DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

The Potential of Deep Learning based Computer Vision in Warehousing Logistics

Das Potenzial Deep Learning basierter Computer Vision in der Intralogistik

Jérôme Rutinowski *[0000-0001-6907-9296]*
Hazem Youssef *[0000-0002-7197-9127]*
Anas Gouda *[0000-0002-3062-5656]*
Christopher Reining *[0000-0003-4915-4070]*
Moritz Roidl *[0000-0001-7551-9163]*

Chair of Material Handling and Warehousing
Faculty of Mechanical Engineering
TU Dortmund University

This work describes three deep learning based computer vision approaches, that hold the potential to increase the degree of automation and the productivity of common warehousing procedures. These approaches will focus on: the re-identification of logistical entities, especially when entering and leaving the warehouse; the multi-view pose estimation of logistical entities to track and to localize them on the shop floor; and the category-agnostic segmentation of items in a bin for robotic grasping.

[Keywords: Deep Learning, Computer Vision, Re-Identification, Pose Estimation, Object Segmentation]

Diese Arbeit beschreibt drei Deep-Learning-basierte Computer-Vision-Ansätze, die das Potenzial haben, den Automatisierungsgrad und die Produktivität gängiger Lagerverfahren zu erhöhen. Diese Ansätze konzentrieren sich auf: die Re-Identifizierung von logistischen Einheiten, insbesondere beim Betreten und Verlassen des Lagers; die Multiview-Positionsschätzung von logistischen Einheiten, um sie in der Fabrik zu verfolgen und zu lokalisieren; und die kategorienunabhängige Segmentierung von Artikeln in einem Behälter für das Greifen durch einen Roboter.

[Schlüsselwörter: Deep Learning, Computer Vision, Re-Identifikation, Pose Estimation, Objekt Segmentierung]

1 INTRODUCTION

Sensors and cameras enable a computer-understandable capture of reality. Their data enables evaluations and predictions in order to better understand reality. According to [1] the seamless recording and visibility of what is currently happening is the basis of the fourth industrial revolution and therefore for the automation of processes that still involve human operators. The ability to design a highly automated warehouse is of tremendous interest to the logistics industry [2]. This is the case because a high degree of automation could make logistics processes more efficient, hence increasing the operating margin by reducing either process expenses, duration, or both [3]. Improving logistics processes requires knowledge of what is happening, e.g., in a warehouse, at a given point in time. This knowledge is acquired by observation. Nowadays, observation based on computer vision is superior to human observation in many cases but is not yet widely used. Therefore, this contribution presents three ways in which computer vision based observation can be applied, providing a greater degree of automation for the handling of logistics objects around the warehouse, thereby improving warehouse performance.

The first application is the deep learning based re-identification of individual logistical entities using their inherent, visual characteristics. This method entails the use of multiple cameras, which record logistical entities, such as load carriers, at different points along the supply chain. In doing so, images of the recorded entities are stored remotely and can later on be fetched when needed, as to re-identify a previously recorded instance of a specific given entity (i.e., in the case of a pallet, not only would it be detected but the specific pallet would be identified as the unique entity that it is). This can be of great use for standardized yet non-serialized logistical entities, such as Euro-pallets, that could previously not be uniquely identified without relying on artificial features such as barcodes [4].

The second application is category-agnostic object segmentation for robotic grasping [5]. The goal of this application is to segment and categorize a variable number of yet unseen objects, as would be the case for a real warehouse, in which numerous different object types exist. Shape, color, and handling properties of these objects may remain unknown until the given object reaches the point in the supply chain at which it is handled by a robot. The use

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

of category-agnostic object segmentation goes against the more common method of segmenting only object types that are known beforehand. Such a generalized method would be useful for the deployment in differing warehousing environments, with the same exact model

Finally, the third application constitutes a multi-view pose estimation approach for localizing logistics objects. In logistics facilities, many objects are localized using checkpoints which can either result in blind spots during material flow operations or constrain the motion of moving objects, such as mobile robots, into inflexible layouts. The aim of this application is to globally localize objects on the shop floor using a system of monocular RGB cameras. Doing so could mitigate the aforementioned problems of checkpoints and inflexible navigation paths. Recent deep learning based approaches for pose estimation [6, 7] significantly outperform classical methods, making them a preferable choice. Such approaches require large amounts of data that is manually annotated. Techniques that enable the automated annotation of data are rare and usually limited in performance, but provide substantial reduction in the preparation time of the entire localization pipeline.

This contribution presents in detail the above mentioned applications of computer vision based observation and their potential benefits for the warehousing sector. In addition, we will discuss the importance of data and problem driven solution development.

2 State of the art in warehousing logistics and computer vision

This section aims at laying out the research that is considered relevant by the authors for the approaches subsequently presented in this work. The relevant literature is divided into logistics processes related research and computer vision related research, under which it will be further subdivided by each topic

2.1 Relevant warehousing processes

This subsection splits warehousing processes into three broad, non-holistic categories, namely the inbound and outbound flow of goods and the tracking and handling of material.

2.1.1 Inbound and outbound flow of goods

Material flow is defined as "[…] the interlinking of all processes in the extraction, processing and distribution of goods within defined areas. Material flow includes all forms of the passage of work objects through a system" [8]. A system, in this context, is to be understood as a given area with an input and an output [9]. In this case, focusing on the internal flow of material, this means the reception and issuing of goods.

To ensure a flawless internal flow of materials, certain storage and conveying systems as well as corresponding loading aids are necessary. Storage systems are used for the planned storage of goods, fulfilling several tasks, such as quantity balancing [10]. Conveyor systems, on the other hand, serve to move goods within the system [10]. To move goods efficiently through the various conveyor and storage systems, standardized loading aids are used. The most commonly used loading aid is the Euro-pallet (dimensions of 800 x 1200 mm), to which all elements of the material flow system are adapted [10–12]. This also includes, e.g., small load carriers, which are standardized according to VDA standard 4500 [13]. With a maximum length of 600 mm and a maximum width of 400 mm, as well as smaller sizes corresponding to the respective system dimensions, the surface of Euro-pallets can be utilized in an efficient manner [13].

