Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 232 (2024) 1961–1971

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Industry 4.0 and Smart Manufacturing

Simplifying Robot Grasping in Manufacturing with a Teaching Approach based on a Novel User Grasp Metric

Matteo Pantano[a,b,*], Vladislav Klass[a], Qiaoyue Yang[a], Akhil Sathuluri[c], Daniel Regulin[a], Lucas Janisch[a], Markus Zimmermann[c], Dongheui Lee[d,e]

[a]*Technology Department, Siemens Aktiengesellschaft, D-81739 Munich, Germany*
[b]*Human-Centered Assistive Robotics, Technical University of Munich, D-80333 Munich, Germany*
[c]*Laboratory for Product Development and Lightweight Design, Technical University of Munich, D-85748 Garching, Germany*
[d]*Autonomous Systems, Technische Universität Wien, AT-1040 Vienna, Austria*
[e]*Institute of Robotics and Mechatronics, German Aerospace Center, D-82234 Wessling, Germany*

Abstract

The manufacturing industry is undergoing rapid evolution, necessitating flexible and adaptable robots. However, configuring such machines requires technical experts, which are hard to find, especially for small and medium enterprises. Therefore, the process needs to be simplified by allowing non-experts to configure robots. During such configuration, one key aspect is the definition of objects' grasping poses. The literature proposes deep learning techniques to compute grasping poses automatically and facilitate the process. Nevertheless, practical implementation for inexperienced factory operators can be challenging, especially if task-specific knowledge and constraints should be considered. To overcome this barrier, we propose an approach that facilitates teaching such poses. Our method, employing a novel user grasp metric, combines the operator's initial grasp guess given by a 3D spatial device with a state-of-the-art deep learning algorithm, thus returning reliable grasping poses but simultaneously close to the operator's initial guess. We compare this approach against commercial grasping pose definition interfaces through a user test involving 28 participants and against state-of-the-art deep learning grasp estimators. The results demonstrate a significant improvement in system usability (+24%) and a reduced workload (-16%). Furthermore, our experiments reveal an increased grasp success rate when utilizing the user grasp metric, surpassing state-of-the-art deep learning grasping estimators.

© 2024 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 5th International Conference on Industry 4.0 and Smart Manufacturing

Keywords: operator empowerment; spatial interaction; robotic grasping; usability; small medium enterprises; cobots

* Corresponding author.
E-mail address: matteo.pantano@siemens.com

1877-0509 © 2024 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 5th International Conference on Industry 4.0 and Smart Manufacturing
10.1016/j.procs.2024.02.018

1. Introduction

Collaborative robots (cobots) pose to be an enabling technology for Small and Medium Enterprises (SMEs) as long through their simple-to-use user interfaces (UI) can enable non-experts to program robotic behaviors and improve the efficiency and productivity of the company [1]. However, lately, their adoption entered a market sobering phase, and the number of adopted cobots started to decrease dramatically [2, 3]. Previous research already identified that cobots face three barriers to adoption: interfaces, safety, and design [4]. However, in the design area, cobots are especially suffering when they need to deal with constant changes in production, which require the robot to adapt to changes. In this regard, the computation of grasping poses is one of the main pain points for robotics in High-Mix Low-Volume (HMLV) manufacturing due to high demands on robot capabilities in handling several geometries with low a priori object knowledge [5].

Nevertheless, robust grasping pose selection is complex, especially in industrial domains [6]. This problem was initially solved in research by calculating the grasp stability and equilibrium based on part knowledge and the End-Effector (EE) Tool Center Point (TCP) geometries. However, these approaches are often limited by the required a priori information [6]. To overcome these limitations, the research community has recently developed data-driven approaches based on Deep Learning (DL). The most prominent examples use Convolutional Neural Networks (CNN) to estimate the best grasping poses [7–9]. To do so, the algorithms process depth or point cloud data to estimate different grasping poses and their quality considering either suction or antipodal grasping [10, 11]. However, despite the promising results, these approaches are still bound to research, and a gap exists in how DL can be transitioned into the industrial domain [6]. Therefore, analytical methods or user selections are still the ones used in high-quality camera vision systems in the industrial domain (e.g., Keyence™) when defying grasping poses [12]. Therefore, industrial practitioners are often required to select the grasping pose based on previous robotics knowledge and task-specific information through point clicks inserted via the UI of the camera vision system. This can limit non-expert factory operators to change the process parameters of camera vision systems due to the required knowledge, thus limiting their problem-solving capability and reducing the appealing factor of collaborative robots [13].

To solve this issue and empower novices to configure high-quality camera vision systems with task-specific knowledge, this work presents a method to record and fine-tune an initial grasp guess given by an operator through a spatial interaction with a 3D device. More precisely, the contributions of this work are two-fold. First, a method to gather the initial grasp guess from an operator with a 3D spatial device and transform the position in object coordinates considering the limitations of the depth camera. Second, a novel user grasp metric (UGM) which can be used to bias state-of-the-art CNN to sample robust grasps within proximity to the initial grasp guess.

The remainder of this paper is organized as follows to present our contributions. Initially, the state-of-the-art is introduced in Sec. 2. Afterward, the methodology to sample an initial grasp guess and select a reliable pose via the UGM is described in Sec. 3. Next, experiments with virtual evaluations, physical evaluations with known objects, and user tests are presented in Sec. 4. Finally, discussions and conclusions of our approach are given in Sec. 5 and Sec. 6.

2. Related Works

2.1. Analytical robotic grasping

Analytical robotic grasping was the first to be introduced in the robotics domain to compute grasp poses. The approach is based on the kinematics and dynamics of the system; hence, the computation of grasp poses considers the object's stability during and after the grasp action [6]. A comprehensive review of analytical methods is available in [14]; however, for clarity, it is helpful to underline that analytical methods have the following pitfalls. First, they have a high modeling complexity; second, they bear some practical inconsistencies with restrictive assumptions; and finally, they often compute just the grasp stability and not the grasp location [6, 15]. Therefore, with the emergence of DL, data-centric methodologies have been employed to surmount the limitations associated with analytical techniques.

