i = \sum_i 1/\left(\sigma_i^2 + \hat{\tau}_i\right)$ scales with the number of measurements the voxel received. Although we

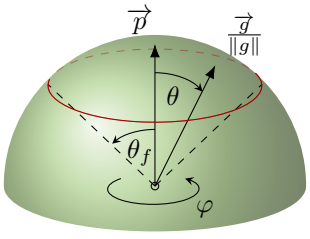

Fig. 5: Polar representation of the vector p connecting the antipodal points and the friction cone. Equation (19) evaluates the likelihood of $g/\|g\|$ being within the friction cone.

model the measurements according to (1), the grasp probability estimation is similar for all Gaussian noise assumptions.

VIII. Collision Checks

If (20) is satisfied, the corresponding pair of antipodal points is likely in force closure and potential poses on the circle perpendicular to p with radius equivalent to the gripper depth need to be evaluated. Inspired by [5], we use the TSDF to resolve this redundancy and eliminate potential collisions with the environment at the same time. As the TSDF is a spatial representation of the environment, we can project the gripper into the volume and check occupied SDF values for collision. While [5] uses the mesh of the end effector, we uniformly discretize the volume of the end effector and the fingers in the set of points EE. This conveniently allows to adapt the discretisation to the resolution to the TSDF.
We uniformly distribute a set of grasp poses T_v on the upper semicircle with angular distances of $10\,^\circ$. All poses that either occupy unobserved voxels or ones with negative SDF values are discarded. Of the remaining ones only the grasp pose T_g with the greatest sum of truncated distances is considered for further evaluation

$$T_g = \operatorname{argmax}\left(\sum_i t_i\right), \quad t = \text{TSDF}\left(T_v \cdot EE\right) \quad . \tag{21}$$

Compared to using the surface vertices as [5], a volumetric representation leads to a more reliable collision evaluation since the occupied space is represented uniformly whereas a mesh can over- or under represent segments and even leave small collisions in low resolution areas undetected. Eq. (21) usually leads to the approach vector being perpendicular to the objects surface, which we consider the desired result since this configuration leaves the largest margin for errors during the approach phase of the robot.

IX. Simulated experiments

We evaluate the proposed approach in two different setups simulated with the physics engine *PyBullet*, which allows us to test our approach over a large number of randomly generated scenes. The first setup was adopted from [2], where randomly chosen objects are iteratively placed at random positions on the table in front of a *Franka Panda* arm in an upright manner. Positions that end up in collision with already placed objects are rejected and resampled until either four objects are placed or a maximum number of attempts is reached. The objects are mostly cylindrical or box-like and large enough to lead to significant occlusions and potential collisions. The second setup has a similar configuration except for the type and number of objects. For each iteration we only use a single object randomly chosen from a subset[1] of the YCB dataset [6], which are more complex compared to the first setup. The subset provides simplified collision mesh files and physics parameters that are fine-tuned and tested for robotic manipulation tasks in *PyBullet*. Since grasps that require less friction are arguably better and in order to make both setups more discriminative we set the lateral friction coefficients of all objects to 0.2 compared to the setting of 1.0 in [2] and the original range of the YCB subset [0.3, 0.8]. The lateral friction of the fingers was left fixed at a constant value of 0.5 for all experiments. For both setups the depth images were imposed with Gaussian noise according the the noise model of [1]. The bounding box of the target object is provided by the physics simulator. A grasp was considered a success if the object could be lifted by 10 cm. Although each grasp was checked for inverse kinematic solutions and for collision-free placement of the end effector, occasionally *MoveIt* failed to find a plan. These cases were removed from the results. All tests were performed using a *NVIDIA 3090 RTX* graphics card and an *Intel i7-9700K* CPU. Table I lists the used parameters of our approach. We use the following metrics for performance evaluation:

- Success Rate (**SR**): Ratio of runs where the target was successfully grasped.
- Failure Rate (**FR**): Ratio of runs where a grasp was detected, but failed during execution.
- Abortion Rate (**AR**): Ratio of runs where no grasp on the target object was found.
- Search time (**ST**): Time elapsed between receiving a bounding box and returning a grasp configuration.
- Distance (**D**): Distance traveled by the end effector before opting for grasp execution.

We compare our results against those of [2] (vgn) where we left the parameters at their default values except for the window size which was set to 25 since the authors report the highest success rate for this setting. Additionally, we

TABLE I: Parameters used for the experiments.