In addition to the material flow itself, an information flow, which is often considered separate in the literature [9, 10], is also of importance. The information flow includes accompanying, subsequent, or preceding information for material flow control and regulation as well as supplementary data for administrative tasks. Sensors, computer-aided processing and further automation already play a role here [9]. This also includes the processes required to track logistical units. For the unambiguous tracking of logistical entities along their lifecycle, such as the load carriers already mentioned, reliable identification must be ensured. Various global standards exist for this purpose, e.g., barcodes or RFID technology [14, 15].

An alternative to subsequently applied identification features would be the camera-based identification of load carriers based on their inherent, visual features (i.e., surface structure, color patterns). This would in turn reduce the need for manual processes like attaching labels or scanning barcodes. In addition, it would thus be possible to programmatically unite goods to their respective load carriers so that they would remain jointly identifiable.

2.1.2 Material handling and grasping

In pick and place applications, most process steps are still performed by hand. Even though deep learning is capable of fully automating such tasks, it is still not widely adopted. The reason for this limited usage is the separate thinking between computer vision research and logistics. Research focuses on accuracy while logistics focuses on practicality. This separation makes the industry adopt older, classical non-learning based approaches over learning-based ones even though learning-based approaches provide higher degrees of performance. This is in part because classical non-learning approaches are easier to integrate. Hence, deep learning models should also be built with practicality in mind.

Universal Robots ActiNav, Pickit 3D, Swisslog ItemPiQ are examples of software products for robotic

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

grasping and bin picking that use non-learning algorithms but are still widely used due to their ease of integration, requiring no re-training or tailoring process.

2.1.3 Material tracking

Although material tracking approaches target moving goods in material flow operations, other logistical entities might leverage an existing tracking system in other forms (and thus make the advantages of using a tracking system twofold). Autonomous Mobile Robots (AMRs) and autonomous forklifts are among the logistical entities that could not only benefit from a performance enhancement, but also receive a solution to persistent problems in their operation using a tracking system. For AMRs or autonomous forklifts to deliver goods between two points, the navigation problem has to be solved repeatedly during the operation. The main component in robot navigation is localization [16], where the robot has to localize itself, along with the goal and any encountered obstacles, in order to navigate the environment successfully. However, using on-board sensors only to navigate the environment would eventually lead to a drift in the location of the robot [17]. On the other hand, Automated Guided Vehicles (AGVs) can only move through pre-defined paths, which are established using markings on the ground or an RFID network embedded in the infrastructure. Such approaches lead to a closed solution for the navigation problem of AGVs, but at the cost of flexibility. In dynamic warehouses or automated production facilities where humans work closely alongside humans, having fixed navigation paths is sub-optimal in terms of the navigation time and consequently in the cost of delivering the goods. To mitigate the drift problem in AMRs and the navigation rigidity issues in AGVs, a global but flexible sensor system is needed. Multi-camera systems offer a solution to such problems by acting as remote sensors that do not interfere with the robot's motion and thus allow free navigation in the process area. By possessing a global view of the shop-floor, drifts during the navigation of the robots could also be corrected by repeatedly updating the robot's position in a global manner.

Camera systems have been used extensively for the purpose of localizing and tracking mobile robots by acting as global sensors. The authors of [18] localize a mobile robot as well as perceived obstacles using a set of three cameras that are viewing the operation area from up top, mounted at a distance from one another (wide baseline) for coverage and have overlapping fields of view. The approach computes a visual hull around the obstacles, assuming a cylindrical model to create an occupancy grid of the area. A tracking algorithm, that is based on a particle filter, then uses the occupancy grid as an appearance model to keep track of the robot and the obstacles across the frames. Finally, to guarantee the distinction of the robot's identity from obstacles, the robot's odometry is used for validation. In [19], the authors use surveillance cameras mounted in narrow hallways to localize robots and obstacles. The approach targets the interaction of humans and robots in narrow passages, such as those occurring in warehouses, that could lead to deadlocks or unsafe traversal of humans. The authors resolve the coordination problem on a high level by using logical interlocks to provide suitable commands for robots in navigating the hallways when other robots or humans are detected. Other approaches, as in [20], also use multi-camera surveillance systems to localize robots in narrow hallways. The images from a two-camera system were transformed to bird's eye-view images using homography and artificial feature points that were added manually to the environment (using a checkerboard placard overlaid on the area). A common region of interest from both transformed images was then extracted. The robot is then segmented out from a binary version of the input images after applying a threshold. The real location of the robot is then obtained by projecting the actual floor image onto the projected plane containing a contour that surrounds the robot. The abovementioned methods illustrate the underlying capabilities of extending a camera-based tracking system into mobile robot localization.

2.2 Computer vision tasks in warehousing logistics

This subsection describes the computer vision tasks considered to be integral for efficient, modern warehousing logistics, namely re-identification, bin picking, and object pose estimation.

2.2.1 Identification

When using one or multiple cameras to record a subject or a set of subjects of interest, data is created that can be used in different ways. For one, the data could be used to detect movement, in the sense of a frame-to-frame anomaly detection, i.e., a change in scenery [21, 22]. Beyond the detection of movement, the presence of a certain set of subjects could be detected, as in class-based object detection [23]. A similar approach would be the simple classification of recorded images into predefined classes or clusters. Finally, for instances in which not only the detection or classification of a certain subject is of interest, but an intraclass distinction is to be made, identification comes into play [24]. Defined as the process of identifying previously recorded subjects over a network of cameras [24], identification is used in the area of pedestrian or vehicle surveillance. It also holds other uses, such as animal identification or, in this very case, the identification of logistical entities.