2.2. Learning-based robotic grasping

Learning-based approaches, also known as data-centric methods, utilize sensor data of the target object as input and train a machine learning algorithm to estimate a robotic grasp pose, thus removing the need for an explicit model of the underlying system. Learning-based approaches can be categorized into model-based and model-free. In model-based methods, like in [7, 16], prior knowledge of the object is available during the training stage (e.g., CAD models of target objects). On the contrary, model-free methods, like in [17, 18], assume that no prior knowledge is known. A detailed review of these methods is available in [6, 7, 19]. A noteworthy example of a model-based approach is Dex-Net [16]. Dex-Net samples grasp candidates in the image space and evaluates these candidates by a Grasp Quality Convolution Neural Network (GQ-CNN). Dex-Net 4.0 achieves grasp success rates with unknown objects of up to 95% [11], and it is considered a state-of-the-art approach [6]. Despite the high success rate of model-based approaches, model-free ones have recently emerged to address scenarios where a system model is unavailable. (e.g., robotic kitting) [20]. Often, these approaches rely on Fully Convolutional Networks (FCN) to estimate dense pixel-wise grasping poses, thus skipping the grasp sampling step [17, 21, 22]. In contrast to the model-based, model-free approaches are faster because FCNs have few parameters, and the input image size to the FCN is flexible. However, despite the improved accuracy with unknown objects, the performances are still lower compared to model-based approaches [6].

While recent advancements in machine learning have achieved impressive success rates across various applications, there remains scope for further improvement. One area that calls for attention is enhancing algorithms' capacity to integrate task semantics, boundaries, and environmental factors, which are essential prerequisites for real-world applications. Therefore, even if a grasp pose computed by the algorithm is reliable, unintended collisions are still possible during object placement or movement [6]. For the sake of clarity, the following example can be considered. A robotic system needs to pick up a knife but needs prior exposure to knives in its training dataset. In this situation, using an antipodal EE could result in the most optimal grasp location on the handle, ensuring a secure hold. However, if the robot is required to hand over the knife to an operator, this configuration could be potentially dangerous. Therefore, it would be advantageous for the operator to impose specific constraints on the robot, such as constraining the grasp region on the knife's blade, ensuring a safer handover motion.

To the best of the authors' knowledge, the only methods known in the state-of-the-art to mitigate these issues are the ones proposed by [23–25]. However, both approaches depend on training task-specific grasp pose estimation algorithms and acquiring additional task information through human demonstration or knowledge representation. Therefore, a gap exists on how to embed task-specific requirements given by an inexperienced operator, with grasps computed by already existing learning-based robotic grasping pose computation algorithms avoiding retraining for each task.

2.3. Grasping pose labeling

One way to embed task-specific information is to label grasp poses, a common task when creating datasets for machine learning algorithms like the Cornell Grasp Dataset (CGD) [26]. Grasping poses can be labeled through pixel-wise annotation or oriented bounding boxes via point-clicks, utilizing expert knowledge or simulations. However, inexperienced operators may find this method laborious and time-consuming, similar to other point-click labeling approaches [27, 28]. Additionally, task semantics may be lost if labeling is done through simulation without knowledge of the final application.

Other labeling approaches that explore spatial interaction at the work cell can be used to simplify the process. Two examples from the literature include [29], who proposed using a digital pen to specify grasp location on technical drawings printed on paper, and [30], who proposed using the location and direction of a pointed finger to extract object information but only for the object location and not the grasping pose. However, these approaches should have considered human inaccuracies or the development of DL approaches in robotic grasping. Thus, there is still a need for better spatial interactions that can enable task semantics integration while reducing labeling efforts.

To solve these issues, this work proposes a method to label objects using spatial interaction that robotic novices can use. The method leverages the good performances of data-driven approaches and constrains their selection to specific areas using a novel metric UGM.

3. Methodology

3.1. Problem statement

Humans are inaccurate regarding meticulous position definitions [31]. Hence, while defining grasp points, operators should be provided with tools that minimize errors while allowing them to influence the area where the grasp should occur based on their understanding of task-related environmental factors. In other words, if a user defines an initial grasp guess $\mathcal{P}, \in \mathbb{R}^3$ on a surface $\mathcal{S}$ of an object $\mathcal{O}$, $\mathcal{S} \in \mathbb{R}^3 \wedge \mathcal{S} \in \mathcal{O}$, where the grasp should be executed, and $\mathcal{P},$ does not represent a precise position of a grasp due to human inaccuracies. It is possible to define that a point $\mathcal{P} \in \mathbb{R}^3$ contained in an area of radius $\epsilon \in \mathbb{R}^3$ is an accurate and a reliable grasp point satisfying the task-related environmental factors. This can be then expressed with Eq. (1).

$$\mathcal{B}(\mathcal{P},, \epsilon) = \{\mathcal{P}, \in \mathbb{R}^3 \text{ t.c } |\mathcal{P} - \mathcal{P},| < \epsilon\} \tag{1}$$

Therefore, there is a need to define a metric that can find the point $\mathcal{P} \in \mathbb{R}^3$ in the neighborhood $\mathcal{B}(\mathcal{P},, \epsilon)$ which has the highest probability of grasp success. To propose this metric, first Sec. 3.2 describes how a point $\mathcal{P}, \in \mathbb{R}^3$ can be sampled from the user. Second, Sec. 3.3 describes how depth images for obtaining grasp quality are sampled. Finally, Sec. 3.4 describes the metric and how it is used to obtain final grasping poses fusing the initial grasp guess and a grasp quality given by a DL algorithm.