TSDF size		0.3^3 m^3
Voxel count per side		80
Policy rate		5 Hz
Number of view candidates	$\lvert T_c \rvert$	16
Maximum number of views		100
Probability threshold	P_{thr}	85%
Linear velocity		5 cm/s
Gripper Force		10 N
Critical ramp angle	arctan(f)	20°
Initial τ estimation	τ_0	0.9
Initial v value	v_0	0.0375
Utility weight	γ	0.05

TABLE II: Results for both simulated setups.

Setup	Policy	SR	FR	AR	ST (s)	D (m)
clutter	$ours$	92 %	7 %	1 %	6.2	0.20
	$ours_{se}$	91 %	8 %	1 %	5.6	0.18
	$ours_{rs}$	86 %	7 %	7 %	6.4	0.21
	vgn	55 %	37 %	8 %	8.3	0.29
	$vcpd$	52 %	48 %	-	4.5	0.12
	$o.w.vcpd_{10}$	58 %	42 %	-	3.7	0.08
	$o.w.vcpd_{20}$	73 %	27 %	-	5.7	0.17
	$o.w.vcpd_{30}$	87 %	13 %	-	7.8	0.25
single	$ours$	90 %	8 %	2 %	7.1	0.22
	vgn	77 %	21 %	2 %	8.9	0.30
	$vcpd$	51 %	49 %	-	5.1	0.15
	$o.w.vcpd_{10}$	58 %	42 %	-	3.8	0.08
	$o.w.vcpd_{20}$	64 %	38 %	-	5.8	0.17
	$o.w.vcpd_{30}$	75 %	25 %	-	7.9	0.26

[1] *https://github.com/eleramp/pybullet-object-models*

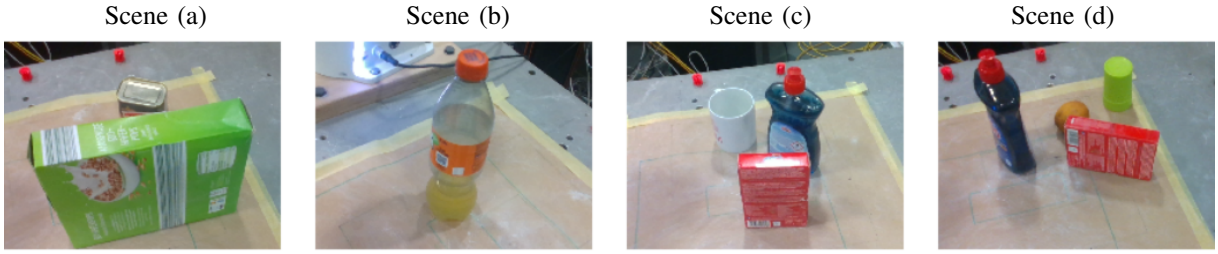

Fig. 6: Images of the initial camera view for each tested setting. In scene (a) the target object is the corned beef can behind the cornflakes box, in scene (c) the dishwashing liquid and in scene (d) the lemon in the image center.

compare our results against those of Cai et al. [5] ($vcpd$) who propose a closed loop approach where a neural network is leveraged to predict the quality of contact points. Their control strategy commands linear movement towards the grasp pose with highest estimated quality until a distance threshold is reached without an exploration strategy. If appropriate grasping possibilities are initially occluded, the algorithm might settle for insufficient ones. Therefore, we additionally implemented the network as grasping evaluator ($o.w.vcpd_n$) in our pipeline in order to measure the performance gain introduced by our exploratory method. Due to the lack of a built-in quality threshold we force the exploration for n views before opting for grasp evaluation and execution.

In order to evaluate effectiveness of the employed information gain formulation we additionally implemented two other formulations in the proposed pipeline ($ours_{se}$, $ours_{rs}$). The first is the summed entropy, which is similar to eq. (7) but without averaging over the visible voxels. The second one is the rearside formulation, as utilized by [3], which counts the number of visible voxels with negative SDF values. The mean results of $\sim$240 trials in both setups can be seen in table II. Table III depicts noteworthy computation times of algorithm components. Since $vcpd$ performs maximum selection of estimated grasp qualities as well as collision scores its abortion rate is at a constant 0%. The success rate of trials where we implemented as an alternative grasping evaluator ($o.w.vcpd_n$) into our approach show a significant improvement compared to the original closed-loop approach $vcpd$. Our utility function (8) often computed trajectories similar to the one visible on the right in figure (4) which lead to more complete object representations and helped the network to make informed decisions, where in case of $n = 20$ the slightly increased distance and search time are negligible. In the cluttered setup the performance of our probabilistic force closure grasp evaluator is comparable to $o.w.vcpd_{30}$. However, our approach shows a significant advantage in the single setup where the objects are more complex. This might be related to the fact that the network of [5] was trained using primitive-shaped objects which were also used in the cluttered setup and highlights the invariance of our approach to the used object category.