In order to identify objects, distinct methods can be employed. A common and straightforward one would be the use of 2D or 3D codes. These could be linear codes, such as barcodes, or two-dimensional codes, such as DataMatrix codes or QR codes [15]. Their advantage is the standardized and well-established way in which they can

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

be used. In this sense, they offer a low entry barrier, both in terms of development and in terms of costs and complexity. On the other hand however, their functionality is reliant on their legibility, which hinges on the material that they are applied to and in turn on the amount of wear and tear that it is exposed to during its life cycle.

Furthermore, sensors and transmitters could be used, presenting a more complex, cost-intense but durable and reliable alternative to codes [25]. The pairing of an RFID sensor and transmitter would be one such option. [15, 25]

Finally, the surface structure of a subject of interest could be used as its identifying feature. This approach could be used at a high level of detail and granularity, as proposed in methods such as FIBAR [26] or PaperSpeckle [27], in which the microscope-level surface structure of a subject would be analyzed. At this level, all objects on earth that would come to mind should be unique and thus allow for them to be distinguished from one another. However, the use of such methods necessitates special sensor equipment, a laboratory environment, and a dedicated recording setup.

Given these limitations, the exploitation of surface structures detectable by means of ordinary vision systems, such as a camera, would be advantageous, as long as the aim is the deployment in an industry environment. Therefore, subjects that hold certain visually detectable idiosyncrasies are suited for this type of identification. This could be humans and their inherent visual features, their gait, etc., or objects that are made out of materials that have a unique surface structure, i.e., wood [28–31]. These visual features however, would have to be extracted from the image dataset, which could be done by means of deep learning techniques. The advantage of such an approach would be the use of widely used sensors and the low setup cost and complexity. On the other hand, deep learning techniques are often considered to be black boxes and can therefore be experimental in nature and need a great amount of data and expertise to be deployed correctly and proficiently.

2.2.2 Object segmentation for grasping tasks

A former trend in the deep learning community was to build end-to-end deep learning models. While many researchers, such as [32] showed that it is possible to teach a model a very specific task from end to end, these models are not practical to use. They have to be trained using an in-house collected dataset for that very specific task and the very specific environment. They can also not be used for any other similar task.

That end-to-end trend evolved into rather divided problems where object detection and object grasping are two different modules. While this helped to integrate deep learning in the industry it still misses an important aspect, which is that deep learning models produced by researchers are not directly useful for industry usage. They have to go through a time-consuming tailoring process for each application, for each setup.

The current trend in deep learning is to divide the problems into even smaller modules. An example of such a problem is the unseen object segmentation problem, i.e., [33–36]. Instead of of the regular object segmentation of having a closed set of objects or goods that the deep learning model can segmented, unseen object segmentation focuses on building deep learning models that can segment any type of number of objects without classifying those objects.

While this work is going in the modularized direction it is still not clear how much of the work each module should do. In addition, only the visual information is used, neglecting other sources of information in the practical world. As discussed later, we propose a pipeline for the modularization and discuss how different information sources would help in the practical world.

2.2.3 Object pose estimation

In order to estimate the pose of objects in an indoor environment, multiple methods have been proposed in recent literature [6, 37–39]. One prominent approach is cosypose [6], where the authors rely on monocular images from multiple image streams to estimate the poses of objects within a scene. The approach initially detects regions where objects of interest are present, then uses the collected detections to estimate the pose from individual images. The poses from the camera streams are then aggregated with a joint bundle adjustment algorithm to produce a more accurate pose for the objects present. Other approaches such as [39] treat the pose estimation task as a classification problem. The approach relies on finding corresponding keypoints between multiple RGB image streams which are then paired with a 3D model of the object. Individual pose predictions are then obtained from each pairing. The generated pairings are then assimilated by a neural network which generates the final pose parameters.

The discussed approaches rely on large annotated datasets to achieve the pose estimation task. Such data is usually manually annotated in a tedious manner. The T-LESS dataset [40], for example, has been collected and annotated using a special setup comprised of an RGBD camera, a monocular RGB camera, and a turntable that moves the objects at predefined intervals. The objects were matched to their 3D model counterparts by manual alignment. Moreover, datasets that target the logistics domain are scarce. Being one of the few datasets in the logistics domain, the LOCO dataset [41] captures objects in realistic industrial scenarios and comprises objects that are commonly involved in material flow operations such as pallets and forklifts. The dataset contains 152,421 2D bounding box annotations that were manually added.

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

3 Methodological considerations

This section describes the methods used to put the approaches detailed in section 2.2 into practice. In doing so, the used software as well as the experimental setups will be described.

3.1 Re-identification methods

As discussed in Section II, different identification methods exist. The authors assume that the use of computer vision based deep learning methods allow for the reliable deployment of identification methods in the industry. As established, this is due to the adaptability of deep learning methods and the increasing ease with which large amounts of data can be gathered. Therefore, we propose the use of deep learning based re-identification methods for the purpose of re-identifying logistical entities. In this context, we have so far focused on the application of these methods on Euro-pallets and their chipwood pallet blocks. Due to the unique pattern obtained during the creation of these pallet blocks, they are well suited for the use as idiosyncratic features or fingerprints of their respective pallet.

For this specific task, our approach (as can be seen in Fig. 1) is the following: First, a pallet on a conveyor belt would pass by a set of two cameras, positioned at a 90° angle in relation to the conveyor belt. These cameras would provide an object detection model, trained on pallet blocks, with a continuous video stream. In this case, a YOLO based solution is employed, which detects pallet blocks, saves the frame in which the detection takes place and the pallet block is entirely in the field of view, and then crops these images to the predicted bounding boxes. Subsequently, these images are processed by a re-identification model, which works as a feature extractor. In this sense, the pallet block images are processed into a vectorized form, that contains the features of the respective image, that the model considers to be the defining attributes of the image. These vectors are then stored in a data frame, in which they can ultimately be compared to one another in terms of their similarity. Thus, the vectors that are considered to be the least dissimilar are matched and predicted to be images of the same pallet block, assuming that another image of the same pallet block already exists. In this context, the image that would be compared to the remaining images would be taken from a query set, while the remaining images would be taken from a gallery set. These sets can differ in their size and distribution. For the time being, the re-identification process is conducted under the assumption of a closed set. That is, when matching a query image with the gallery, it is assumed, that a correct match ought to be found. This means, that no novelty detection takes place at this point in time. The results of the re-identification process are evaluated through ranked accuracy and F-Scores, conforming to standard literature in state-of-the-art research.

