2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)
August 23-27, 2021. Lyon, France

Visual Scene Understanding for Enabling Situation-Aware Cobots

Markus Eisenbach, Dustin Aganian, Mona Köhler, Benedict Stephan,
Christof Schröter and Horst-Michael Gross

***Abstract*— Although in the course of Industry 4.0, a high degree of automation is the objective, not every process can be fully automated – especially in versatile manufacturing. In these applications, collaborative robots (cobots) as helpers are a promising direction. We analyze the collaborative assembly scenario and conclude that visual scene understanding is a prerequisite to enable autonomous decisions by cobots. We identify the open challenges in these visual recognition tasks and propose promising new ideas on how to overcome them.**

I. Introduction

For many small and medium-sized enterprises (SMEs), large and highly specialized production lines are not flexible enough to justify their high investment. Instead, these SMEs can generate their profits with versatile assembly in small quantities [1]. In this scenario, full automation is rarely achieved and some complex assembly sequences can be executed faster and cheaper by a worker. To optimize the process in the course of Industry 4.0, robots can execute some simple assembly and transport tasks, which may require close interaction with humans. This can speed up the assembly process, not through high automation, but through teamwork and parallelization.

However, robots and humans usually work in separate areas. If a human gets close to a robot, the robot stops immediately to avoid injury. Furthermore, robots only perform tasks in which every production step and every object involved is known, which is a severe limitation in flexible production. Therefore, the goal of the next stage in Industry 4.0 is the cooperation between humans and situation-aware collaborative robots (cobots) [2]. To enable situation-aware decisions of the cobot, visual scene understanding plays a key role, which is the focus of our paper.

II. Collaborative Assembly Scenario

To further define the context in which our cobot has to work, in the following, we present our collaborative assembly scenario. Our cobot has access to multiple cameras observing the workspace. After observing the assembly process a few times, the cobot should be able to anticipate when tools or workpieces are needed (see Fig. 1). The cobot should act as a helper, performing transport tasks between warehouse and workplace and handing over tools and objects as needed.

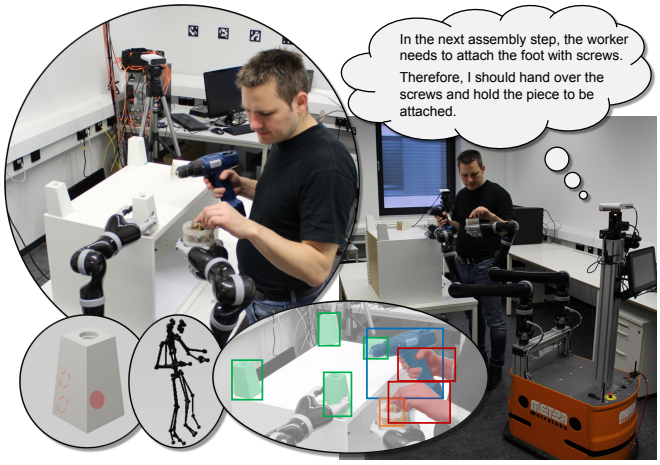

Fig. 1. The cobot is assisting the worker at our prototypical workplace in our vision lab. To be able to autonomously plan helpful actions, the cobot must understand the assembly process, and thus must analyze its visual input. This includes detecting objects, recognizing human actions, estimating 3D shapes, and predicting adequate grasp poses.

Obviously, for this scenario, the cobot must master many tasks, such as navigation and actuator control. But besides these complex robotic tasks, first of all, the cobot must recognize objects and people in the scene and understand the assembly process. The cobot must know what the current assembly state is and what it can do to help the worker.

III. Visual Scene Understanding

Recognizing the current assembly state is especially challenging because it requires robust object detection and human action recognition. This should also be achieved even if certain objects or human actions are very similar or not present in training data. But scene understanding for manipulation is also extremely difficult, as it requires adequate 3D shape and grasp pose estimation combined with uncertainty estimation to avoid accidents. Therefore, in order to make the robot aware of the situation and let it act as a helper, initially, scene analysis should be the main focus and all subtasks should be addressed by powerful machine learning approaches. When applying state-of-the-art approaches for visual scene understanding, certain shortcomings must be addressed, as described below.

A. Object Detection

As occurring objects vary, extensive scenario-specific labeling of training data is impractical. A promising direction in our scenario is few-shot object detection [3]. It leverages

This work has received funding from the Carl Zeiss Foundation as part of the project Engineering for Smart Manufacturing (E4SM) – Engineering of machine-learning-based assistance systems for data-intensive industrial scenarios under grant agreement no. P2017-01-005

C. Schröter is with MetraLabs GmbH, 98693 Ilmenau, Germany www.metralabs.com

M. Eisenbach, D. Aganian, M. Köhler, B. Stephan and H.-M. Gross are with Neuroinformatics and Cognitive Robotics Lab, Ilmenau University of Technology, 98693 Ilmenau, Germany www.tu-ilmenau.de/neurob

knowledge from training on base categories with abundant training data in order to detect objects from novel categories with only a few labeled instances per category. Its application in an industrial setting has yet to be implemented.

