Visuo-Haptic Grasping of Unknown Objects based on Gaussian Process Implicit Surfaces and Deep Learning

Simon Ottenhaus, Daniel Renninghoff, Raphael Grimm, Fabio Ferreira and Tamim Asfour

***Abstract*— Grasping unknown objects is a challenging task for humanoid robots, as planning and execution have to cope with noisy sensor data. This work presents a framework, which integrates sensing, planning and acting in one visuo-haptic grasping pipeline. Visual and tactile perception are fused using Gaussian Process Implicit Surfaces to estimate the object surface. Two grasp planners then generate grasp candidates, which are used to train a neural network to determine the best grasp. The main contribution of this work is the introduction of a discriminative deep neural network for scoring grasp hypotheses for underactuated humanoid hands. The pipeline delivers full 6D grasp poses for multi-fingered humanoid hands but it is not limited to any specific gripper. The pipeline is trained and evaluated in simulation, based on objects from the YCB and KIT object sets, resulting in a 95 % success rate regarding force-closure. To prove the validity of the proposed approach, the pipeline is executed on the humanoid robot ARMAR-6 in experiments with eight non-trivial objects using an underactuated five finger hand.**

I. Introduction

Grasping objects is a central capability for humanoid robots, as it is a prerequisite of object manipulation. Grasping is a challenging task, which has been approached from many directions in the past, including different robots, hands, sensors, and algorithms [1]. In particular, grasping with humanoid hands is more challenging than grasping with grippers, since it requires the generation of full 6D poses for multi-fingered hands [2]. Robots which have to work unstructured and partially unknown environments must be able to deal with incomplete and imprecise object models as well as noisy sensor data for successful grasping. Humans learn grasping in the early development stage and fuse visual and tactile information to transfer grasps from one object to another [3].

Inspired by human grasping capabilities, we present a complete pipeline for visuo-haptic grasping of unknown objects leveraging a deep learning approach, as shown in Fig. 1. After capturing a point cloud, the robot gathers information about the unseen sides of the object by tactile exploration. Thereafter the visual and tactile information is fused using Gaussian Process Implicit Surfaces, as proposed by Björkman et al. [4]. A skeleton and a surface-based grasp planner are employed to generate grasp hypotheses based on the estimated surface of the object. The key contribution

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 643950 (SecondHands) and the Helmholtz Association, Project ARCHES (contract number ZT-0033).

The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany, {simon.ottenhaus, asfour}@kit.edu

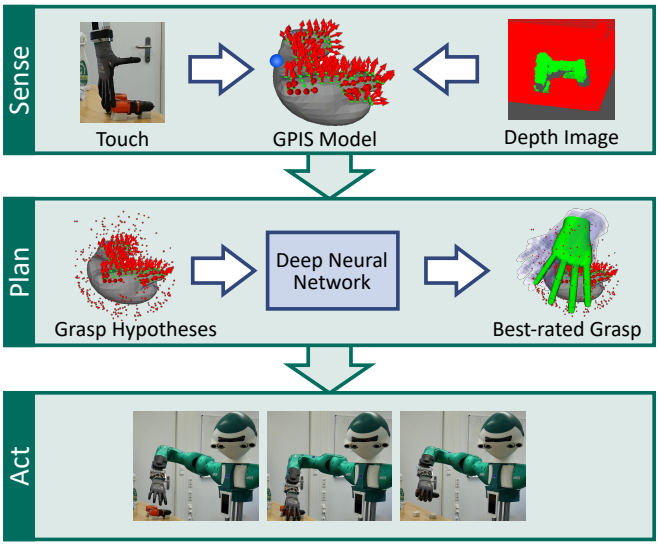

Fig. 1: Proposed Visuo-Haptic Grasping Pipeline: The humanoid robot ARMAR-6 captures a point cloud of an unknown object from the front. The robot's hand touches the object from the back. The information is fused using Gaussian Process Implicit Surfaces (*GPIS*) where visual data is depicted with as arrows and the tactile contact is shown as a blue dot. Multiple grasp hypotheses are generated from the estimated surface, shown as red dots. The neural network rates the hypotheses based on a local 3D voxel-grid view that is aligned with the grasp pose. The best-rated grasp candidate (green hand) is selected for grasp execution.

of this work lies in a discriminative deep neural network (DNN) that is inspired by VoxNet [5]. However, instead of classification, we use the DNN for scoring the grasp hypotheses based on a local view of the estimated object model. This enables the pipeline to reliably discard invalid grasp hypotheses, leaving only grasp hypothesis with high success probability for execution.

