A Knowledge-augmented Concept for Programming by Demonstration based on Hand-Object Actions

Junsheng Ding*, Alexander Perzylo*, Liangwei Zhou*

***Abstract*—In this paper, we introduce work towards a knowledge-augmented concept for Programming by Demonstration (PbD) for industrial assembly processes. The hand-object actions in various assembly tasks are abstracted with the semantic model of grasp types based on the Web Ontology Language (OWL) containing task context from the grasp taxonomy stored in a Knowledge Base (KB). A Long Short-Term Memory (LSTM) network is used for recognizing the grasp types from the hand skeleton in human demonstrations. The semantic process description enables the generation of human assembly processes with a sequence of hand-object actions and their conversion to a robot-suitable process variant, which can be executed based on the skill description of a workcell. We showcase the concept with different assembly steps for a ball bearing as an example of matching between human grasp types and robotic skills.**

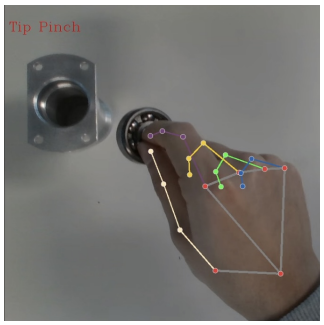
I. Introduction

Small and medium-sized enterprises (SMEs) focus on producing individualized or small-batch products that require frequent adaptation to the changing demands of customers and markets. However, the current industrial robots are primarily designed for large-scale production and require expert knowledge in robot programming. They are not well-suited for productions in SMEs, where the robots need to be frequently re-programmed by the product change. Thus, the development of more user-friendly robot programming methods that are easier to learn and more intuitive is necessary to facilitate the adoption of robot-based automation in SMEs.

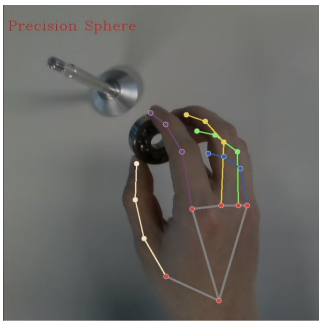
In this work, we extend our intuitive robot programming paradigm [1] to facilitate robot programming in SMEs with a *Programming by Demonstration* (PbD) framework for mechanical assembly processes, which uses the semantic model of hand-object actions for a more intuitive robot programming on the task level based on an user's demonstration of the task.

PbD enables robots to learn new skills from human demonstration with passive observation, which has been investigated in manufacturing for different tasks such as pick-and-place, peg-in-hole, and assembly operations [2]. The operator can perform the demonstration in high degree of freedom (DOF) with their body requiring almost no extra training. Yet the challenges in PbD with passive observation lie in correctly recognizing human actions, converting them to robot motions, and reproducing the actions on the robot in a correct sequence [3]. Interactive task learning (ITL) [4], [5] focuses on the learning of the task concept on a symbolic level instead of only the task itself, which facilitates the learning of new assembly processes in highly flexible sequences including a huge variety of products in SMEs.

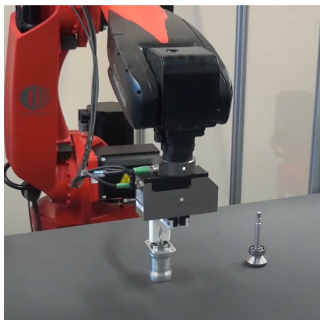
*J. Ding, A. Perzylo, L. Zhou are with fortiss, Research Institute of the Free State of Bavaria associated with Technical University of Munich, Germany. Please direct all correspondence to ding@fortiss.org.



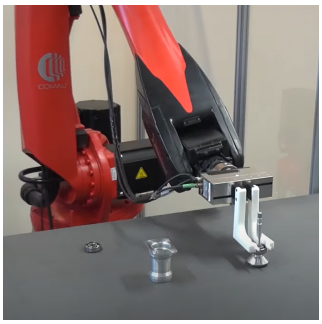
(a) Tip Pinch



(b) Precision Sphere



(c) Cylindrical Grasp Inside



(d) Cylindrical Grasp Outside

Fig. 1: In the assembly steps of the same *BallBearing* on a *MechanicalPipe* or a *MechanicalTree*, the performed grasp types (a) (b) contain the task context for matching the corresponding skill of the gripper(c) (d).