3.2. Spatial point labeling via user input

To assist the user in defining an initial grasp guess $\mathcal{P},$, a cost-effective 3D spatial device is selected. This device employs photodiode sensors and several infrared emitting stations, making it an affordable alternative to camera-based systems [32]. Specifically, the Tracepen™ pen from Wandelbots was selected. Using such a device, the projection of point coordinates onto multiple camera images is done effortlessly and without relabeling, as evidenced in [28, 33]. However, the device coordinates are expressed in reference to the emitting stations [32]. Therefore, homogeneous transformations are necessary to obtain its coordinates in camera or robot frames, as shown in Fig. 2. To solve this problem, referring to the figure nomenclature, the transformation between the robot base and emitting station ${}^{t}T_{r}$ is calculated following the approach of [34]. Subsequently, the 3D point is projected onto the 2D images captured by the camera through Eq. (2).

$$\begin{bmatrix} u' \\ v' \\ w' \end{bmatrix} = K\, {}^{r}T_{c}\, {}^{t}T_{r}\, P_{i}^{t}, \ \forall i \tag{2}$$

Where u, v are pixel coordinates of a point i which belongs to the surface $\mathcal{S}$ of an object $\mathcal{O}$, $u = u'/w'$ and $v = v'/w'$ with w' as the camera scaling factor. K is the intrinsic camera matrix, ${}^{r}T_{c}$ is the camera extrinsic matrix calculated as in [35] and P_{i}^{t} is the initial grasp guess in tracker coordinates.

3.3. Creation of depth images and sampling of grasp pose candidates

Grasp pose candidates must be sampled after obtaining the initial grasp guess in camera coordinates. For this task, we propose to use Dex-Net 4.0 [11] as long it can predict grasping poses and their quality, select which is the best gripper type between antipodal and suction, and it is known to scale well with unknown objects [18]. However, the literature reports that Dex-Net 4.0 best performs when the camera is in the same position as in the training dataset [11, 18]. To overcome the challenge of aligning the camera setup with Dex-Net 4.0, we propose an alternative approach where depth images for sampling grasp pose candidates are obtained through a synthetic pipeline. In this method, first synthetic depth images using a Blender pipeline [36] are generated utilizing the CAD model of the object with several virtual depth cameras positioned in similar means as in the Dex-Net 4.0 training set. Second, the depth images are evaluated by Dex-Net 4.0 to compute a collection of potential virtual grasp poses and their quality.

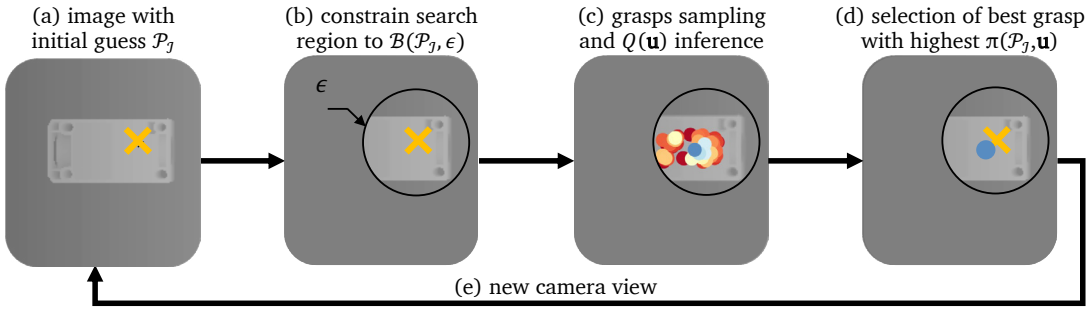


Fig. 1. Qualitative representation of the sampling process using Dex-Net 4.0. First a synthetical camera image is taken alongside its projection of the initial grasp guess (a). Second, for speeding the computation time, the image is masked to the search region close to the user input $\mathcal{B}(\mathcal{P}_\prime, \epsilon)$ (b). Third, Dex-Net 4.0 is used to sample grasps and computed their raw quality $Q(\mathbf{u})$ (c). Fourth, the grasp u with highest $\pi(\mathcal{P}_\prime, \mathbf{u})$ is selected and saved (d). Finally, the process is reiterated for multiple camera views to ensure the best $\pi_f(\mathcal{P}_\prime, \mathbf{u})$ in the boundary $\mathcal{B}(\mathcal{P}_\prime, \epsilon)$(e).

3.4. User grasp metric

With the user's initial grasp guess and a set of potential virtual grasp poses along with their corresponding quality available, the subsequent objective is to combine this information and derive a reliable final grasp pose. This is accomplished by evaluating the virtual grasp poses based on their quality and proximity to the user's initial grasp guess. To achieve this goal, a UGM is used. We propose the UGM as a geometric average between the sampled grasp quality $Q(\mathbf{u})$, given by Dex-Net, and the inverse of the normalized distance from the sampled grasp to the user grasp guess. This selection is driven by the need to maximize the grasp quality while ensuring the vicinity of the area marked by the user. This can be mathematically formulated with Eq. (3).

$$\pi(\mathcal{P}_\prime, \mathbf{u}) = \sqrt{Q(\mathbf{u})\left(1 - \frac{d(\mathbf{u}, \mathcal{P}_\prime)}{\epsilon}\right)} \tag{3}$$

Where $\pi(\mathcal{P}_\prime, \mathbf{u})$ is the UGM, $\pi(\mathcal{P}_\prime, \mathbf{u}) \in \mathbb{R}$, $\mathbf{u}$ is a grasp pose according to [16], $\mathbf{u} = (p, o) \in \mathbb{R}^3 \times \mathcal{S}^1$, $Q(\mathbf{u})$ is the grasp quality according to [16] $Q(\mathbf{u}) \in \mathbb{R}$, $\mathcal{P}_\prime$ is the initial grasp guess in 3D space $\mathcal{P}_\prime \in \mathbb{R}^3$, $d(\mathbf{u}, \mathcal{P}_\prime)$ is the euclidean distance in meters between the initial grasp guess and the sampled grasp $d(\mathbf{u}, \mathcal{P}_\prime) \in \mathbb{R}$. Moreover, ϵ is the radius in meters of the largest object size the EE mounted on the robot can grasp $\epsilon \in \mathbb{R}$.

