Automated Item Picking - Machine Learning Approach to Recognize Sports and Fashion Articles

Moritz Weisenböhler, Michael Walz, and Christian Wurll

Karlsruhe University of Applied Sciences
moritz.weisenboehler@hs-karlsruhe.de
christian.wurll@hs-karlsruhe.de

Abstract. Automated Item Picking (AIP) systems can be used to improve the quality of order picking processes and to reduce the cost. This work presents an AIP solution for the sports and fashion sector. The system consists of an industry robot, a customized gripper and an AI-based perception. A convolutional neural network is trained for identification and localization of various apparel articles and shoeboxes under real conditions.

Keywords: automated item picking, computer vision, object detection, machine learning, neural networks

1 Introduction

With the worldwide increase in sales figures for e-commerce, the challenges for associated logistics are also rising. Nevertheless, many logistical fulfillment processes are still largely manual and labor-intensive. The use of automated order picking systems for repetitive processes can increase the quality and flexibility of the process, while reducing the costs. However, these systems are rarely used due to the great challenges in terms of perception and the required flexibility. In many applications a large variety of objects has to be handled. The rapid progress in AI can offer appropriate solutions for these challenges.

In a cooperation project between the Karlsruhe University of Applied Sciences, Adidas AG, KUKA AG, Zimmer GmbH and Roboception GmbH a prototypical solution is developed to automate the order picking process for sports and fashion articles. The primary goal is to fully automatically detect and pick shoeboxes and plastic-wrapped apparel from a tote. Special challenges here are handling shoeboxes with opened lids, the chaotic arrangement in the tote and the high number of inhomogeneous articles. This paper focuses on the use of machine learning methods for object detection.

2 Related Work

The breakthrough of machine learning in image processing was achieved by Krizhevsky et al. in 2012. With AlexNet, a Convolutional Neural Network (CNN) for image classification, they clearly won the prestigious competition ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [1]. Numerous further developments followed, such as the *GoogLeNet* designed by Szegedy et al. with the so-called *"Inception"* architecture, which won at the ILSVRC 2014 [2].

In addition to image classification, object localization is a central task in computer vision. While image classification only assign a label to the whole image, localization

approaches detect the position and class of multiple objects in an image. The region-based CNN (R-CNN) developed by Girshick et al. laid the foundation for modern object localization [3]. Further developments such as the *Fast R-CNN* [4] and *Faster R-CNN* [5] improved in particular the recognition performance and the speed of the method. Besides the classification and object localization, segmentation is another possible application of CNNs. He et al. developed an architecture known as *Mask R-CNN*, allowing segmentation at pixel level. *Mask R-CNN* is an extension of the *Faster-R-CNN* algorithm [6].

Today CNNs have become the standard for most object detection tasks, as the winners of well-known Amazon Robotics Challenge (ARC) illustrate [7]. For example the winning teams of 2016 (TU Delft) and 2017 (Queensland UT) both used CNNs for perception [8] [9].

3 System Overview

The developed automated item picking system consists of the following components: the robot (control) and the PLC, the hybrid gripper and the perception system, all managed by the central flow control (see Figure 1).