As another way of recognizing objects, which has not been explored yet, we propose to apply a class-agnostic object detector [4], which aims to detect objects irrespective of their semantic class. The resulting detections can then be fed into a module for re-identifying previously seen objects.

B. Human Action Recognition

Typically, in action detection and recognition, actions are classified only on already known classes. Although there are several hundred different classes present in current datasets, e.g. [5], only few of them are relevant for industrial assembling. Furthermore, the few available assembly datasets, such as [6], are rather small.

However, in advance it is unclear which actions a person will perform when building a new workpiece and post-labeling and post-training of the performed actions is impractical. Therefore, as a novel idea, we formulate this action recognition as a re-identification problem. By recognizing actions from previous assembly processes, the current assembly state can be identified.

To enable action re-identification, we prefer to use skeletons rather than RGB images as they generalize better for different persons and environments. Since some assembly actions can only be distinguished by the movement of the hands, we extend the full body skeleton with estimated hand skeletons. Moreover, tables or the workpiece itself lead to occlusion of skeleton joints. Therefore, we make use of multiple cameras around the workplace to predict complementary skeletons.

C. 3D Shape Estimation

Model-based methods for grasp pose estimation expect a 3D shape description. However, in general training data for 3D shape estimation [7] are synthetic and often rendered without textures and background. Even if textures and background were present, widely used methods generalize badly to new object categories [8]. Therefore, we propose to estimate uncertainties [9] and adapt these predictors by post-training on the most uncertain objects in order to specialize to scenario-specific objects.

D. Grasp Pose Estimation

Grasp Pose Estimation either expects 3D shapes of the object (see Sec. III-C) or processes depth images directly. However, for the latter methods, common datasets, such as [10], contain images with only few labeled grasp poses. Yet, providing an exhaustive amount of labeled grasp poses is impossible due to an infinite amount of possible grasp poses. This may result in possible contradictions in the dataset where a similar input is labeled with different grasps poses. A neural network trained on these contradictions would average these poses resulting in possibly unfeasible or unstable grasps poses. Current work [11] only partially solves this problem. We propose a more promising way to resolve existing contradictions by estimating the aleatoric uncertainty (a.k.a. data uncertainty). This way, we can either reject estimates with high uncertainty or incorporate the uncertain estimates into the training process of the self-learning cobot.

E. Uncertainty Estimation

To avoid accidents, uncertainties should be estimated [9] and handled for all visual scene understanding tasks. This includes uncertainties in the detected positions of objects and the worker, uncertainties in features for re-identification of assembly actions, uncertainties in 3D object shape estimates, and uncertainties in estimations of grasp poses. Unfortunately, methods for robust uncertainty estimation in deep learning models are still in their infancy, but are essential for a practical implementation of safe cobots. Therefore, transferring uncertainty estimation for classification tasks to detection, grasp pose estimation, etc., will also be in the focus of our future work.

IV. Conclusion

In the context of Industry 4.0, cobots as helpers become more relevant, especially in versatile manufacturing. Since the cobot should act as an autonomous helper, we propose to focus on visual scene understanding as an important first step. We identified several challenges when applying current state-of-the-art methods to an industrial setting. To overcome these challenges, we propose promising new directions:

- As occurring objects vary in versatile manufacturing, we propose few-shot learning, which allows detection with little training data.
- To cope with new objects and human actions, which are unknown during training, we propose to formulate object detection and human action recognition as a re-identification problem.
- For a more generalizing action recognition to varying persons and environments we propose to use multi-view full body and hand skeletons instead of image data.
- To combat generalization deficits in 3D shape estimation and contradictions when training grasp pose detection, we propose to estimate and handle uncertainties.

References

[1] A. Radziwon *et al.*, "The smart factory: Exploring adaptive and flexible manufacturing solutions," *Procedia Engineering*, 2014.
[2] M. Strehlitz, "Tempolimit für Roboter," *VDE dialog*, 2018.
[3] Q. Fan *et al.*, "Few-shot object detection with attention-RPN and multi-relation detector," in *CVPR*, 2020.
[4] A. Jaiswal *et al.*, "Class-agnostic object detection," in *WACV*, 2021.
[5] L. Smaira *et al.*, "A short note on the kinetics-700-2020 human action dataset," *arXiv*, 2020.
[6] Y. Ben-Shabat *et al.*, "The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose," in *WACV*, 2021.
[7] N. Wang *et al.*, "Pixel2Mesh: Generating 3D mesh models from single RGB images," in *ECCV*, 2018.
[8] M. Tatarchenko *et al.*, "What do single-view 3D reconstruction networks learn?" in *CVPR*, 2019.
[9] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016.
[10] "Cornell dataset," http://pr.cs.cornell.edu/grasping/rectdata/data.php.
[11] D. Morrison *et al.*, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv*, 2018.