II. Related Work

Since grasping is such a diverse and highly active field in robotics, we limit the related work to approaches most relevant to the proposed visuo-haptic grasping pipeline. We discuss publications regarding three different challenges: fusion of visual and tactile sensor information, encoding of point clouds for deep learning through voxel grids in particular, and lastly, both grasp planning and metrics for model-based grasp generation and evaluation.

In their survey, Bohg et al. propose that through *Interactive Perception (IP)* rich sensory data can be obtained [6]. Applied to grasping, IP recognizes dependencies between the tactile and visual perception, as these modalities have

to be fused to facilitate successful grasping. A prominent strategy has been the use of Gaussian Process Implicit Surfaces (GPIS) for estimating object surfaces [7]. GPIS has been widely applied to fuse visual and haptic perception in one joint surface estimate [4, 8, 9], while other approaches exist, e. g. Extended Kalman filter with assumed object symmetry [10]. Maldonado et al. fuse an RGB-D point cloud with observations from a proximity sensor [11].

Recently 3D voxel grids became popular for encoding point clouds for deep neural networks (DNNs). Varley et al. merge tactile and depth information into a shared occupancy map that serves as input to a convolutional neural network (CNN), which generates object geometry hypotheses [12]. By feeding a DNN with the voxel grid representation of a point cloud the full 3D shape of an object can be predicted allowing next-best-view estimation [13]. Yan et al. also use Voxel grids for learning grasps in simulation [14]. Wang et al. use a 2D color image and shape priors as input to their deep neural network to generate a rough 3D shape that is updated by tactile signals [15].

Grasp synthesis can be divided into analytic and data-driven approaches. Analytic approaches rely on kinematic or dynamic grasp simulation, resulting in planners such as *GraspIt!* [16] or *Simox* [17]. Grasp candidates are typically evaluated in the wrench space [18] or by calculating the force closure probability under random perturbations of the grasp [19]. Bohg et al. [1] have subsumed a number of data-driven contributions for grasping unknown objects. Relying on low-level features, Object-Action Complexes (OAC) can be learned by visual feature extraction processes [20]. This system was extended by texture features [21]. Morrison et al. generate grasp candidates from on depth images, based on learning methods [22]. Global object shape can also be used to facilitate 2D contour extraction for grasping [23].

Other approaches employ grasp candidate simulation based on object primitives [24, 25] or shape completion [26]. The generation of grasps based on mean curvature skeletons, local surface structure, and alignment of the hand have also been investigated [27].

In accordance with Interactive Perception, there also exist approaches that additionally use tactile feedback to improve on the visual information. Hsiao et al. [28] use top, side and high point grasps and a reactive grasping heuristic to generate grasp hypotheses. Schiebener et al. [29] use pushing actions to verify visually perceived object hypotheses and apply grasping after successful object segmentation. In addition, further approaches exist that also use interaction to improve scene segmentation [30 – 32].

The use of vision-based deep learning methods for grasping unknown objects has recently led to significant advances [22]. Several approaches use CNNs for frame-based grasp detection that solve either regression or region-proposal classification problems where grasps are encoded as rectangles [22, 33 – 36]. The methods usually vary in the number of outputs (best grasp pose, multiple ranked poses) or in the type of data used (real or simulated, image or depth image). Calandra et al. use tactile sensory data in addition to visual information in order to learn re-grasping policies for jaw-grippers with a CNN [37].

Contributions not limited to parallel jaw-grippers but extended to three- [38] or full-fingered [2] end-effectors have been proposed. Other works have investigated generative models to improve data efficiency of deep robotic grasping and addressed generalization from simulation to real-world scenarios [39, 40]. Based on self-supervised trial and error, deep reinforcement learning methods have been used to learn grasping policies for a jaw-gripper [41] or for multi-fingered hands in simulated scenarios [42] based on visual data.

III. Object Shape Estimation based on GPIS

In the following, we briefly introduce the concept of Gaussian Process Implicit Surfaces (GPIS) and its extension to include surface normal observations for estimation of the object surface. A more detailed explanation can be found in our previous work [43] and in the original GPIS paper [7].