By utilizing the UGM and selecting grasp poses $\mathbf{u}$ with the highest $\pi(\mathcal{P}_\prime, \mathbf{u})$, we enable the operator to influence grasp selection while ensuring a high level of reliability. Let us consider an example: Suppose a user identifies a corner of an object as the initial grasp guess. Attempting to grasp directly at that point would be impractical. In other words, if a sampled grasp pose $\mathbf{u}$ is near the initial guess, the corresponding $\pi(\mathcal{P}_\prime, \mathbf{u})$ value should be low. The UGM achieves this by multiplying the sampled grasp quality $Q(\mathbf{u})$ (which tends to be low for corner points) by the inverse of the distance to the initial grasp guess. However, as $Q(\mathbf{u})$ improves on a flat surface near the object's corner, the UGM assigns higher values to $\mathbf{u}$ closer to the initial grasp guess. Consequently, the system can select the closest and most reliable grasp from the available set.

After having identified the optimal and nearest grasp, the next step is to convert the virtual grasp from the depth image in the virtual domain to the corresponding representation in the real image; for doing this, the UGM is accompanied by a novel pipeline, Fig. 1 shows the main steps of the procedure. Initially, the initial grasp guess is projected in the synthetic depth image using the method described in Sec. 3.2 see Fig. 1a. Afterward, a region of boundary $\mathcal{B}(\mathcal{P}_\prime, \epsilon)$ is defined in the vicinity of the initial grasp guess with size equal to maximum graspable object size, see Fig. 1b. Third, possible grasps in the $\mathcal{B}(\mathcal{P}_\prime, \epsilon)$ are sampled using the Dex-Net analytical grasping poses sampler, see Fig. 1c. Fourth, UGM is employed for selecting the grasp with the highest probability of success and closest to the initial grasp guess, i.e., the grasp with the highest UGM, see Fig. 1d. Finally, the process is iterated for multiple views, generated with the synthetic pipeline described in Sec. 3.3, until the best grasp pose is selected, see Fig. 1e.

Once the best $\mathbf{u}$ is found, the synthetically generated grasp pose is converted back to the real world via the transformations shown in Fig. 3. These transformations work on two main assumptions. On the one hand, the object's CAD is placed in the origin of the synthetic world, and the synthetic depth camera is placed at a known distance to the

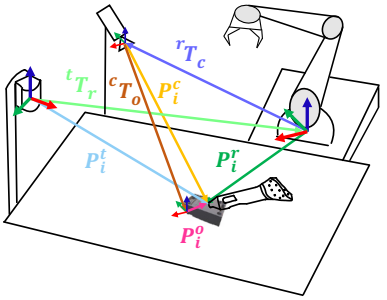


Fig. 2. Illustration of the different coordinate frames involved in the definition of a grasp guess. A point marked by the motion-tracked pen (P_i^t) is transformed to a point in the image coordinate frame (P_i^c) through ${}^{r}T_{c}$ and ${}^{t}T_{r}$. Afterward, the point can be transformed in the object coordinate system (P_i^o) through ${}^{c}T_{o}$ to ensure that conventional bin-picking systems can use it.

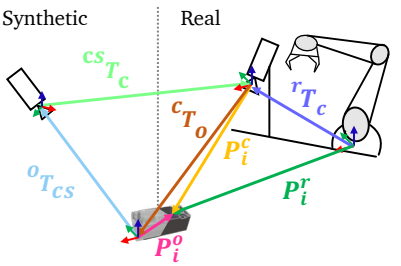


Fig. 3. Transformations involved for the generation of a synthetic depth image knowing the object digital model, its location in the real world (${}^{c}T_{o}$), and its location in the synthetic world (${}^{o}T_{cs}$).

object (${}^{o}T_{cs}$). On the other hand, the position of the real object is known in the camera mounted on the work cell (${}^{c}T_{o}$). Through these, the transformation between the robot and the synthetic camera (${}^{cs}T_{c}$) can be calculated, thus allowing the transformation of synthetically generated grasp positions in the robot coordinate frame through Eq. (4).

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = K^{-1}\,{}^{r}T_{c}\,{}^{c}T_{o}\,{}^{o}T_{cs} \begin{bmatrix} u' \\ v' \\ w' \\ 1 \end{bmatrix} \tag{4}$$

4. Experiments

To validate the methodology and understand the impact of using this approach, two sets of experiments in a robotic cell with a custom EE [37, 38] were conducted. First, a user study for evaluating the operators' perceived usability and workload as described in Sec. 4.1. Second, virtual and physical experiments to evaluate the quality of grasps with the proposed UGM as outlined in Sec. 4.2.

4.1. Usability and effort

For testing both the perceived usability and the workload, between subjects user tests were conducted. In the experiment, the subjects had to perform labeling with four different systems. More precisely, our method was compared against industrial grasp labeling interfaces from Roboception™, Photoneo™ and Mech-Mind™. These three were selected due to the different price ranges to better represent the market, and such are usually graphical user interfaces (GUI), which require several point labels to define grasp points [39]. Therefore, initially, the users were introduced to the systems and had time to familiarize themselves with them by looking at a walk-through video showing the steps to perform. Afterward, the participant had to perform the task of grasp labeling with a suction EE for a defined industrial object of dimensions 75 mm x 30 mm x 40 mm; see Fig. 7 for a picture of the object. Finally, system usability scale (SUS) [40], NASA-TLX [41], and labeling time were recorded. For the sake of clarity, the experiment design and the experimental set-up are shown in Fig. 4 and Fig. 5.

The study involved 28 participants, who were evenly distributed among the four systems. These users had a medium experience in grasp labeling, as determined by their self-assessment on a scale from low (never used grasp labeling) to high (use grasp labeling often). The average age of the participants was M = 27.3 yrs, SD = 3.40.