With the development of human action recognition in machine vision [6], the vision-based methods brought advantages to the passive observation method with a more versatile use. Yet the challenge lies in the lack of a general dataset that covers all the hand-object actions in different manufacturing processes. However, research efforts tried to unify grasp types of human hands in manufacturing processes with grasp taxonomies since 1986 [7]. Despite of the differences in human actions for the specific tasks in manufacturing, the grasp types of the human actions can be generally described with the palm-thumb position and the type of contact with the object [8]. With the grasp type described in a taxonomy as a general classification of the labels in human action recognition, the hand-object action dataset is simplified in the structure and thus also the collection and annotation process. Additionally,

the grasp types can be semantically modeled containing the context of tasks and the type of contact, with which the task-level information of the hand-object actions are inferred and consequently ease the mapping of a robotic process for its execution.

In this paper, we present our work towards a PbD concept for our knowledge-based robot system that semantically integrates the grasp types with their task-level information as the basis for understanding hand-object interactions in assembly processes. With several *PickAndPlace* tasks from an assembly process involving a *BallBearing*, as shown in Fig. 1, we showcase the steps from the composition of human-object actions with grasp type recognition and the generation of the platform-independent human assembly process, to the transformation into a robotic process based on skill descriptions.

II. Related Work

A. Programming by demonstration with grasp taxonomies

Grasp taxonomies [7], [8] are widely investigated in PbD to understand human grasp types and reproduce the grasps with robotic grippers. [9] presented and evaluated three methods for grasp recognition, where the hand posture during the grasp sequence or the hand trajectory with the hand rotation is firstly separately considered and then combined. In the following work [10], a method for the generation of the approach vector for robotic grasps based on object information and human demonstration was introduced. In [11], [12], human demonstrations were performed in a Virtual Reality (VR) environment for the convenient calculation of contact points and normals of human grasps. Upon this information, the human grasps are then categorized based on a grasp taxonomy, and the robot pregrasp planning is generated from the human grasp trajectory.

Most research in PbD involving grasp taxonomies focuses on the generation of parameters or trajectories of robotic grasps, where the task-level information of the grasp types in the human demonstration is still missing. In this work, we consider the grasp taxonomy as the backbone for understanding human actions. With the information of human grasps related to the performed tasks, the hand-object actions are exploited for the reproduction of robotic grasps on a task level.

B. Vision-based hand action recognition

With the progress of machine learning and computer vision in recent years, different methods for image-based hand action recognition have been developed. In [6], [13], recent approaches and datasets in Human Action Recognition are summarized and reviewed. Among the machine learning methods, *recurrent neural network* (RNN) with *Long Short-Term Memory* (LSTM) is the focus on skeleton-based methods either for body action or hand gesture recognition [14], [15], [16].

The vision-based methods for hand action recognition need to be trained on datasets that cover a large variety of actions. [17] annotated video sequences of complete assembly processes with fine-grained and coarse actions, such as *PickUp*, *PutDown*, and *Screw*. But this dataset only covers the actions in the assembly process of 101 toy trucks and cannot be directly reused for industrial assembly processes of different products. Another type of dataset is the collection of short clips consisting of single actions. [18] introduced a 3D hand pose dataset in RGB-D data and hand pose with 21 joint positions covering 45 daily hand action types. Although the dataset only contains daily action types and can thus not be directly used in industrial assembly processes, this study offered a taxonomy for hand-object actions and their relationship with grasp types. We follow the dataset structure from this work for the recognition of grasp types from sequences of hand skeletons and combine an LSTM due to its good performance in skeleton-based hand action recognition.

III. Concept

The difficulty in understanding the assembly process for industrial products lies in the high flexibility in task orders performed by humans and the fast-changing products from SMEs with different assembly procedures. As mentioned in Section II, despite the good performance of vision-based methods, the process of collecting and annotating datasets for hand-object actions recognition is elaborate and not suitable for manufacturing SMEs. Additionally, the vision-based method offers only the classification of the grasp type, where the rich task-level information for the understanding of hand-object actions are often neglected.