GPIS combines Gaussian processes (GP) with implicit functions for surface estimation. The *Implicit Surface Potential function (ISP)* f is defined for each point in $\mathbb{R}^3$ and can be used to calculate the estimated surface.

$$f(\boldsymbol{x}) : \mathbb{R}^3 \rightarrow \mathbb{R} \begin{cases} = 0, & \boldsymbol{x} \text{ on the surface} \\ > 0, & \boldsymbol{x} \text{ inside} \\ < 0, & \boldsymbol{x} \text{ outside} \end{cases} \tag{1}$$

$$S = \{\boldsymbol{x}, f(\boldsymbol{x}) = 0\} \tag{2}$$

We used observed points on the surface and added outside points. Furthermore, we add normal observations by setting the gradient of the ISP $\nabla f(\boldsymbol{p}_i)$.

$$f(\boldsymbol{p}_i) = \begin{cases} 0, & \boldsymbol{p}_i \text{ observed surface point} \\ -1, & \boldsymbol{p}_i \text{ additional point outside} \end{cases} \tag{3}$$

$$\nabla f(\boldsymbol{p}_i) = (n_{i,1}, n_{i,2}, n_{i,3})^T \ . \tag{4}$$

In this work, we use the GPIS estimate in two different ways. 1) We triangulate the 0-level set of the ISP to get the estimated surface of the object, which is used by the grasp planners to generate grasp hypotheses. 2) Furthermore, we sample the ISP of the GPIS in a 3D voxel grid that is used as an input of the deep neural network.

IV. Visuo-Haptic Grasping Pipeline

We propose a visuo-haptic grasping pipeline consisting of eight stages: visual and tactile perception, sensor fusion, grasp hypotheses generation, filtering, scoring, selection, and execution. Each pipeline step is explained in detail in the following. The full pipeline is depicted in Fig. 2.

In the *Visual Perception* (S_1) pipeline step the robot points the depth camera at the object and captures a depth image, which is transformed to a point cloud. We assume that the object rests on a flat supporting plane, therefore it can easily be segmented using RANSAC [44]. After denoising, the visual perception computes normals from the point cloud, resulting in a set of points with normals on the front and top surface of the object.

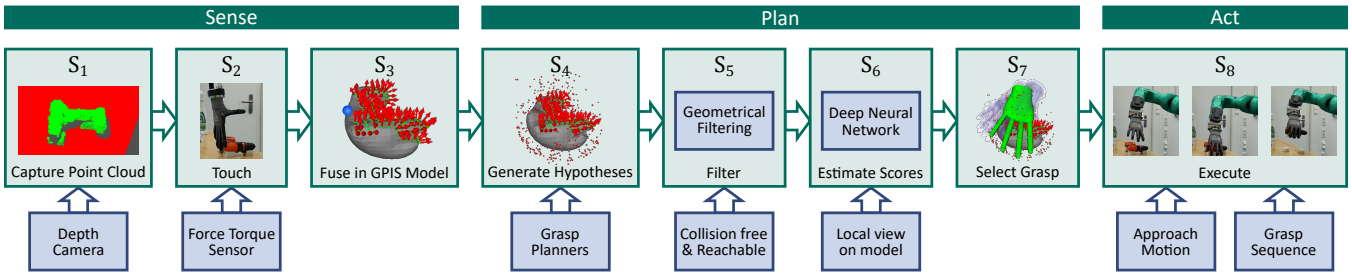

Fig. 2: Proposed visuo-haptic grasping pipeline integrating sensing, planning, and acting.

During *Tactile Exploration* (S_2) of the object, several points from the unseen sides are chosen for exploration. We use a simplified version of the next-best-touch algorithm that we presented previously [45]. We infer tactile contact from forces in the force torque sensor located in the wrist of the robot since the used hand does not provide tactile sensors. In simulation, the contacts are calculated based on the collision between the object mesh and the hand geometry.

A GPIS estimate model fuses tactile and visual data in the third pipeline stage (S_3). Visual points are introduced using contact positions and surface normals, where the normals define the local gradient of the GPIS *Implicit Surface Potential (ISP)*. In order to speed up the matrix inversion necessary for GPIS we reduce the visual point cloud to about 100 points with methods provided by the PCL library [46]. The tactile contacts are encoded as points on the surface, defining the ISP value to 0 on the surface. Additionally, the ISP is constrained by adding outside observations far away from the object.

The *Grasp Hypotheses Generation* (S_4) stage triangulates the GPIS model using the marching cubes algorithm. Based on this estimated surface model, the skeleton [27] and the surface-based grasp planner, provided by Simox [17], generate grasp hypotheses.

The *Geometric Grasp Filtering* (S_5) stage ensures that the generated grasp hypotheses are collision-free and reachable. We employ a simple filtering method regarding grasp position and orientation, so that remaining grasps are reachable and do not collide with the supporting surface.

The *Grasp Hypotheses Scoring* (S_6) stage estimates the scores of the remaining grasps. For each grasp hypothesis, a local voxel grid view of the ISP of the GPIS is generated and fed to the neural network, which predicts the success probability of the grasp hypothesis (P_G) between 0 and 1 for each grasp.

The *Grasp Selection* (S_7) stage selects the grasp with the highest predicted success probability, shown as a green hand, while other feasible grasp hypotheses are shown as light blue hands (see (S_7) in Fig. 2).