All participants completed the test, yielding the following results. Regarding usability and workload, our approach demonstrated statistically significant enhancements in the SUS and NASA-TLX measures, specifically when compared to Mech-Mind™. More precisely, a Mann-Whitney-U-test was applied as long the preconditions of the t-test did

not hold, and it reported $p < 0.05$ (CI = 95%) for both the questionnaires as shown in Fig. 6. Unfortunately, no statistically significant differences were observed when comparing our method to Roboception™ and Photoneo™. Hence, our method can be considered comparable to these systems regarding usability and workload. However, the results differed in terms of the required time. The test reported M = 11.00s, SD = 3.31, M = 465s, SD = 135.42, M = 792.14s, SD = 193.64, and M = 1335.29s, SD = 286.71 for ours, Roboception™, Photoneo™ and Mech-Mind™ respectively. This data also reported a $p < 0.05$ (CI = 95%) with the Mann-Whitney-U-test; therefore, it is possible to conclude that our method provides a faster approach for defining a grasp.

4.2. Grasp quality

Two sets of additional experiments were conducted to validate the grasp quality prediction using our method. The first set involved a virtual evaluation, comparing the grasp quality calculated using Eq. (3) between raw Dex-Net with an object mask and our approach. The second set involved a physical evaluation, assessing the same metrics and the grasp success rate.

In the virtual experiments, 19 objects were chosen from various industrial datasets [12, 37, 42] see Fig. 7 for a list of things. A random uniform distribution was employed in the synthetic rendering tool described in Sec. 3.3 to generate a simulated initial grasp guess within the object area. Then, the UGM and the distance from the simulated initial grasp guess to the final grasp were calculated using the generated depth images. The results of these computations are depicted in Fig. 8. Regarding the suction EE, our approach demonstrated a statistically significant higher UGM (M=34.06%, SD=24.00) compared to the raw Dex-Net (M=29.16%, SD=27.15) when a Mann-Whitney-U test

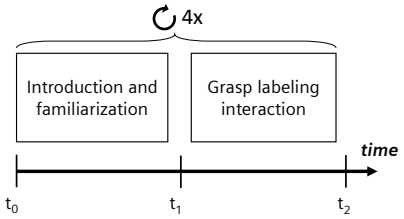


Fig. 4. Experimental procedure. Following an initial introduction and familiarization with the grasp labeling between t_0 and t_1, every participant in the study proceeded to complete the grasp labeling task from t_1 to t_2. The time taken for labeling, calculated as (t_2 - t_1), was documented. Subsequently, at time t_2, the user assessed the interaction's usability and workload by completing the provided questionnaires.

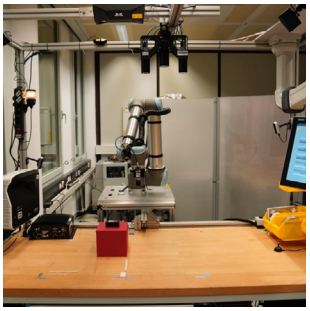
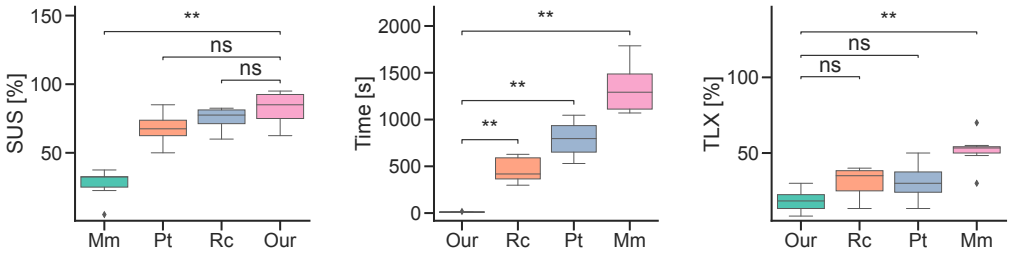


Fig. 5. Robot work-cell set-up. The cell was equipped with industrial camera systems, and the test object was placed under them. The user had to perform the grasp labeling using either the user interfaces of the camera systems or the method proposed by this work.



(a) SUS (b) TLX (c) Time

Fig. 6. Results of the user test of conventional grasp labeling versus our approach. The statistical significance of the results is displayed with asterisks. If no statistical significance is found, *ns* is displayed, or no statistical marking is shown. *Mm* stands for Mech-Mind™ Mech-Eye Pro S 1000M, *Pt* stands for Photoneo™ Phoxi M, and *Rc* stands for Roboception™ rc_viscore, *Our* is the method proposed in this paper.

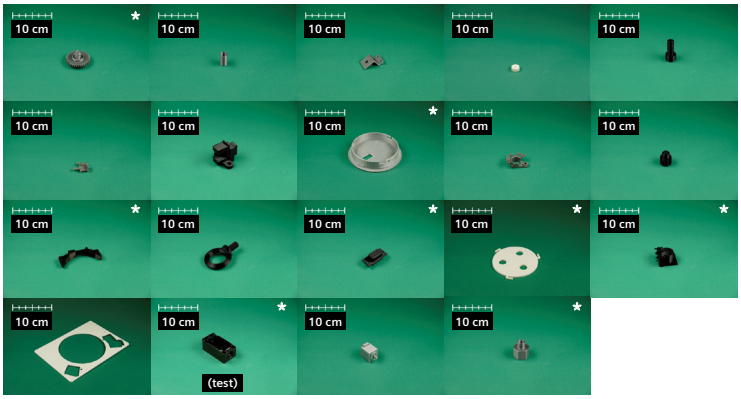
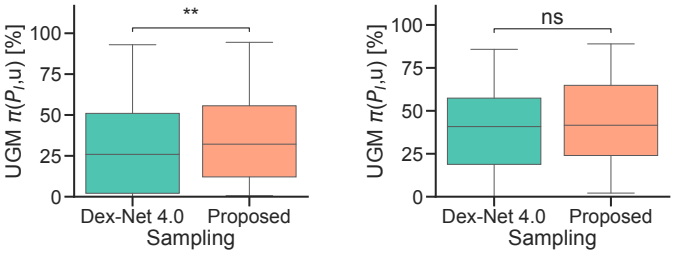


Fig. 7. Industrial components used for the evaluations. 19 objects from various industrial datasets were selected for the virtual evaluation. However, a subset of them (marked with an asterisk) was utilized for the physical evaluation. Within this subset, the object used for the user test is shown (marked with test).