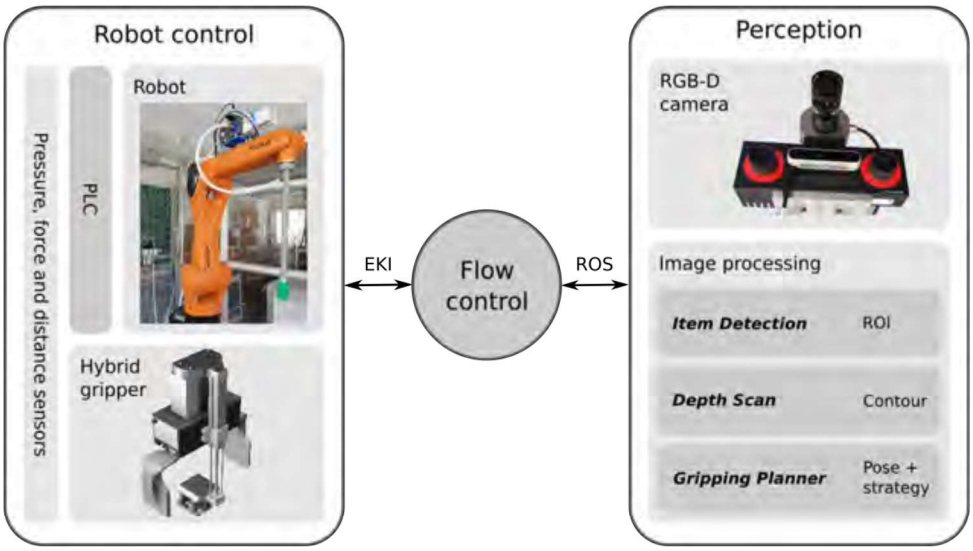

Fig. 1: Schematic system overview.

A KUKA *KR 10 AGILUS* is selected as the main actuator. With a payload of 10 kg, this robot is able to handle all relevant products and is fast enough to reach the desired cycle times, while having a good cost-benefit ratio. All sensors, including pressure, force and distance control, are connected to a Beckhoff PLC directly interfacing with the robot. The robots movements are not programmed directly in the robot control, but provided by the flow control over an Ethernet interface (EKI).

The gripper is designed for the problem specific items: plastic bags and shoeboxes. Especially the handling of the shoeboxes with unfixed lids was a main challenge for the system. Based on the *Zimmer* pneumatic 2-jaw gripper *GP400*, a hybrid gripping system with an extendable cylinder, multiple suction cups and two coated clamping jaws is developed and evaluated. The resulting gripper can handle various types of plastic-wrapped clothes and perform orientation related shoebox picks without opening lids.

Being designed very modular, the perception system allows using any depth sensor, providing RGB images and a Robot Operating System (ROS) interface for image acquisition. Mainly used and tested are the *Roboception rc_visard 160* with a RandomDot projector and the *Intel RealSense^{TM} D415*. To calculate the optimal gripping item and gripping position the acquired images are passed through a multi step process. In a first stage the *Item Detection* module proposes a region of interest (ROI) by a CNN detecting individual items in the RGB image. The ROI is transferred to the *Depth Scan* module, which uses the depth image to determine exact contour if the item. Finally, the *Gripping Planner* selects the best positioned item, calculates its pose and determines an appropriate movement strategy depending on the environment.

4 Item detection with machine learning

The goal of the perception system is to identify and locate items in a tote. Therefore, different CNNs are trained and evaluated by a total of 15 different sports articles. The machine learning solution is based on state-of-the-art models and realized with Google's *TensorFlow* framework.

4.1 Data collection

The data collection process is based on the experience of the Delft Robotics team, winner of the ARC 2016. Hernandez et al. divide the process into two different subgroups, a basic and a tote set, in order to train the CNN in two stages [10]. The basic set only contains images of individual objects and is used to train the CNN on all generic object characteristics. Afterwards, the pre-trained CNN is fine-tuned with the tote set consisting of scenes of multiple objects placed in one container. A total of more than 8,500 RGB images (8,200 basic and 300 tote) are produced. The setup for data generation contains a *Roboception rc_visard 160* and an *Intel Real Sense SR300* 3D camera, as well as individual controllable lighting modules.

4.2 Training

Various state-of-the-art neural networks are trained and tested, with three models being considered in more detail: SSD MobileNet, Faster R-CNN Inception and Faster R-CNN Inception ResNet. SSD MobileNet is a high speed model specially developed for mobile applications. The latter two are based on the processing-intensive but more precise Faster R-CNN approach. They differ by using different submodels to classify the regions of interest.

As explained in subsection 4.1, all training processes are divided into a pre-training and a fine-tuning part. The first part typically takes several hours up to a day. Due to the comparatively small training set the latter part finishes in less than one hour. All neural networks are trained on a single *Nvidia GeForce GTX 1080 Ti* graphics card.