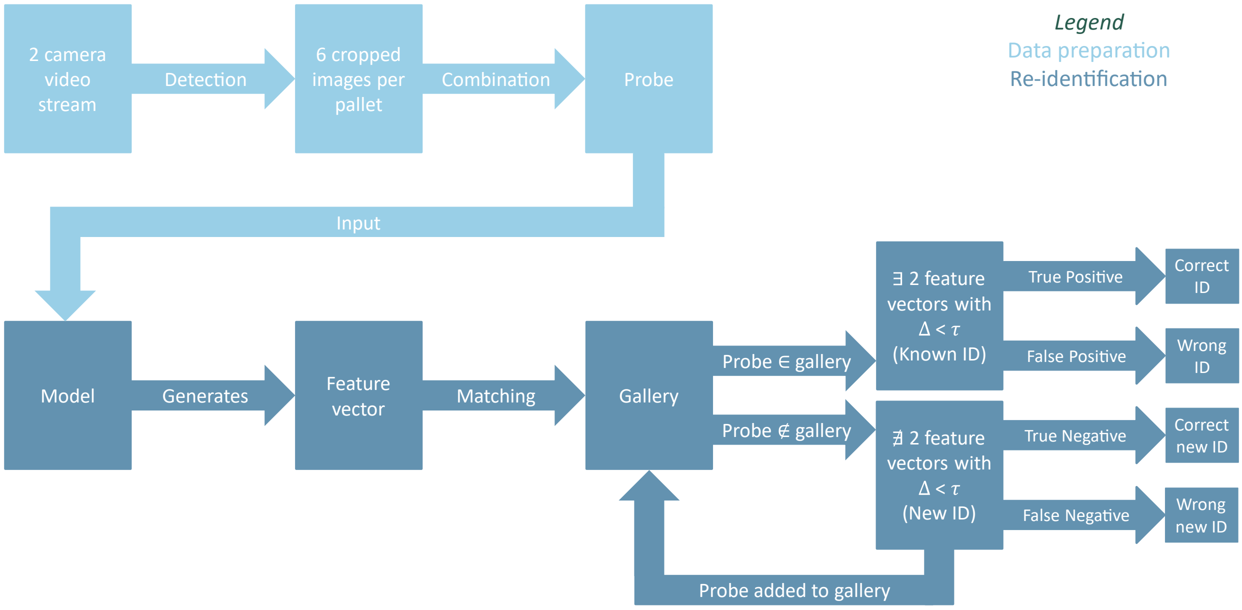

Fig. 1: The proposed pallet re-identification workflow

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

3.2 Object segmentation methods

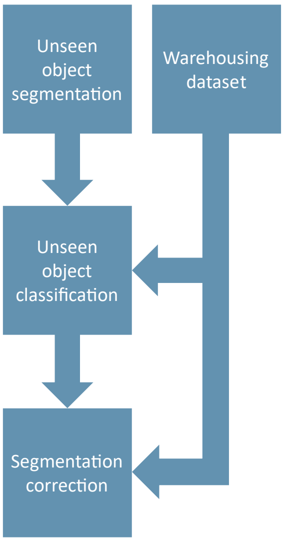

Fig. 2: The proposed pipeline for the object segmentation and classification task.

Computer vision problems can benefit from taking the logistical perspective into consideration. Object segmentation is one such problem, rethinking how the architecture of deep learning models can be modified as to improve two points. The first point is considering a priori information from other sources and the second one is the network design for practical deployment.

First, the most common case for object segmentation is to only use images for the segmentation and neglecting other sources of information. Warehousing databases are a source of such a priori information. Typically, such databases contain the precise number of objects and their types. An object segmentation pipeline can use such information to compare against the output of the model segmentation with the database and make a correction step. Therefore, if the number of the segmented masks does not match the number of objects in the database then a join and split method should be used to correct the segmentation such as the method proposed in [42]

The second point is to split the object segmentation into two steps (segmentation and classification). This can be achieved using the object types stored in the database. This modularization would help deep learning models to be directly exported from research to application with little to no effort. The segmentation network should be a model for unseen object segmentation. The classification model compares the segmented masks to a pre-stored image of an object in the warehouse database.

The reason for which a model for unseen object segmentation is needed is that a model that is trained for unseen object segmentation on different environments could be deployed to the industry nearly instantaneously but a model that is trained on a specific dataset of objects could only segment this set of objects and would require fine-tuning before deployment. Similarly deep learning models for unseen object classification could match those unclassified object masks to pre-captured images of the target object.

Fig. 2 shows the building blocks for the proposed segmentation pipeline. The pipeline keeps the two concepts in mind, first integrating the a priori knowledge to reduce deployment effort and take advantage of other information sources that can be used to improve the segmentation and verify its output and second to split the segmentation process into smaller modules so that models trained in research could be used in real world scenarios.

3.3 Object pose estimation methods

Deep learning techniques have recently shown significant performance improvement over classical methods, particularly in vision based applications. However, as discussed previously, the performance of deep learning techniques is heavily reliant on the existence, amount, and quality of the data used. Deep learning approaches often require large amounts of data that are manually annotated in a laborious manner. In this work, we automatically annotate data captured in large industrial settings using an RGB monocular camera system combined with a motion capturing system for the purpose of object pose estimation. Our method was used to collect and annotate a sample dataset.