(a) suction (b) antipodal

Fig. 8. Results of the virtual evaluations. To the left, the suction user grasp metric, and to the right, the antipodal one. The statistical significance of the results is displayed with asterisks. If no statistical significance is found, *ns* is displayed, or no statistical marking is shown. The proposed method shows statistical significance in selecting grasps with higher UGM in the case of a suction EE.

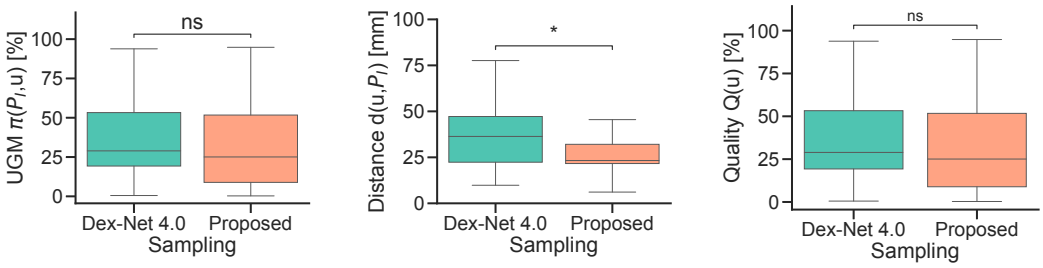
was conducted $p < 0.05$ (CI= 95%), preconditions for the t-test did not hold. However, for an antipodal EE, the distributions of raw Dex-Net and our method showed no statistical difference.

In the physical experiments, a subset of 8 objects from the previous set was used due to limitations given by the suction EE, mainly on the requirements of a large enough flat area for executing the pick. The evaluation results are shown in Fig. 9. When our method is compared against raw Dex-Net, the UGM does not decrease significantly $p > 0.05$ (CI = 95%). Still, the distance between the sampled grasp center and the initial grasp guess reduces significantly $p < 0.05$ (CI = 95%) when a Mann-Whitney-U-test is applied; preconditions for the t-test did not hold. Therefore, our approach can select grasp positions closer to the initial grasp guess. Moreover, in the same set of experiments, the grasp success rate resulted in 83.3% for Dex-Net and 87.5% for our approach.

5. Discussions and outlook

5.1. User impact

The user test showed that our based on the 3D pointing device and UGM can improve both the usability and the workload. More precisely, considering that the SUS scored an average of 82.25%, our method can be accepted by a broad user pool [43] if employed as in this scenario. Additionally, the method achieved an average workload of 19.17%, which can be considered low. Last, the proposed method allows a much faster grasp definition compared to



(a) user grasp metric (b) distance to initial grasp guess (c) raw grasp quality

Fig. 9. Outcomes of the physical evaluation for the suction EE. To the left is the UGM; in the center is the distance to the initial grasp guess; to the right is the quality given by Dex-Net 4.0. The statistical significance of the results is displayed with asterisks. If no statistical significance is found, *ns* is displayed, or no statistical marking is shown. The proposed method is able to select grasps closer to the user's initial grasp guess, maintaining the same quality as the raw Dex-Net 4.0.

existing industrial solutions. Considering these findings, it is possible to view our method as an excellent way to teach grasping poses, especially for novice users. Therefore, if SMEs adopt this method, higher acceptance compared to existing systems can be ensured. This could simplify the integration of vision-based systems in collaborative robotics applications, thus enlarging the set of applications that SMEs can address using cobots, answering some of their needs for HMLV manufacturing [5].

5.2. Grasp reliability

The virtual evaluations for the suction EE showed that by using the proposed UGM it is possible to constraint the selection of grasping poses in the vicinity of the initial grasp guess as seen in Fig. 8a. This has been reflected in the lower distance of the grasp pose from the initial grasp guess in the suction EE physical evaluation as seen in Fig. 9b. Unfortunately, the UGM reported no statistical difference. To further investigate this point the raw grasp quality $Q(\mathbf{u})$ is compared and the results are shown in Fig. 9c. From this result, which was not found to be statistically significant, it is then possible to understand that our method can force to select grasps close to the initial grasp guess while maintaining the same average raw grasp quality of Dex-Net and probably more tests are needed on the physical evaluation to find statistical difference. Therefore, the method, considering its high usability, can help non-experts in SMEs to define robust grasp positions which satisfy the requirements at the shop-floor when using suction grippers thus reducing the adoption barriers in industrial domains [6]. However, further investigations will be needed to understand how this method behaves in case of antipodal grippers. Therefore, the authors encourage readers to investigate the proposed UGM $\pi(\mathcal{P}_\prime, \mathbf{u})$ in different scenarios with the code and supplementary material made available on GitHub[1].

6. Conclusions

This study introduces a technique to enhance advanced learning-based grasp estimation approaches, such as Dex-Net, by incorporating human preferences in an intuitive way that allows inexperienced users to teach object-grasping poses. This is accomplished through a novel user grasp metric and a spatial teaching method. To understand the benefit of this approach, we validated the method with virtual evaluations, physical evaluations, and user tests. The user tests showed that usability and workload are improved compared to commercial grasp teaching methods. Moreover, the virtual assessments show that the proposed user grasp metric can select grasp poses close to the initial grasp guess given by the operator. Last, the physical evaluation showed the same grasp quality as Dex-Net. Therefore, using the user grasp metric and spatial inputs can be beneficial to obtain robust grasp poses close to the user's initial guess while improving user experience. Unfortunately, the proposed method assumes that 6D pose detection is available,

[1] https://github.com/matteopantano/Dex-Net-userInput

and the physical evaluation was performed with only eight objects. Therefore, with this work, we show that incorporating human preferences via spatial interactions can benefit the user experience as long it can enable non-experts to interact with vision systems, thus addressing some adoption barriers of collaborative robotics in Small and Medium Enterprises. However, further research is necessary for testing the approach with other end effectors like antipodal ones. Therefore, readers are invited to try our method with the available code and supplementary material.