The beneficial effect of entropy based information gain formulation on our approach is evident in results table II. The binary distinction into known and unknown volume did not motivate our algorithm to improve the reconstruction and smearing effects of the TSDF had a greater impact on the performance, leading to a higher failure rate. Additionally, the robot could get stuck when the rear side has been fully explored and still no stable grasp has been found, leading to an increased number of aborted runs. Although the proposed IG formulation shows a slightly lower abortion rate than I_{ae} the results are not conclusive and further tests in more diverse setups might be necessary.

X. Real-world experiments

The intended main benefit of our algorithm is the ability to handle severe surface dependent noise, which is difficult to reproduce in simulation. Figure (6) presents four scenarios that highlight the advantages of our approach and showcase the challenges due to partial occlusion and obstacles (scene a, c, d), transparency (b, c) and reflections (a) in the assistive-robot context. With the exception of the lemon, recordings of all objects show strong sensor noise that varies across the surface. This leads to reconstructions that are heavily distorted in some sections and emphasises the advantage including the estimation variance in the modeling of grasp quality. Again, we compare our approach against (vgn) and ($vcpd$) and list the results in table IV. Our approach shows a comparatively steady performance across all scene. It focused on the comparatively noise-free regions, e.g head and belly of the bottle in scene b and c or the sides of the box in scene a and reliably avoided the distorted regions with significantly higher estimation variance.

XI. Conclusion

We presented a closed-loop grasp system based on a probabilistic, truncated signed distance function, a next-best-view planner based on entropy considerations and a probabilistic grasp detection algorithm to achieve 6-dof grasping of unknown household objects with real-time performance. The system is able to deal with the various sources of error in assistive robotics, such as heavy, non-constant sensor noise and occlusions and does neither require prior object knowledge or a dataset tailored to the use case. The presented path planning method is not dependent on an initial grasp proposal, it instead determines the trajectory by entropy considerations. This allows our algorithm to consistently find stable grasps where others perform worse. Simulated as

TABLE III: Mean durations [ms] and std. deviations.

Integration of depth image	4.9 ± 2.2
Evaluate contact points	15.4 ± 8.3
Grasp selection	7.8 ± 3.2
Information gain computation	21.2 ± 4.7
Complete policy update	38.2 ± 11.8

TABLE IV: Results from real world experiments.

Scene	Policy	SR	FR	AR	ST (s)	D (m)
(a)	$ours$	9 / 10	0 / 10	1 / 10	14.5	0.52
	vgn	6 / 10	4 / 10	0 / 10	13.1	0.38
	$vcpd$	3 / 10	7 / 10	0 / 10	11.0	0.39
	$o.w.vcpd_{60}$	8 / 10	2 / 10	0 / 10	15.2	0.55
(b)	$ours$	10 / 10	0 / 10	0 / 10	18.6	0.67
	vgn	8 / 10	2 / 10	0 / 10	13.5	0.40
	$vcpd$	1 / 10	9 / 10	0 / 10	9.5	0.32
	$o.w.vcpd_{60}$	6 / 10	4 / 10	0 / 10	16.5	0.59
(c)	$ours$	10 / 10	0 / 10	0 / 10	16.7	0.57
	vgn	9 / 10	1 / 10	0 / 10	16.3	0.55
	$vcpd$	3 / 10	7 / 10	0 / 10	10.2	0.35
	$o.w.vcpd_{60}$	7 / 10	3 / 10	0 / 10	16.2	0.59
(d)	$ours$	9 / 10	2 / 10	0 / 10	14.4	0.52
	vgn	8 / 10	0 / 10	2 / 10	16.0	0.50
	$vcpd$	5 / 10	6 / 10	0 / 10	12.2	0.45
	$o.w.vcpd_{60}$	9 / 10	2 / 10	0 / 10	14.1	0.51

well as real-world experiments show the effectiveness and reliability of the approach.

REFERENCES

[1] Min Sung Ahn et al. "Analysis and Noise Modeling of the Intel RealSense D435 for Mobile Robots". In: *16th International Conference on Ubiquitous Robots (UR)* (2019), pp. 707–711.