Extending our knowledge-based digital engineering concept [19], hand-object actions based on a grasp taxonomy III-A are semantically modeled with the OWL 2 Web Ontology Language and stored in a Knowledge Base (KB), where all relevant knowledge about an assembly process – from objects, tasks, and workcells with capability descriptions – are

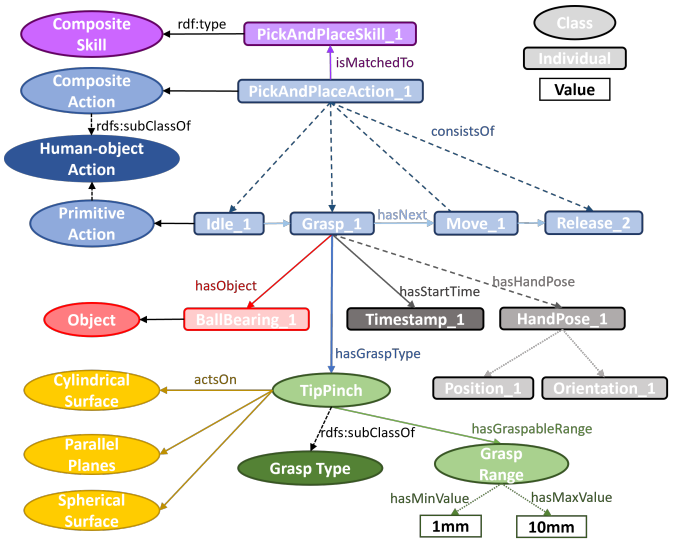


Fig. 2: Examplary semantic model of a hand-object action with task context from the grasp type; A sequence of *PrimitiveActions* forms a *CompositeAction* of *PickAndPlace* for matching of human actions and robotic skills at different levels.