References

[1] M. T. Ballestar, Á. Díaz-Chao, J. Sainz, J. Torrent-Sellens, Knowledge, robots and productivity in SMEs: Explaining the second digital wave, Journal of Business Research 108 (2020) 119–131. doi:10.1016/j.jbusres.2019.11.017.

[2] T. Kopp, A. Schäfer, S. Kinkel, Kollaborierende oder kollaborationsfähige Roboter? Welche Rolle spielt die Mensch-Roboter-Kollaboration in der Praxis?, Industrie 4.0 Management 2020 (2) (2020) 19–23. doi:10.30844/I40M_20-2_S19-23.

[3] International federation of robotics (Ed.), World Robotics 2021: Industrial Robots, VDMA Services GmbH, Franckfurt, 2021.

[4] V. Villani, F. Pini, F. Leali, C. Secchi, Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications, Mechatronics 55 (2018) 248–266. doi:10.1016/j.mechatronics.2018.02.009.

[5] C. Schlette, A. G. Buch, F. Hagelskjær, I. Iturrate, D. Kraft, A. Kramberger, A. P. Lindvig, S. Mathiesen, H. G. Petersen, M. H. Rasmussen, T. R. Savarimuthu, C. Sloth, L. C. Sørensen, T. N. Thulesen, Towards robot cell matrices for agile production – SDU Robotics' assembly cell at the WRC 2018, Advanced Robotics (2019) 1–17doi:10.1080/01691864.2019.1686422.

[6] J. P. C. Souza, L. F. Rocha, P. M. Oliveira, A. P. Moreira, J. Boaventura-Cunha, Robotic grasping: From wrench space heuristics to deep learning policies, Robotics and Computer-Integrated Manufacturing 71 (2021) 102176. doi:10.1016/j.rcim.2021.102176.

[7] M. Fujita, Y. Domae, R. Kawanishi, G. A. G. Ricardez, K. Kato, K. Shiratsuchi, R. Haraguchi, R. Araki, H. Fujiyoshi, S. Akizuki, M. Hashimoto, A. Causo, A. Noda, H. Okuda, T. Ogasawara, Bin-picking Robot using a Multi-gripper Switching Strategy based on Object Sparseness, in: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), IEEE, Vancouver, BC, Canada, 2019, pp. 1540–1547. doi:10.1109/COASE.2019.8842977.

[8] M. Fujita, Y. Domae, A. Noda, G. A. Garcia Ricardez, T. Nagatani, A. Zeng, S. Song, A. Rodriguez, A. Causo, I. M. Chen, T. Ogasawara, What are the important technologies for bin picking? Technology analysis of robots in competitions based on a set of performance metrics, Advanced Robotics (2019) 1–15doi:10.1080/01691864.2019.1698463.

[9] A. ten Pas, M. Gualtieri, K. Saenko, R. Platt, Grasp Pose Detection in Point Clouds, The International Journal of Robotics Research 36 (13-14) (2017) 1455–1473. doi:10.1177/0278364917735594.

[10] E. Solowjow, I. Ugalde, Y. Shahapurkar, J. Aparicio, J. Mahler, V. Satish, K. Goldberg, H. Claussen, Industrial Robot Grasping with Deep Learning using a Programmable Logic Controller (PLC), in: 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), IEEE, Hong Kong, Hong Kong, 2020, pp. 97–103. doi:10.1109/CASE48305.2020.9216902.

[11] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, K. Goldberg, Learning ambidextrous robot grasping policies, Science Robotics 4 (26) (2019) eaau4984. doi:10.1126/scirobotics.aau4984.

[12] J. Yang, Y. Gao, D. Li, S. L. Waslander, ROBI: A Multi-View Dataset for Reflective Objects in Robotic Bin-Picking, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Prague, Czech Republic, 2021, pp. 9788–9795. doi:10.1109/IROS51168.2021.9635871.

[13] P. Heikkilä, S. Aromaa, H. Lammi, T. Kuula, Framework of Future Industrial Worker Characteristics, in: 9th International Conference on Human Interaction and Emerging Technologies - Artificial Intelligence and Future Applications, 2023. doi:10.54941/ahfe1002927.

[14] R. M. Murray, Z. Li, S. S. Sastry, A Mathematical Introduction to Robotic Manipulation, 1st Edition, CRC Press, 2017. doi:10.1201/9781315136370.

[15] M. A. Roa, R. Suárez, Grasp quality measures: Review and performance, Autonomous Robots 38 (1) (2015) 65–88. doi:10.1007/s10514-014-9402-3.

[16] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, K. Goldberg, Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics (Aug. 2017). arXiv:arXiv:1703.09312.

[17] D. Morrison, P. Corke, J. Leitner, Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach (May 2018). arXiv:arXiv:1804.05172.

[18] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. Chavan Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, A. Rodriguez, Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching, The International Journal of Robotics Research 41 (7) (2022) 690–705. doi:10.1177/0278364919868017.

[19] K. Kleeberger, R. Bormann, W. Kraus, M. F. Huber, A Survey on Learning-Based Robotic Grasping, Current Robotics Reports 1 (4) (2020) 239–249. doi:10.1007/s43154-020-00021-6.

[20] Z. Kootbally, C. Schlenoff, B. Antonishek, F. Proctor, T. Kramer, W. Harrison, A. Downs, S. Gupta, Enabling robot agility in manufacturing kitting applications, Integrated Computer-Aided Engineering 25 (2) (2018) 193–212. doi:10.3233/ICA-180566.