Our automated annotation pipeline requires two sources of information, namely an RGB camera system for collecting images and 6-dimensional positions of the objects of interest such as those provided by a motion capturing system. At our research facility, a large motion capturing system, comprised of 48 cameras, covers a shop-floor-like area. The area is commonly perceived by an 8-RGB camera system. Our experimentation area is shown in Fig. 3.

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

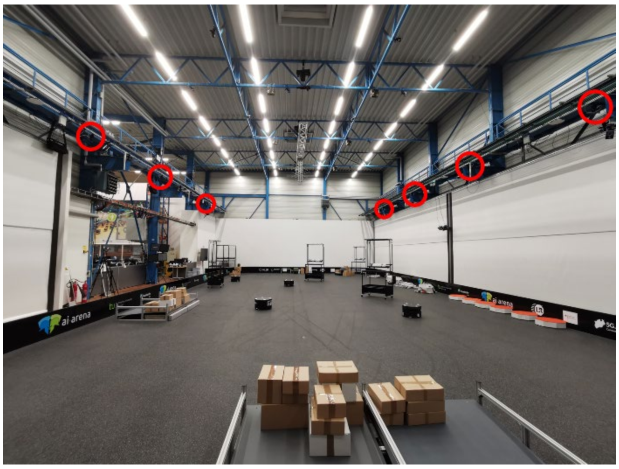

Fig. 3: Camera system installed at our research facility. Locations for seven of the eight installed cameras are circled in red.

Our approach to automated annotation is comprised of three main phases: camera calibration and localization, calculation of relative poses, and post-processing. Initially, each RGB camera is calibrated to obtain the intrinsic parameters. The cameras are then localized within the world frame of the motion capturing system which acts as a common frame for both systems. Since the object pose estimation task is concerned with the relative pose of the object with respect to the camera, the relative poses are calculated using a linear transformation chain. Finally, in the post processing phase, the annotation of the RGB images is generated by projecting the 3D models of the objects of interest onto the images using their ground truth poses obtained by the motion capturing system. We discuss the details of the three phases next.

First, individual RGB cameras are calibrated using the methods described in [43], using a checkerboard pattern. A range of 35 to 60 pattern images were taken per camera that vary heavily in terms of the distance to the camera. The resulting intrinsic parameters per camera are then stored. A modified version of the checkerboard pattern that is tracked by the motion capturing system is then used to localize the cameras. The tracking is done by attaching retro-reflective markers to the checkerboard pattern. The localization can then be performed by extrapolating the 3D points existing on the checkerboard grid in each of the captured frames. Since the 3D points are measured with respect to the motion capturing system's world frame, and since the 2D corresponding pixel locations are available from the intrinsic camera calibration step, the location of the camera in the world frame of the motion capturing system could be obtained by passing the sets of 3D and 2D points to the solvePnP algorithm [44].

Second, the obtained camera locations and their intrinsic parameters could be used to calculate the poses of the objects of interest with respect to the cameras, i.e, perform object pose estimation. Given the locations of the object as well as the cameras, in the global frame of the motion capturing system, the relative location of the object with respect to each of the cameras can be calculated using a linear transformation chain that links the motion capturing system, the objects of interest, and the camera under investigation. The linear transformation chain can be represented by the following equation:

$$H_{obj}^{cam} = (H_{cam}^{mc})^{-1} H_{obj}^{mc} \qquad (1)$$

where $H_{(destination)}^{(source)}$ represents the homogeneous transformation from a source frame to a destination frame. $cam, obj, and\ mc$ represent the camera, object, and motion capturing system frames, respectively. For example, H_{obj}^{cam} describes how the object frame is situated within the camera frame.

Finally, in the post-processing step, the annotations of the images are automatically generated. The annotations include image masks and 2D bounding boxes for objects of interest. Image masks are produced by rendering the pre-designed 3D models of the objects of interest at their relative poses with respect to the camera. The rendered models are then projected from the 3D space onto the image using the camera intrinsic parameters that were previously stored. The 3D points are projected to their pixel locations using the camera projection matrix [45] described by the following equations:

$$x = P\,X \qquad (2)$$

where X is a 4 x 1 vector of a point location in 3D space, x is a 3 x 1 vector of pixel locations, and P is the projection matrix defined as:

$$P = K\,[R \mid t] = K\,R\,[I \mid R^T\,t] \qquad (3)$$

where K is the 3 x 3 camera matrix describing the intrinsic parameters of the camera. R and t are the 3 x 3 rotation matrix and 3 x 1 translation vector of the ground truth pose of the object, respectively. A 2D bounding box is then fitted around the generated masks to produce the final annotations in a fully automated manner.

4 RESULTS AND ADVANCES

This section will present the current results and advances of the methods presented in the preceding section.

4.1 RE-IDENTIFICATION ADVANCES

So far, two datasets have been created that represent substantial first steps towards putting the re-identification of logistical entities into practice (see Fig. 4). The first dataset, pallet-block-502 [46], contains 5,002 images of pristine pallet blocks, provided by EPAL. 10 images per

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

pallet block were taken in five predetermined perspectives and with two different lighting conditions.

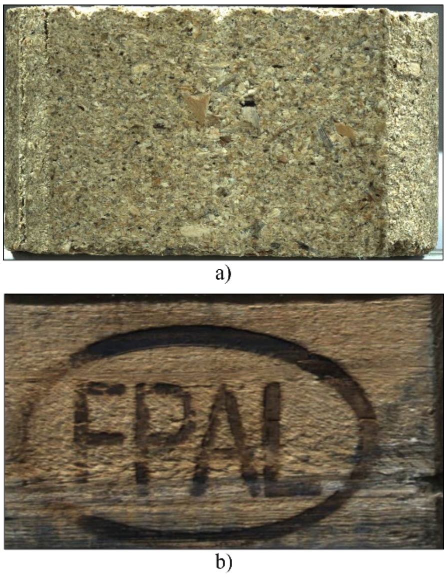

Fig. 4: Example images from both datasets a) ID 28 of pallet-block-502 and b) ID 3047 of pallet-block-32965.