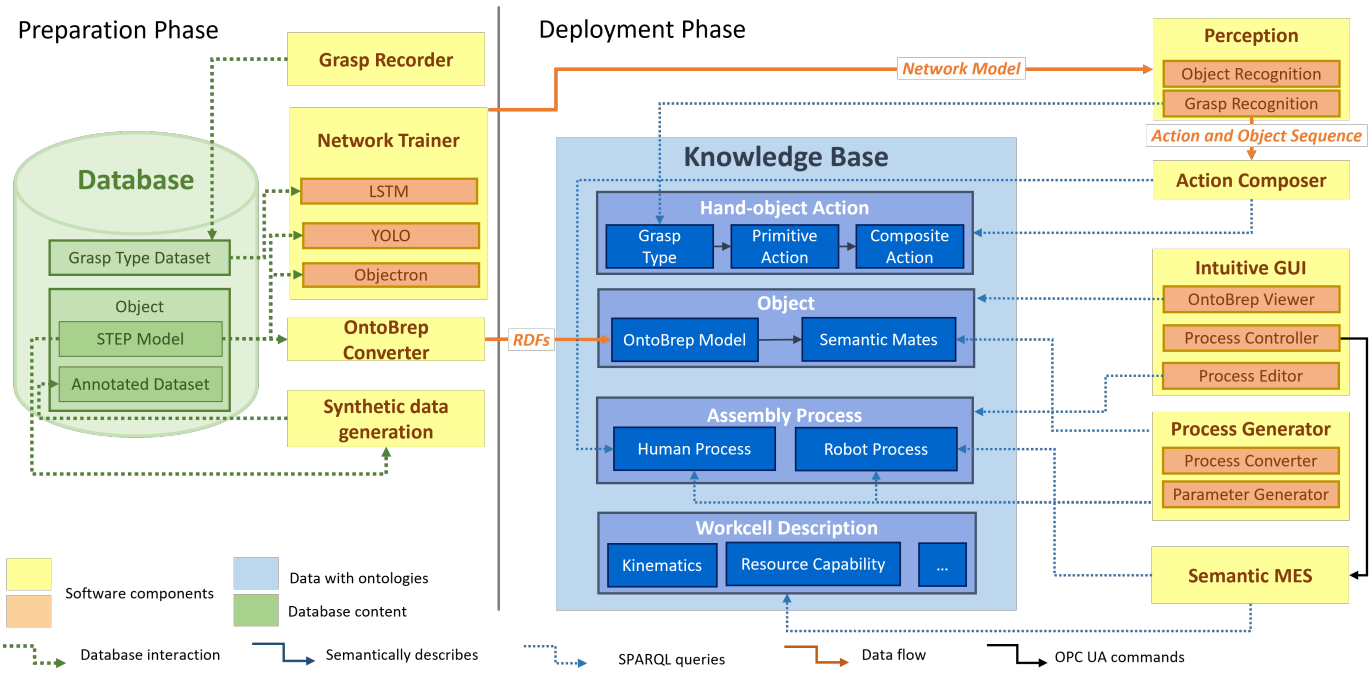


Fig. 3: An overview of the proposed PbD paradigm with the Knowledge Base (KB) on the right for storing the ontology-based semantic models, the peripheral software components in yellow interacting with the KB via SPARQL queries. The preparation phase on the left contains a database with CAD models and datasets for training the network as a prerequisite.

interconnected as shown in Fig. 3. The peripheral software components use SPARQL 1.1 (SPARQL Protocol and RDF Query Language) requests to query and manipulate the models in the KB, in order to support the hand-object action recognition and the specification of human assembly processes.

Other information is stored in a database. This includes the original CAD models of objects, synthetic image datasets for object recognition (Section III-B2), and sequences of hand skeleton data for grasp type recognition (Section III-B1). Several software components are developed, as described in Section III-B, facilitating the conversion between different data types from the KB and the database.

A. Hand-object action

The hierarchical structure of the grasp taxonomy is naturally easy to be semantically modeled with ontologies, which allows a semantic connection of the human grasp types to the task-related information for understanding the hand-object actions in various industrial assembly processes, such as the type of surfaces that it can act on, and the required grasp range for a given object's diameter as shown in Fig. 2.

Furthermore, the semantic model of grasp types contains additional information for generating parameters of a robotic process from [19], such as the matching of human grasp types and robotic grasp skills. An example in the assembly process is that the two grasp types of *TipPinch* and *PresicionSphere* on the same ball bearing can offer different contexts for the assembly tasks as shown in Fig. 1 due to the constraints on either the outer or the inner ring of the ball bearing. Although the dexterous human hand can also perform the *PrecisionSphere* grasp on the outer ring for the assembly of the ball bearing and the mechanical tree, the user was instructed to only grasp the part of the object with no constraint for the precise definition of the contact.

The hand-object actions are also defined in different layers for reusability and flexibility. As the tasks of a robotic process were defined on different layers [20], a set of *Primitive Actions* (*Grasp*, *Move*, *Release*) in an order can also form a *CompositeAction* of *PickAndPlace*, which also enables a homogeneous description at a different level for human actions, robotic skills, and assembly tasks.

B. Perception

While the development of novel perception algorithms is not the focus of this work, we still face the challenges of recognizing the hand-object actions for specific tasks and the fast-changing products in SMEs, where the resources for the collection and manual annotation of datasets for specific work-pieces and human actions are typically missing. In this work we develop two automated pipelines for grasp type recognition and object recognition based on RGB-D image from an Intel D435 depth camera. With the basic information of grasp types or object types from the compact yet robust networks, other information, that is necessary for the generation and the reproduction of the human assembly processes, can be inferred from the semantic description of objects [21] and hand-object actions (Section III-A) in our KB.