[21] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morena, P. Qu Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, A. Rodriguez, Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Brisbane, QLD, 2018, pp. 3750–3757. doi:10.1109/ICRA.2018.8461044.

[22] V. Satish, J. Mahler, K. Goldberg, On-Policy Dataset Synthesis for Learning Robot Grasping Policies Using Fully Convolutional Deep Networks, IEEE Robotics and Automation Letters 4 (2) (2019) 1357–1364. doi:10.1109/LRA.2019.2895878.
[23] M. Sun, Y. Gao, GATER: Learning Grasp-Action-Target Embeddings and Relations for Task-Specific Grasping, IEEE Robotics and Automation Letters 7 (1) (2022) 618–625. doi:10.1109/LRA.2021.3131378.
[24] Y. Liu, K. Qian, X. Xu, B. Zhou, F. Fang, Grasp Pose Learning from Human Demonstration with Task Constraints, Journal of Intelligent & Robotic Systems 105 (2) (2022) 37. doi:10.1007/s10846-022-01650-z.
[25] Z. Zhang, Z. Jiao, W. Wang, Y. Zhu, S.-C. Zhu, H. Liu, Understanding Physical Effects for Effective Tool-use (Jun. 2022). arXiv:arXiv:2206.14998.
[26] Yun Jiang, S. Moseson, A. Saxena, Efficient grasping from RGBD images: Learning using a new rectangle representation, in: 2011 IEEE International Conference on Robotics and Automation, IEEE, Shanghai, China, 2011, pp. 3304–3311. doi:10.1109/ICRA.2011.5980145.
[27] C. Sager, C. Janiesch, P. Zschech, A survey of image labelling for computer vision applications, Journal of Business Analytics 4 (2) (2021) 91–110. doi:10.1080/2573234X.2021.1908861.
[28] A. Caporali, M. Pantano, L. Janisch, D. Regulin, G. Palli, D. Lee, A Weakly Supervised Semi-Automatic Image Labeling Approach for Deformable Linear Objects, IEEE Robotics and Automation Letters 8 (2) (2023) 1013–1020. doi:10.1109/LRA.2023.3234799.
[29] J. Pires, T. Godinho, R. Araújo, Using digital pens to program welding tasks, Industrial Robot: An International Journal 34 (6) (2007) 476–486. doi:10.1108/01439910710832075.
[30] S. van Delden, M. Umrysh, C. Rosario, G. Hess, Pick-and-place application development using voice and visual commands, Industrial Robot: An International Journal 39 (6) (2012) 592–600. doi:10.1108/01439911211268796.
[31] D. A. Norman, The Design of Everyday Things, revised and expanded edition Edition, Basic Books, New York, New York, 2013.
[32] A. K. T. Ng, L. K. Y. Chan, H. Y. K. Lau, A low-cost lighthouse-based virtual reality head tracking system, in: 2017 International Conference on 3D Immersion (IC3D), IEEE, Brussels, 2017, pp. 1–5. doi:10.1109/IC3D.2017.8251910.
[33] D. Gregorio, A. Tonioni, G. Palli, L. Di Stefano, Semiautomatic Labeling for Deep Learning in Robotics, IEEE Transactions on Automation Science and Engineering 17 (2) (2020) 611–620. doi:10.1109/TASE.2019.2938316.
[34] W. Zhang, X. Ma, L. Cui, Q. Chen, 3 Points Calibration Method of Part Coordinates for Arc Welding Robot, in: C. Xiong, Y. Huang, Y. Xiong, H. Liu (Eds.), Intelligent Robotics and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 216–224.
[35] V. Lepetit, F. Moreno-Noguer, P. Fua, EPnP: An Accurate O(n) Solution to the PnP Problem, International Journal of Computer Vision 81 (2) (2009) 155–166. doi:10.1007/s11263-008-0152-6.
[36] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, H. Katam, BlenderProc (2019). doi:10.48550/ARXIV.1911.01911.
[37] M. Pantano, A. Blumberg, D. Regulin, T. Hauser, J. Saenz, D. Lee, Design of a Collaborative Modular End Effector Considering Human Values and Safety Requirements for Industrial Use Cases, in: G. Palli, C. Melchiorri, R. Meattini (Eds.), Human-Friendly Robotics 2021, Vol. 23 of Springer Proceedings in Advanced Robotics, Springer International Publishing, Cham, 2022, pp. 45–60. doi:10.1007/978-3-030-96359-0_4.
[38] M. Pantano, Y. Pavlovskyi, E. Schulenburg, K. Traganos, S. Ahmadi, D. Regulin, D. Lee, J. Saenz, Novel Approach using Risk Analysis Component to Continuously Update Collaborative Robotics Applications in the Smart, Connected Factory Model, Applied Sciences 12 (11) (2022) 5639. doi:10.3390/app12115639.
[39] F. Kaynar, S. Rajagopalan, S. Zhou, E. Steinbach, Remote Task-oriented Grasp Area Teaching By Non-Experts through Interactive Segmentation and Few-Shot Learning (Mar. 2023). arXiv:arXiv:2303.10195.
[40] J. Brooke, SUS - A Quick and Dirty Usability Scale, CRC Press, 1996.
[41] S. G. Hart, L. E. Staveland, Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, in: Advances in Psychology, Vol. 52, Elsevier, 1988, pp. 139–183. doi:10.1016/S0166-4115(08)62386-9.
[42] K. Kleeberger, C. Landgraf, M. F. Huber, Large-scale 6D Object Pose Estimation Dataset for Industrial Bin-Picking (Dec. 2019). arXiv:arXiv:1912.12125.
[43] A. Bangor, P. T. Kortum, J. T. Miller, An Empirical Evaluation of the System Usability Scale, International Journal of Human-Computer Interaction 24 (6) (2008) 574–594. doi:10.1080/10447310802205776.