5 Results

The three most relevant models are compared in terms of inference speed (duration on GPU and CPU) and accuracy (see Table 1).

Table 1: Comparison of three CNN-models for object localization

Criteria	SSD MobileNet	Faster R-CNN Inception	Faster R-CNN Inception ResNet
Duration CPU	0,174 s	1,532 s	20,353 s
Duration GPU	0,020 s	0,109 s	0,495 s
Precision mAP	41,19 %	82,12 %	92,01 %

In terms of inference speed, SSD MobileNet shows the strength of its comparatively lightweight design. Even without a GPU, the inference takes less than 0.18 seconds processing time per image. The Faster R-CNN Inception takes about 1.5 seconds for a calculation, which is almost nine times longer than SSD MobileNet. On average, Faster R-CNN Inception ResNet takes more than 20 seconds for a calculation. This is obviously too slow in terms of a targeted detection cycle time of about one second. Here the use of a GPU is required for shortening the processing time to about half a second.

For the evaluation of the detection accuracy, an excluded test set of roughly 10% of the total images are used. With a mAP of only 41%, the fast SSD MobileNet is not suitable for a secure item detection. Better results are achieved using the Faster R-CNN networks. Inception ResNet achieves the best mAP of 92%, while Inception scores slightly lower (mAP of 82%).

The Faster R-CNN Inception ResNet provides good forecasts for scenes containing deformable plastic-packed articles (see Figure 2a). A high recognition reliability can be achieved in particular if the objects lay clearly visible in the tote. In the case of concealed or strongly deformed objects, the CNN shows weaknesses. However, even for humans, it is difficult to discern all objects in these constellations. For scenes with several shoeboxes the Faster R-CNN Inception ResNet shows also good results. These are reliably localized even for challenging arrangements with boxes lying very close to each other or opened lids (see Figure 2b).

(a) Tote scene with socks.

(b) Tote scene with shoeboxes.

Fig. 2: Example results for Faster R-CNN Inception ResNet.

The use of a orthogonal bounding box is a disadvantage for the precise contour detection of inclined shoeboxes. Although the detected object is completely enclosed by the bounding box, it does not provide enough information about the exact location and rotation of the object. This is why the downstream Depth Scan module is required (see

section 3). The algorithm uses the corresponding depth image to determine the contour of an item inside the detected ROI.

6 Conclusion

The present paper examined machine learning techniques to detect shoeboxes and deformable plastic-packed articles in realistic scenarios. Different neural networks were trained and evaluated to overcome the analyzed challenges. Based on the Faster R-CNN Inception ResNet, the developed solution was able to reliably detect and locate different objects by means of bounding boxes. The reproducible learning process can be quickly transferred to new objects with sufficient training data available.

Further investigations will concentrate on the reliabilities of different degrees of item distinction categories, the detection of unmanageable arrangements and items (e.g. opened shoeboxes) and the detection of shoebox poses. Especially the latter task requires the detection of the lid and the direct involvement of the corresponding depth data.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6) (2012) 84–90
2. Szegedy, C., Liu, W., et al.: Going deeper with convolutions. 2015 IEEE CVPR (2015) 1–9
3. Girshick, R., Donahue, J., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. (2013)
4. Girshick, R.: Fast r-cnn. (2015)
5. Ren, S., He, K., et al.: Faster r-cnn. towards real-time object detection with region proposal networks. (2015)
6. He, K., Gkioxari, G., et al.: Mask r-cnn. (2017)
7. Wurman, P., Romano, J.: The amazon picking challenge 2015 [competitions]. IEEE Robot. Automat. Mag. **22**(3) (2015) 10 – 12
8. Milan, A., Pham, T., et al.: Semantic segmentation from limited training data. (2018)
9. Hernandez, C., Bharatheesha, M., et al.: Team delft's robot winner of the amazon picking challenge 2016. (2016)
10. Hernandez Corbato, C., Bharatheesha, M., et al.: Integrating different levels of automation. lessons from winning the amazon robotics challenge 2016. IEEE Trans. Ind. Inf. (2018)