During *Grasp Execution* (S_8) the robot calculates an approach vector for the grasp. Then, the hand is controlled in Cartesian velocity mode while the force torque sensor is monitored. When a force is detected, the hand is closed and thereafter the object is lifted. In the simulation, the grasp execution is replaced by moving the hand to the target pose, closing the fingers until contact and then calculating the grasp metric.

V. Scoring Grasp Hypotheses with a DNN

The main contribution of this work is a data-driven grasp metric, that scores grasp hypothesis for underactuated humanoid hands. The grasp metric is implemented as a deep convolutional neural network (subsequently simply referred to as DNN). The network follows a discriminative approach where the input is the current world state (local view of the observed surface points) and the pose of the grasp hypothesis. Based on these inputs, the DNN predicts the success probability of grasp execution by the robot (P_G). In Fig. 3 the network architecture is depicted.

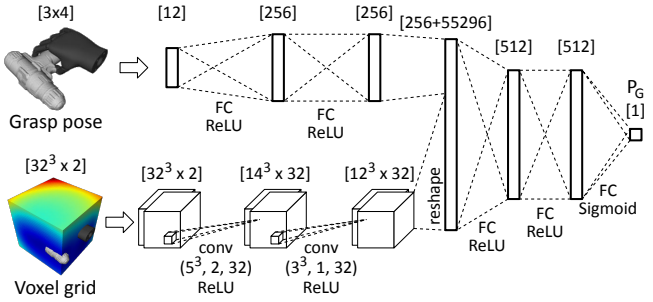

Fig. 3: Structure of the deep neural network: The grasp pose is encoded relative to the object center as the top three rows of the homogeneous pose matrix (3×4). Two fully connected layers preprocess the grasp pose. The local view on the estimated model is encoded in a 3D voxel grid and processed by two 3D convolutional layers $conv(d, s, f)$, where d is the kernel size, s is the stride and f denotes the number of filters. At each voxel center, two features are observed: The *Implicit Surface Potential (ISP)* and the distance to the closest point observed on the surface. The information is fused in two fully connected layers, resulting in the success probability of grasp execution.

For training data generation, the stages $S_1 \ldots S_4$ of the pipeline are executed in simulation. The training object set is comprised of objects from the KIT object database [47] and the YCB object and model set [48]. The simulation loads a random object mesh model and generates a point cloud via a simulated depth camera (S_1). The point cloud is segmented, reduced and normals are computed. Stage 2 adds between one and five touches, which are fused with the visual information in a GPIS model (S_3). Stage 4 then generates grasp hypotheses based on the estimated GPIS model. For each hypothesis, the ground truth P_G is determined by calculating the force-closure probability of the grasp against the ground truth mesh of the object under small random

perturbations. A 32^3 voxel grid with a side length of $30\,\mathrm{cm}$ is generated. The grid is aligned with the grasp pose such that the hand is always in the center of the grid while the relative orientation between hand and grid is fixed. At each voxel center, two features are observed: The ISP of the GPIS and the distance to the closest observed surface point. Aligning the voxel grid with the grasp pose has multiple advantages: The captured information is of high relevance to the grasp success probability since force-closure is only dependent on the local geometry of the object. The voxel grid alignment allows the network to omit the transformation of the object geometry into the correct alignment so that learning can focus on the relevant features. Hence, the network can transfer local part-specific grasps from one object to another object having similar local geometry. Furthermore, a local grid, enclosing only the hand, has a finer resolution than a similar-sized global grid, enclosing the whole object.

Overall, we generated $1.6\,\mathrm{million}$ training samples. For test and validation, we generated 200.000 samples from a separate set of objects not included in the training. The test objects used in the evaluation can be seen in Fig. 4. Each object was randomly rotated and scaled to augment the object set and to increase the training diversity. We used a $50\,\%$ dropout rate and a learning rate of 10^{-4}.

VI. Evaluation

We present a complete pipeline for grasping unknown objects, where the main novelty lies in the rating of grasp hypotheses based on tactile and visual data with a deep neural network. In the evaluation, we want to focus on the effectiveness of the deep neural network (pipeline stage S_6) as well as the tactile exploration impact (S_2). This poses two main questions: What is the benefit of the proposed deep learning model? How does adding tactile exploration to the visual perception improve the grasp success probability? To address these questions, we chose an object set comprised of 12 unseen test objects that were not used for training. The object models were taken from the YCB and KIT object datasets and can be seen in Fig. 4. We separate the test objects into fully unknown objects and unknown, but familiar objects. Unknown objects ("Power Drill" and "Spray Bottle") differ significantly from the objects in the training set, whereas familiar objects have similar a shape as objects in the training set.