1) Hand-object action recognition: The standard LSTM model from TensorFlow[1] is used to classify the grasp types from the grasp taxonomy [8] with a time series of hand skeletons. A hand skeleton of 21 joints with their positions of $[x, y, z]$ is generated from the RGB image with Mediapipe Hands [22] as shown in Fig. 1. Despite of the compact structure from the standard LSTM model from Tensorflow without further modifications, a 96.7% recognition rate on the test set is achieved with an average Frames per Second (FPS) of 15 tested on a computer with an Intel i7-9850H CPU and an Nvidia Quadro T2000 GPU. In order to improve the recognition rate, we define the exact grasp types and give instructions to the user to imitate the hand shape for the grasp type. Such as if the hand is idle with no action performed, the fingers should be fully opened to avoid being misclassified as some grasp type. In addition to the recognition of static grasp types, the LSTM can be extended for the recognition of continuous actions such as screw tightening due to its authentic property for processing time series.

Apart from the grasp type, other information based on the grasp are also recorded and further semantically stored in the KB, such as the timestamp of the start/end of the action and the normalized hand pose in the world coordinate system, which is necessary for inferring the interaction object (Section III-B2).

With a sequence of recognized grasp types, the primitive actions of *Grasp*, *MoveHand*, and *Release* can be extracted upon change of grasp types, such as a *Grasp* can be extracted when the grasp type changes from *Idle* to *TipPinch*.

2) Object recognition: The original CAD models of the workpieces are utilized for the generation of a photo-realistic synthetic dataset using Ominiverse Issac[2] and the object recognition with YOLO v3 [23] for rapid iteration with new parts. So far, we have tested the pipelines with our 3 object types of *Mechanical Pipe*, *Mechanical Tree*, and *Ball Bearing* as depicted in Fig. 1. We achieved a recognition rate of 98% on the synthetic test set with in total 10,000 images. The bounding box from YOLO in the image coordinate system is used to define the interacting object with the hand.

In certain assembly scenarios, the 6-dimensional pose instead of only the type of product is necessary for defining the constraints of the objects, such as the assembly step of a *BallBearing* on a *MechanicalPipe*, with multiple possibilities given by the Semantic Mates [24] for the cylindrical constraints. Objectron [25] can be integrated for detecting the object poses, which can also be trained with the synthetically generated dataset from Omniverse Issac.

C. Process description

A human assembly process consists of a sequence of semantic hand-object actions on a task level in Section III-A and the interacted objects from Section III-B2. By connecting semantic information regarding the grasp type with geometric information from an OntoBREP model [21], the geometric constraints of the objects in each assembly task can then be determined with the help of Semantic Mates [24]. Such information is used for matching the tasks and the provided skills of the software and hardware components in the workcell. Not only the gripper type, e.g., a parallel gripper or a vacuum gripper, can be determined by matching the grasp ranges of various grasp types and the geometric information of the objects' graspable surfaces, but the specific skill for the execution of the task can also be determined in certain cases. An example of matching the grasp skills from the inside and outside of the ball bearing is given as in Fig. 1, where the human explicitly defines the graspable surfaces. In this case, the approach vector of the robotic grasp can be performed vertically from the top, but can be calculated with the hand pose and the object pose in more complicated assembly steps.

D. Graphical user interface

Our graphical user interface (GUI) is implemented based on the Angular[3] framework and communicates with the KB via a REST API. The fundamental function of the GUI in the overall PbD paradigm is to enable the user to modify the assembly process with the necessary information for the correct generation and execution of the robotic process, that is either required due to multiple results from the matching of Semantic Mates [24] or partially covered and not recognized objects (Section III-B). Furthermore, an OntoBrep Viewer is developed and integrated into the GUI for the intuitive interaction with the objects on a semantic level, such as choosing the outer ring on the ball bearing. With the missing information manually filled, the task sequence of the assembly process can also be adjusted with *Add*, *Delete*, or *Edit Parameter*, so that a robotic process composed of a sequence of tasks can be executed in a desired order by the semantic manufacturing execution system (sMES) [19] with skills offered by the components in a workcell wrapped with the OPC UA (Open Platform Communications Unified Architecture) middleware providing an unified interface [26].