[2] Michel Breyer et al. "Closed-Loop Next-Best-View Planning for Target-Driven Grasping". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2022), pp. 1411–1416.

[3] Michel Breyer et al. "Volumetric Grasping Network: Real-time 6 DOF Grasp Detection in Clutter". In: *Conference on Robot Learning*. 2021.

[4] Junhao Cai et al. "Real-time Collision-free Grasp Pose Detection with Geometry-aware Refinement Using High-Resolution Volume". In: *IEEE Robotics and Automation Letters* PP (2022), pp. 1–1.

[5] Junhao Cai et al. "Volumetric-based Contact Point Detection for 7-DoF Grasping". In: *ArXiv* abs/2209.06675 (2022).

[6] Berk Çalli et al. "Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols". In: *ArXiv* abs/1502.03143 (2015).

[7] Rui Chen, Jing Xu, and Song Zhang. "Comparative study on 3D optical sensors for short range applications". In: *Optics and Lasers in Engineering* (2022).

[8] Jeffrey A. Delmerico et al. "A comparison of volumetric information gain metrics for active 3D object reconstruction". In: *Autonomous Robots* 42 (2018), pp. 197–208.

[9] Minghao Gou et al. "RGB Matters: Learning 7-DoF Grasp Poses on Monocular RGBD Images". In: *IEEE International Conference on Robotics and Automation (ICRA)* (2021), pp. 13459–13466.

[10] Phillip M Grice and Charles C Kemp. "Assistive mobile manipulation: Designing for operators with motor impairments". In: *RSS Workshop on Socially and Physically Assistive Robotics for Humanity*. 2016.

[11] Marcus Gualtieri and Robert W. Platt. "Viewpoint selection for grasp detection". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2016), pp. 258–264.

[12] Georg Halmetschlager-Funek et al. "An Empirical Evaluation of Ten Depth Cameras: Bias, Precision, Lateral Noise, Different Lighting Conditions and Materials, and Multiple Sensor Setups in Indoor Environments". In: *IEEE Robotics & Automation Magazine* 26 (2019), pp. 67–77.

[13] Simon Kriegel et al. "Efficient next-best-scan planning for autonomous 3D surface reconstruction of unknown objects". In: *Journal of Real-Time Image Processing* 10 (2015), pp. 611–631.

[14] Haoxiang Ma and Di Huang. "Towards Scale Balanced 6-DoF Grasp Detection in Cluttered Scenes". In: *Conference on Robot Learning*. 2022.

[15] Jeffrey Mahler et al. "Learning ambidextrous robot grasping policies". In: *Science Robotics* 4 (2019).

[16] Douglas Morrison, Peter Corke, and J. Leitner. "Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach". In: *ArXiv* abs/1804.05172 (2018).

[17] Adithyavairavan Murali et al. "6-DOF Grasping for Target-driven Object Manipulation in Clutter". In: *IEEE International Conference on Robotics and Automation (ICRA)* (2019), pp. 6232–6238.

[18] Kelsey Saulnier et al. "Information Theoretic Active Exploration in Signed Distance Fields". In: *IEEE International Conference on Robotics and Automation (ICRA)* (2020), pp. 4080–4085.

[19] Henry Schaub, Alfred Schöttl, and Maximilian Hoh. "Probabilistic Fusion of Depth Maps With a Reliable Estimation of the Local Reconstruction Quality". In: *IEEE Robotics and Automation Letters* 7 (2022), pp. 11982–11989.

[20] Henry Schaub et al. "Probabilistic Fusion of Depth Maps with Local Variance Estimation". In: *IEEE Sensors* (2023).

[21] Shuran Song et al. "Grasping in the Wild: Learning 6DoF Closed-Loop Grasping From Low-Cost Demonstrations". In: *IEEE Robotics and Automation Letters* 5 (2019), pp. 4978–4985.

[22] Martin Sundermeyer et al. "Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes". In: *IEEE International Conference on Robotics and Automation (ICRA)* (2021), pp. 13438–13444.

[23] Xiaoshuai Zhang et al. "Close the Optical Sensing Domain Gap by Physics-Grounded Active Stereo Sensor Simulation". In: *IEEE Transactions on Robotics* 39 (2022), pp. 2429–2447.

[24] Leon Žlajpah and Tadej Petri. "Kinematic calibration for collaborative robots on a mobile platform using motion capture system". In: *Robotics Comput. Integr. Manuf.* 79 (2023), p. 102446.