A. Baseline and Evaluation Pipeline

Björkman et al. presented in [4] the fusion of tactile and visual data using GPIS for sensor fusion and surface estimation. They propose that grasp planning and execution should be based on the GPIS estimate. In our work, we use this GPIS-based approach as a baseline for comparison. The baseline consists of the full pipeline, where the DNN pipeline stage (S_6) is replaced with a grasp score that is computed on the estimated GPIS model by the employed grasp planner from the Simox framework [17].

During the evaluation, S_8 is replaced by computing the grasp force-closure probability (P_G) in simulation on the ground truth mesh. This is applied for the baseline and for the evaluation of our approach. The grasp force-closure probability is computed by randomly perturbing the grasp position. Force-closure is tested for each position and P_G is calculated from the ratio of force-closure and non force-closure grasps. During the evaluation, we consider a grasp successful if at least $80\,\%$ of the perturbed grasps result in force-closure of the hand and object.

Fig. 4: Object test set for evaluation (1-4 YCB, 5-12 KIT): 1 Power Drill, 2 Apple, 3 Racquetball, 4 Jello, 5 Bottle, 6 Shampoo,7 Spray Bottle, 8 Vitalis Cereal, 9 Tomato Soup, 10 Schaumküsse, 11 Koala Candy, 12 Fruit Drink.

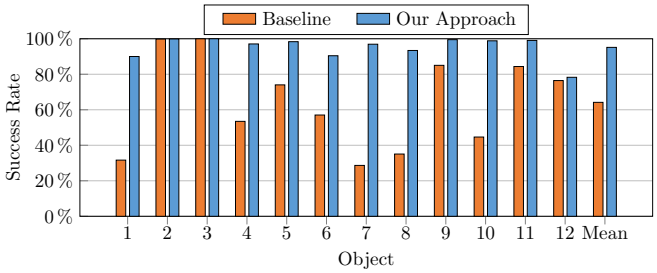

Fig. 5: Comparison of the proposed DNN approach (blue) against the baseline (orange). The DNN outperforms the non-deep learning baseline for complex-shaped objects. The difference is most significant for object 1, 7, and 8.

B. Benefit of the Deep Neural Network

In order to measure the added benefit of the deep neural network, we execute the evaluation pipeline for all 12 test objects. We compare the results of the baseline against the success rates of our approach in Fig. 5. In the following, we analyze the results for ball-shaped objects (2 and 3), cylindrical (5, 6, 9 and 11), boxes (4, 8, 10 and 12) and complex objects (1 and 7).

The baseline approach and the proposed DNN method can easily grasp ball-shaped objects (2 and 3). Hand-sized cylinders (5, 9 and 11) can be grasped by the baseline in most cases; however, the DNN can eliminate almost all grasp failures. The baseline drops to $50\,\%$ success rate for the shampoo (6). The shampoo bottle has an oval shape, resulting in over-sized GPIS estimates when observed from the front. In contrast, the DNN can raise the success rate to $90\,\%$.

The difference in success rate is most prominent for the unknown objects (1 and 7). This can be explained by the properties of the GPIS estimate based on the partial obser-

TABLE I: Grasp success probability depending on the number of touches during tactile exploration.

	Number of exploration actions					
Object	**0**	**1**	**2**	**3**	**4**	**5**
1	70 %	73 %	82 %	68 %	94 %	90 %
2	100 %	100 %	100 %	100 %	100 %	100 %
3	100 %	100 %	100 %	100 %	100 %	100 %
4	88 %	96 %	98 %	98 %	96 %	97 %
5	89 %	99 %	98 %	99 %	99 %	98 %
6	89 %	95 %	92 %	96 %	93 %	90 %
7	77 %	91 %	70 %	74 %	79 %	97 %
8	92 %	88 %	92 %	92 %	94 %	93 %
9	90 %	100 %	99 %	99 %	100 %	99 %
10	96 %	94 %	99 %	93 %	94 %	99 %
11	97 %	99 %	99 %	100 %	99 %	99 %
12	83 %	97 %	97 %	99 %	96 %	78 %
Mean	**89 %**	**94 %**	**94 %**	**93 %**	**95 %**	**95 %**

vations of the object. In these cases the GPIS estimate differs substantially from the real object, leading to many incorrect grasp hypotheses. The baseline approach scores the grasps based on the incorrect GPIS estimate, leading to a low grasp success probability. The proposed DNN approach filters out most of the incorrect hypotheses, therefore increasing the grasp success probability, as can be seen in Fig. 7.