The limitations of this dataset are its use of pristine pallet blocks and its limited size. Subsequently, a new dataset, pallet-block-32965 [47], was recorded in the warehouses of two major German companies. Here, 131,860 images were recorded over a span of multiple months. Two cameras were pointed at conveyor belts at each site. One camera was facing the conveyor belt at a 90° angle, while the other one was positioned at an angle of 120°. This second angle was chosen to have a visually different perspective of the pallet which could still be a realistic recording angle in some industrial scenarios. Per camera (2.4 MP resolution), two images were taken of each recorded pallet block, a couple of frames apart. The camera software parameters were constantly changing (exposure, gain, and level control) and the lighting conditions were changing in function of the lighting in the warehouses. A rendered representation of the laboratory recording setup can be seen in Fig. 5.

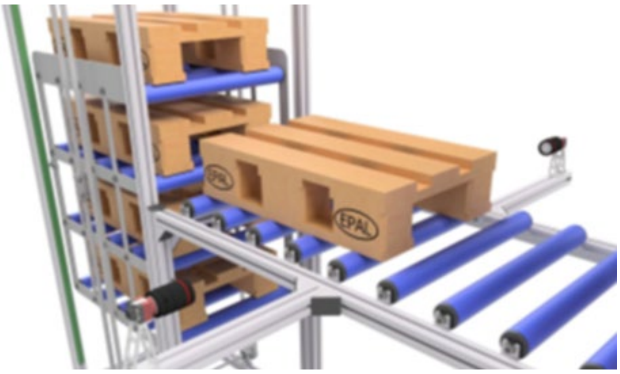

Fig. 5.: Rendering of the demonstrator used for the re-identification workflow (curtesy of Fraunhofer IML).

These two datasets have so far been used for data-driven re-identification approaches. This means, that they were used to train deep learning models, meant to subsequently re-identify pallet blocks. So far, using the abovementioned datasets, rank-1-accuracy scores upwards of 90% could be gathered. These scores were obtained using YOLO based object detection, the PyTorch framework TorchReID, and a residual convolutional neural network (PCB_P4 with a ResNet50 backbone). The experiments conducted so far did not include novelty detection but indicate that a reliable re-identification of closed sets of chipwood pallets is feasible.

4.2 Object segmentation advances

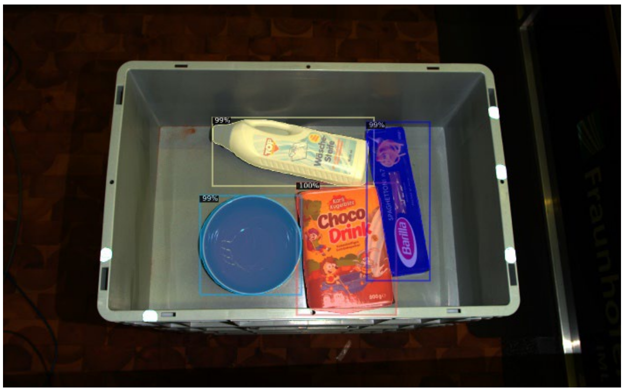

Fig. 6: Unseen objects segmented by our trained model.

Fig. 6 shows a segmentation obtained from our trained model for unseen object segmentation. The model uses a Mask R-CNN model implemented in the Detectron2 framework trained on a synthetic dataset and a real dataset collected by the authors. This segmentation represents the first step of the proposed processing pipeline. The second step in the process would be the classification of the object against all the objects listed in a pre-captured database of object images. As mentioned in the pipeline the database can be used to validate the output of the segmentation. This method would allow the same trained network as the one shown here to be used with any type of objects for bin picking or robotic grasping, eliminating any effort for re-training or fine-tuning.

4.3 Object pose estimation advances

The discussed automated annotation procedure was used to annotate a dataset collected at our research facility. The dataset is comprised of five object categories including pallets, cardboard boxes, movable workstations, mobile robots, and small load carriers. A total of nine instances for the objects were captured by all eight RGB cameras and automatically annotated. A sample overlaid projection of the 3D models for objects of interest in a sample scene is shown in Fig. 7. The grey silhouettes in the figure could be fitted to generate bounding box annotations. Methods mentioned in section 2.2.3 could then be used to estimate the poses of objects of interest.

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

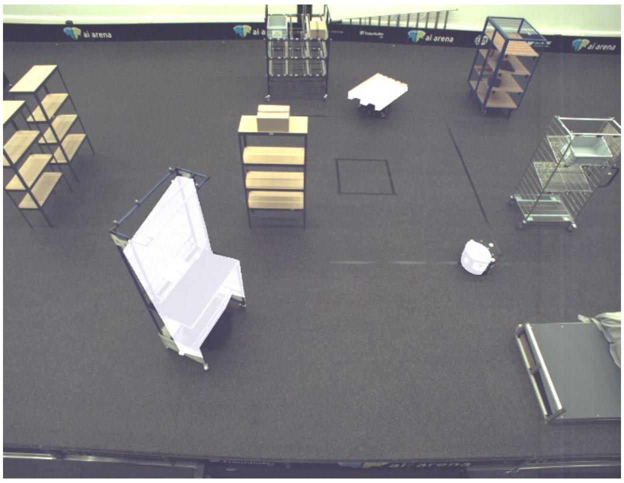

Fig. 7: Overlaid projection of three objects of interest (light grey) onto a single scene image from one camera.

5 Outlook and impact Review

In this contribution, we demonstrated the way in which three distinct deep learning based computer vision methods could be applied to warehousing logistics. These methods, once deployed, would permit the identification of specific logistical entities when entering and leaving the warehouse, could track them around the warehouse and give an estimate on their pose, and could finally handle and grasp before unseen entities, in a reliable manner. The preliminary results obtained during our research indicate, that these methods would pose a benefit to common warehousing environments, in which they could be deployed at a later point in time.