IV. Conclusion

In this work we propose a PbD concept with an LSTM network for grasp type recognition in industrial assembly processes. The semantic model of hand-object actions with the task-level information related to the grasp type is defined in the framework with other necessary software components for interacting with the semantic knowledge of objects, processes, and workcells. A showcase of *PickAndPlace* tasks involving a *BallBearing* is presented with the framework, which can be extended with other actions for different assembly tasks, such as screw tightening. The human process is generated focusing on the task level reproduction with the matching of the robotic skills, where the parametrization of the approach vector and force estimation of each robotic task still need to be calculated from hand and object poses in future works. Experiments with users are to be conducted for the evaluation of the system.

[1]https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM

[2]https://developer.nvidia.com/isaac-sim

[3]https://angular.io/

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952197.

REFERENCES

[1] A. Perzylo, N. Somani, S. Profanter, I. Kessler, M. Rickert, and A. Knoll, "Intuitive instruction of industrial robots: Semantic process descriptions for small lot production," in *2016 ieee/rsj international conference on intelligent robots and systems (iros)*. IEEE, 2016, pp. 2293–2300.

[2] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[3] Z. Zhu and H. Hu, "Robot learning from demonstration in robotic assembly: A survey," *Robotics*, vol. 7, p. 17, 2018.

[4] J. E. Laird, K. Gluck, J. Anderson, K. D. Forbus, O. C. Jenkins, C. Lebiere, D. Salvucci, M. Scheutz, A. Thomaz, G. Trafton, *et al.*, "Interactive task learning," *IEEE Intelligent Systems*, vol. 32, no. 4, pp. 6–21, 2017.

[5] J. R. Kirk and J. E. Laird, "Interactive task learning for simple games," *Advances in Cognitive Systems*, vol. 3, no. 13-30, p. 5, 2014.

[6] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[7] M. Cutkosky and P. Wright, "Modeling manufacturing grips and correlations with the design of robotic hands," in *Proceedings. 1986 IEEE International Conference on Robotics and Automation*, vol. 3, 1986, pp. 1533–1539.

[8] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[9] S. Ekvall and D. Kragic, "Grasp recognition for programming by demonstration," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. IEEE, 2005, pp. 748–753.

[10] ——, "Learning and evaluation of the approach vector for automatic grasp generation and planning," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, 2007, pp. 4715–4720.

[11] J. Aleotti and S. Caselli, "Grasp recognition in virtual reality for robot pregrasp planning by demonstration," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* IEEE, 2006, pp. 2801–2806.

[12] ——, "Grasp programming by demonstration in virtual reality with automatic environment reconstruction," *Virtual Reality*, vol. 16, pp. 87–104, 2012.

[13] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[14] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1227–1236.

[15] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[16] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognition*, vol. 76, pp. 80–94, 2018.

[17] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao, "Assembly101: A large-scale multi-view video dataset for understanding procedural activities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 096–21 106.

[18] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 409–419.

[19] A. Perzylo, I. Kessler, S. Profanter, and M. Rickert, "Toward a knowledge-based data backbone for seamless digital engineering in smart factories," in *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1. IEEE, 2020, pp. 164–171.

[20] A. Perzylo, J. Grothoff, L. Lucio, M. Weser, S. Malakuti, P. Venet, V. Aravantinos, and T. Deppe, "Capability-based semantic interoperability of manufacturing resources: A basys 4.0 perspective," *IFAC-PapersOnLine*, vol. 52, no. 13, pp. 1590–1596, 2019.

[21] A. Perzylo, N. Somani, M. Rickert, and A. Knoll, "An ontology for cad data and geometric constraints as a link between product models and semantic robot task descriptions," in *2015 ieee/rsj international conference on intelligent robots and systems (iros)*. IEEE, 2015, pp. 4197–4203.

[22] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[24] F. Wildgrube, A. Perzylo, M. Rickert, and A. Knoll, "Semantic mates: Intuitive geometric constraints for efficient assembly specifications," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6180–6187.

[25] A. Ahmadyan, L. Zhang, A. Ablavatski, J. Wei, and M. Grundmann, "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7822–7831.

[26] S. Profanter, A. Breitkreuz, M. Rickert, and A. Knoll, "A hardware-agnostic opc ua skill model for robot manipulators and tools," in *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2019, pp. 1061–1068.