In our experiment, we have shown that the proposed DNN approach performs better than or equal to the baseline for all tested objects. In some cases, the DNN outperforms the baseline significantly, reaching almost 100 % success rates. The baseline fails in 35 % of the cases and the proposed approach fails in 5 % of the tested cases, allowing for an average of 7-times fewer grasp failures. Exemplary successful grasp hypotheses generated by the DNN approach are displayed in Fig. 8.

C. Benefit of Tactile Exploration

During execution on the robot, acquiring tactile contact with the objects is a time-consuming endeavor that might move the objects resulting in loss of precision. However, adding more tactile contact points improves the GPIS estimate gradually [4]. Therefore, a trade-off between execution time and model completeness arises. To find an adequate amount of tactile exploration actions, we performed a separate evaluation, where the number of touches was fixed to a predefined amount for each evaluation run. Table I lists the success rates for the test set depending on the number of touches executed during tactile exploration.

The first added tactile contact point increases the success rate considerably while adding more observations hardly improves the success rate on average. Therefore, we opted to perform one exploration action during the validation on the humanoid robot ARMAR-6.

VII. Validation on ARMAR-6

In order to validate our approach, we execute the full pipeline on the humanoid robot ARMAR-6. The robot is located in front of a table with several unknown objects on top, see Fig. 10. The goal is to grasp an object from the table and lift it, to enable further manipulation, e.g. placing. We

Fig. 6: Object set used for validation. Objects from top to bottom and left to right: Aluminum Profile, Hammer, Multimeter, Screw Box, Power Drill, Cutter, Pliers, Spray Bottle.

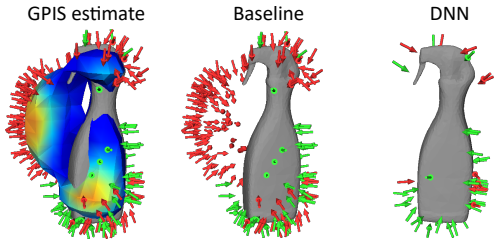

Fig. 7: Comparison of the baseline and the proposed DNN method: Successful grasps are displayed as green arrows, failed grasps are shown as red arrows. The GPIS estimate (colored surface, color denotes GPIS variance, blue: low, red: high) differs from the actual object. The baseline approach plans grasps based on the incorrect surface, leading to many failed grasps. The proposed DNN approach filters out most of the incorrect grasp hypotheses leading to a higher grasp success rate.

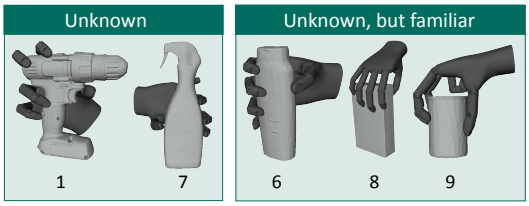

Fig. 8: Examples for successful grasps generated with the proposed approach on three familiar and the two unknown objects from our test set.

use a predefined region of interest in which the robot searches for an object. The object to be grasped lies in this region and should be grasped from the top. For visual perception, we use a Primesense RGB-D camera, located in the head of the robot. The robot grasps with its five-finger hand that is mounted at the end of the robot's 8-DOF arm. At the time of the experiment, the robot's hand did not provide tactile sensing or joint angle encoders. However, we can estimate tactile contacts by inferring contacts from the force torque sensor, located in the wrist. We estimate the contact to be at the fingertip of the middle finger, which can be computed since the hand is fully open during tactile exploration. In addition, the hand is underactuated, controlled by only two motors, one for the fingers and one for the thumb. The robot's hand is similar to the KIT prosthetic hand [49] in design, however larger to fit the size of the robot. Therefore, we

Fig. 9: Exemplary grasping results using the underactuated five-finger hand of ARMAR-6.

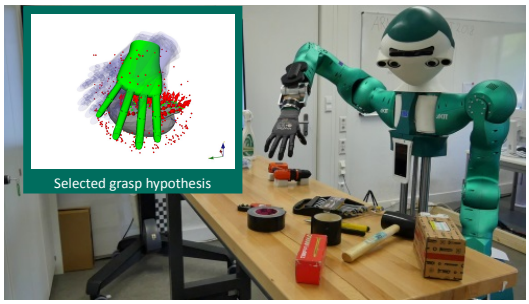

Fig. 10: Validation setup.

TABLE II: Results of the validation on the real robot.

Object	Mass	Lift success?	Attempts
Alum. Profile	1.2 kg	Yes	2
Hammer	0.8 kg	Yes	2
Multimeter	0.4 kg	Yes	1
Screw box	0.4 kg	Yes	2
Power drill	0.9 kg	Yes	2
Cutter	0.3 kg	Yes	2
Pliers	0.3 kg	Yes	1
Spray bottle	0.2 kg	Yes	1

chose to train and execute power grasps only. To enable power grasping from the top we had to raise most objects using small foam blocks. For validation, we used different objects, including the "Power Drill" and the "Spray Bottle". The set of eight different validation objects is shown in fig. 6.