Given the approaches that are presented in this work and the results obtained thus far, we believe, that further development of these methods would be advantageous for the warehousing community. The reliability of such methods is of utmost importance in the industry and as such, more testing and validation has to be performed first. For re-identification, reliable exception handling processes have to be established, potentially involving human operators. This would be important, as to not wrongfully label, e.g., an outbound load unit. For object pose estimation, expanding our approach to other indoor scenes could increase the generalization capability and thus help avoid fingerprinting and overfitting on the current setup. In addition, tracking not only the materials transferred, but also human operators, would add new insights and understanding in terms of monitoring the processes. For object segmentation we showed how unseen object segmentation models can eliminate the need for fine-tuning of deep learning models. The pipeline requires other modules for unseen object classification and segmentation correction to be built. Having such modules would allow for the wide use of the exact same models in different warehouses and production facilities.

References

[1] G. Schuh, R. Anderl, R. Dumitrescu, A. Krüger, and M. ten Hompel, Eds. *Industrie 4.0 Maturity Index: Die digitale Transformation von Unternehmen gestalten,* 2020th ed. (acatech Studie). München, Berlin, Brüssel: acatech - Deutsche Akademie der Technikwissenschaften, 2020.

[2] A. Dekhne, G. Hastings, J. Murnane, and F. Neuhaus, *Automation in logistics: Big opportunity, bigger uncertainty*.

[3] Oxford Economics, *How robots change the world: What automation really means for jobs and productivity*.

[4] J. Rutinowski, C. Pionzewski, T. Chilla, C. Reining, and M. ten Hompel, "Towards Re-Identification for Warehousing Entities - A Work-in-Progress Study," in *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, Vasteras, Sweden, 2021, pp. 1–4, doi: 10.1109/ETFA45728.2021.9613250.

[5] A. Gouda, A. Ghanem, P. Kaiser, and M. ten Hompel, "Object class-agnostic segmentation for practical CNN utilization in industry," in *2021 6th International Conference on Mechanical Engineering and Robotics Research (ICMERR)*, Krakow, Poland, 2021, pp. 97–105, doi: 10.1109/ICMERR54363.2021.9680821.

[6] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "CosyPose: Consistent multi-view multi-object 6D pose estimation," 2020, doi: 10.48550/arXiv.2008.08465.

[7] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition," 2015, doi: 10.48550/arXiv.1505.00880.

[8] *Leitfaden für Materialflussuntersuchungen*, 2689, VDI Verein Deutscher Ingenieure e.V., Feb. 2019.

[9] *Handbuch Logistik,* 3rd ed. (VDI-/Buch]). Berlin, Heidelberg: Springer, 2008.

[10] M. ten Hompel, T. Schmidt, and J. Dregger, *Materialflusssysteme: Förder- und Lagertechnik,* 4th ed. (VDI-Buch). Berlin, Heidelberg: Springer Vieweg, 2018.

[11] Gütegemeinschaft Paletten e.V. "Sichere Lösungen für den Warenverkehr." https://gpal.epal-pallets.org/fileadmin/user_upload/ntg_package/NK_Deutschland_GPAL/03_Produktdownloads/GPAL_Produktmagazin_DE.pdf

Article is protected by German copyright law

[12] *DIN EN 13698-1:2004-01, Produktspezifikation für Paletten_- Teil_1: Herstellung von 800_mm x 1200_mm-Flachpaletten aus Holz; Deutsche Fassung EN_13698-1:2003*, Berlin.

[13] *KLEINLADUNGSTRÄGER (KLT)-SYSTEM*, 4500, Verband der Automobilindustrie e. V. (VDA), May. 2018.

[14] T. Bousonville, *Logistik 4.0: Die digitale Transformation der Wertschöpfungskette* (essentials). Wiesbaden: Springer Fachmedien Wiesbaden, 2017.

[15] GS1 Germany. "Allgemeine GS1 Spezifikationen." https://www.gs1-germany.de/fileadmin/gs1/fachpublikationen/allgemeine-gs1-spezifikation-v22.pdf (accessed Aug. 3, 2022).

[16] D. Fox, W. Burgard, H. Kruppa, and S. Thrun, "A Probabilistic Approach to Collaborative Multi-Robot Localization," *Autonomous Robots*, vol. 8, no. 3, pp. 325–344, 2000, doi: 10.1023/A:1008937911390.

[17] B.-S. Cho, W. Moon, W.-J. Seo, and K.-R. Baek, "A dead reckoning localization system for mobile robots using inertial sensors and wheel revolution encoding," *J Mech Sci Technol*, vol. 25, no. 11, pp. 2907–2917, 2011, doi: 10.1007/s12206-011-0805-1.

[18] D. Pizarro *et al.,* "Robot and obstacles localization and tracking with an external camera ring," in *2008 IEEE International Conference on Robotics and Automation*, Pasadena, CA, USA, 2008, pp. 516–521, doi: 10.1109/ROBOT.2008.4543259.

[19] A. Ravankar, A. Ravankar, Y. Kobayashi, and T. Emaru, "Intelligent Robot Guidance in Fixed External Camera Network for Navigation in Crowded and Narrow Passages," in *Proceedings of the 3rd International Electronic Conference on Sensors and Applications, 15–30 November 2016; Available online: https://sciforum.net/conference/ecsa-3*, 2017, p. 37, doi: 10.3390/ecsa-3-D008.

[20] J. H. Shim and Y. Im Cho, "A Mobile Robot Localization via Indoor Fixed Remote Surveillance Cameras," *Sensors (Basel, Switzerland)*, early access. doi: 10.3390/s16020195}.

[21] K. Stone, J. M. Keller, M. Popescu, T. C. Havens, and K. C. Ho, "Forward looking anomaly detection via fusion of infrared and color imagery," in *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XV*, Orlando, Florida, R. S. Harmon, J. H. Holloway, and J. T. Broach, Eds., 2010, p. 766425, doi: 10.1117/12.851370.