Validation Results

We performed validation experiments with eight different objects, weighing up to 1.2 kg. The robot was able to lift all of the objects, while in some cases multiple attempts were necessary, see Table II. Successful grasps resulted in a firm enclosure of the object, see fig. 9. Failed grasping attempts resulted from the sliding of the fingers when the object could not be enclosed. We chose to raise the objects with small foam blocks, allowing the fingers to enclose the objects, if possible. We argue that these failure modes are not a fundamental limitation of the proposed approach but emerge from uncertainties in perception and insufficient friction between the fingers and the object, e.g. in case of the box, see fig. 9 on the right. Further examples for successful and failed grasps are shown in the accompanying video[1].

VIII. Conclusion

In this work, we presented a visuo-haptic grasping pipeline, leveraging deep learning to estimate grasp success probability. Visual and tactile information is fused with Gaussian Process Implicit Surfaces and grasp hypotheses are generated based on the estimated surface using a skeleton and a surface-based grasp planner. The grasp hypotheses are then scored with a deep neural network that has been trained in simulation. During the evaluation in simulation we achieved a grasp success rate of 95 % regarding force-closure of the hand, tested on 12 unseen objects, including non-trivial shapes like a power drill and a spray bottle. Our approach was able to transfer part-specific grasps from one object to another, since the neural network is presented with a local view of the perceived visual and tactile data, capturing only the relevant local features for grasp assessment. The successful transfer from simulation to the humanoid robot ARMAR-6 was validated in real-world experiments using eight unknown objects, where each object could be lifted successfully.

The used hand is underactuated and does not include joint angle encoders, therefore precision grasps are difficult to execute. Thus, the execution on the robot is limited to power grasps, where the objects are slightly raised using small spacer foam blocks. Currently, we are working on the grasp execution with underactuated hands by exploiting interactions with the environment to eliminate the need for these spacer blocks. Using stochastic gradient descent [50] could be applied in future work for grasp refinement.

[1] http://ottenhaus.de/simon/vhgrasping/

References

[1] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Trans. on Robotics*, vol. 30, no. 2, pp. 289–309, 2014.

[2] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour, "Grasping of unknown objects using deep convolutional neural networks based on depth images," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2018, pp. 6831–6838.

[3] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, p. 345, 2009.

[4] M. Bjorkman, Y. Bekiroglu, V. Hogman, and D. Kragic, "Enhancing visual perception of shape through tactile glances," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2013, pp. 3180–3186.

[5] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928.

[6] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal, and G. S. Sukhatme, "Interactive perception: Leveraging action in perception and perception in action," *IEEE Trans. on Robotics*, vol. 33, no. 6, pp. 1273–1291, 2017.

[7] A. F. Oliver Williams, "Gaussian process implicit surfaces," in *Gaussian Processes In Practice*, 2006.

[8] S. Dragiev, M. Toussaint, and M. Gienger, "Gaussian process implicit surfaces for shape estimation and grasping." in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2011, pp. 2845–2850.

[9] S. Caccamo, Y. Bekiroglu, C. H. Ek, and D. Kragic, "Active exploration using gaussian random fields and gaussian process implicit surfaces," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2016, pp. 582–589.

[10] J. Ilonen, J. Bohg, and V. Kyrki, "Fusing visual and tactile sensing for 3-d object reconstruction while grasping," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2013, pp. 3547–3554.

[11] A. Maldonado, H. Alvarez, and M. Beetz, "Improving robot manipulation through fingertip perception," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2012, pp. 2947–2954.

[12] J. Varley, D. Watkins-Valls, and P. K. Allen, "Multi-modal geometric learning for grasping and manipulation," *Computing Research Repository*, vol. abs/1803.07671, 2018.

[13] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1912–1920.

[14] X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, "Learning 6-dof grasping interaction via deep geometry-aware 3d representations," *arXiv preprint arXiv:1708.07303*, 2017.

[15] 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," *ArXiv e-prints*, 2018.

[16] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, vol. 11, no. 4, pp. 110–122, 2004.

[17] N. Vahrenkamp, M. Kröhnert, S. Ulbrich, T. Asfour, G. Metta, R. Dillmann, and G. Sandini, "Simox: A robotics toolbox for simulation, motion and grasp planning," in *Intelligent Autonomous Systems*. Springer, 2013, pp. 585–594.

[18] C. Ferrari and J. Canny, "Planning optimal grasps," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 1992, pp. 2290–2295.

[19] J. Weisz and P. K. Allen, "Pose error robust grasping from contact wrench space metrics," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2012, pp. 557–562.

[20] D. Kraft, N. Pugeault, E. BAŞESKI, M. POPOVIĆ, D. KRAGIĆ, S. Kalkan, F. Wörgötter, and N. Krüger, "Birth of the object: detection of objectness and extraction of object shape through object–action complexes," *International Journal of Humanoid Robotics*, vol. 5, no. 02, pp. 247–265, 2008.

[21] M. Popovic, G. Kootstra, J. A. Jørgensen, D. Kragic, and N. Krüger, "Grasping unknown objects using an early cognitive vision system for general scene understanding." in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, vol. 11, 2011, pp. 987–994.

[22] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.

[23] A. Morales, P. J. Sanz, A. P. Del Pobil, and A. H. Fagg, "Vision-based three-finger grasp synthesis constrained by hand geometry," *Robotics and Autonomus Systems*, vol. 54, no. 6, pp. 496–512, 2006.

[24] A. T. Miller, S. Knoop, H. I. Christensen, and P. K. Allen, "Automatic grasp planning using shape primitives," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, vol. 2, 2003, pp. 1824–1829.

[25] K. Huebner, S. Ruthotto, and D. Kragic, "Minimum volume bounding box decomposition for shape approximation in robot grasping," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2008, pp. 1628–1633.

[26] D. Schiebener, A. Schmidt, N. Vahrenkamp, and T. Asfour, "Heuristic 3d object shape completion based on symmetry and scene context," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2016, pp. 74–81.

[27] N. Vahrenkamp, E. Koch, M. Wächter, and T. Asfour, "Planning high-quality grasps using mean curvature object skeletons," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 911–918, 2018.

[28] K. Hsiao, S. Chitta, M. Ciocarlie, and E. G. Jones, "Contact-reactive grasping of objects with partial shape information," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2010, pp. 1228–1235.

[29] D. Schiebener, J. Schill, and T. Asfour, "Discovery, segmentation and reactive grasping of unknown objects," in *IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids)*, 2012, pp. 71–77.

[30] J. Kenney, T. Buckley, and O. Brock, "Interactive segmentation for manipulation in unstructured environments," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2009, pp. 1377–1382.

[31] G. Metta and P. Fitzpatrick, "Better vision through manipulation," *Adaptive Behavior*, vol. 11, no. 2, pp. 109–128, 2003.

[32] N. Bergström, M. Björkman, and D. Kragic, "Generating object hypotheses in natural scenes through human-robot interaction," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2011, pp. 827–833.

[33] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2015, pp. 1316–1322.

[34] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Computing Research Repository*, vol. abs/1301.3592, 2013.

[35] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," *Computing Research Repository*, vol. abs/1509.06825, 2015.

[36] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *Computing Research Repository*, vol. abs/1703.09312, 2017.

[37] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *Computing Research Repository*, vol. abs/1805.11085, 2018.

[38] J. Varley, J. Weisz, J. Weiss, and P. Allen, "Generating multi-fingered robotic grasps via deep learning," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 4415–4420.

[39] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," *Computing Research Repository*, vol. abs/1709.07857, 2017.

[40] J. Tobin, W. Zaremba, and P. Abbeel, "Domain randomization and generative models for robotic grasping," *Computing Research Repository*, vol. abs/1710.06425, 2017.

[41] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. A. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," *Computing Research Repository*, vol. abs/1803.09956, 2018.

[42] A. Rajeswaran, V. Kumar, A. Gupta, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *Computing Research Repository*, vol. abs/1709.10087, 2017.

[43] S. Ottenhaus, P. Weiner, L. Kaul, A. Tulbure, and T. Asfour, "Exploration and reconstruction of unknown objects using a novel normal and contact sensor," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2018, pp. 0–0.

[44] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[45] S. Ottenhaus, L. Kaul, N. Vahrenkamp, and T. Asfour, "Active tactile exploration based on cost-aware information gain maximization," *International Journal of Humanoid Robotics*, vol. 15, no. 01, p. 1850015, 2018.

[46] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[47] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models database: An object model database for object recognition, localization and manipulation in service robotics," *Int. J. of Robotics Research*, vol. 31, no. 8, pp. 927–934, 2012.

[48] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *Advanced Robotics (ICAR)*, 2015, pp. 510–517.

[49] P. Weiner, J. Starke, F. Hundhausen, J. Beil, and T. Asfour, "The kit prosthetic hand: Design and control," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3328–3334.

[50] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," *arXiv preprint arXiv:1905.10520*, 2019.