[22] S. D. Bansod and A. V. Nandedkar, "Crowd anomaly detection and localization using histogram of magnitude and momentum," *Vis Comput*, vol. 36, no. 3, pp. 609–620, 2020, doi: 10.1007/s00371-019-01647-0.

[23] H. Süße and E. Rodner, *Bildverarbeitung und Objekterkennung: Computer Vision in Industrie und Medizin* (Lehrbuch). Wiesbaden: Springer Vieweg, 2014.

[24] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep Learning for Person Re-Identification: A Survey and Outlook," *IEEE transactions on pattern analysis and machine intelligence*, early access. doi: 10.1109/TPAMI.2021.3054775.

[25] K. Finkenzeller, *RFID-Handbuch: Grundlagen und praktische Anwendungen von Transpondern, kontaktlosen Chipkarten und NFC,* 7th ed. (Hanser eLibrary). München: Hanser, 2015.

[26] T. Takahashi and R. Ishiyama, "FIBAR: Fingerprint Imaging by Binary Angular Reflection for Individual Identification of Metal Parts," in *2014 Fifth International Conference on Emerging Security Technologies*, Alcala de Henares, Spain, 2014, pp. 46–51, doi: 10.1109/EST.2014.25.

[27] A. Sharma, L. Subramanian, and E. A. Brewer, "PaperSpeckle," in *Proceedings of the 18th ACM conference on Computer and communications security - CCS '11*, Chicago, Illinois, USA, Y. Chen, G. Danezis, and V. Shmatikov, Eds., 2011, p. 99, doi: 10.1145/2046707.2046721.

[28] Di Wu *et al.,* "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, pp. 354–371, 2019, doi: 10.1016/j.neucom.2019.01.079.

[29] Z. Cheng, X. Zhu, and S. Gong, "Face re-identification challenge: Are face recognition models good enough?," *Pattern Recognition*, vol. 107, p. 107422, 2020, doi: 10.1016/j.patcog.2020.107422.

[30] J. C. Hermanson and A. C. Wiedenhoeft, "A brief review of machine vision in the context of automated wood identification systems," *IAWA J*, vol. 32, no. 2, pp. 233–250, 2011, doi: 10.1163/22941932-90000054.

[31] S. Schneider, G. W. Taylor, S. Linquist, and S. C. Kremer, "Past, present and future approaches using computer vision for animal re-identification from camera trap data," *Methods Ecol Evol*, vol. 10, no. 4, pp. 461–470, 2019, doi: 10.1111/2041-210X.13133.

Article is protected by German copyright law

DOI: 10.2195/lj_proc_rutinowski_en_202211_01
URN: urn:nbn:de:0009-14-56018

[32] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, 4-5, pp. 421–436, 2018, doi: 10.1177/0278364917710318.

[33] M. Danielczuk *et al.*, "Segmenting Unknown 3D Objects from Real Depth Images using Mask R-CNN Trained on Synthetic Data," in *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, QC, Canada, 2019, pp. 7283–7290, doi: 10.1109/ICRA.2019.8793744.

[34] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen Object Instance Segmentation for Robotic Environments," *IEEE Trans. Robot.*, vol. 37, no. 5, pp. 1343–1359, 2021, doi: 10.1109/TRO.2021.3060341.

[35] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, "Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation," 2020, doi: 10.48550/arXiv.2007.15157.

[36] S. Back *et al.,* "Unseen Object Amodal Instance Segmentation via Hierarchical Occlusion Modeling," in *2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, 2022, pp. 5085–5092, doi: 10.1109/ICRA46639.2022.9811646.

[37] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," Nov. 2017.

[38] B. Tekin, S. N. Sinha, and P. Fua, "Real-Time Seamless Single Shot 6D Object Pose Prediction," Nov. 2017.

[39] J. N. Kundu, M. V. Rahul, A. Ganeshan, and R. V. Babu, "Object Pose Estimation from Monocular Image Using Multi-view Keypoint Correspondence," in *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part III* (SpringerLink Bücher 11131), L. Leal-Taixé and S. Roth, Eds., Cham: Springer International Publishing, 2019, pp. 298–313.

[40] T. Hodan, P. Haluza, S. Obdrzalek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, 2017, pp. 880–888, doi: 10.1109/WACV.2017.103.

[41] C. Mayershofer, D.-M. Holm, B. Molter, and J. Fottner, "LOCO: Logistics Objects in Context," in *19th IEEE International Conference on Machine Learning and Applications: ICMLA 2020 : 14-17 December 2020, virtual event : proceedings*, M. A. Wani, Ed., Piscataway, NJ: IEEE, 2020, pp. 612–617.

[42] C. Xie, A. Mousavian, Y. Xiang, and D. Fox, *RICE: Refining Instance Masks in Cluttered Environments with Graph Neural Networks*, 2021.

[43] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 11, pp. 1330–1334, 2000, doi: 10.1109/34.888718.

[44] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," *Int J Comput Vis*, vol. 81, no. 2, pp. 155–166, 2009, doi: 10.1007/s11263-008-0152-6.

[45] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press, 2003.

[46] J. Rutinowski, T. Chilla, C. Pionzewski, C. Reining, and M. ten Hompel, "pallet-block-502 – A chipwood re-identification dataset," 2021, doi: 10.5281/zenodo.6353714.

[47] J. Rutinowski, T. Chilla, and C. Pionzweski, "pallet-block-32965 – A chipwood re-identification dataset," 2022, doi: 10.5281/zenodo.6358607.

Jérôme Rutinowski, M.Sc., Research Assistant at the Chair of Material Handling and Warehousing, TU Dortmund University. The focus of Mr. Rutinowski's research lies in re-identification of logistical entities.
Phone: +49 231 755-4831
E-Mail: jerome.rutinowski@tu-dortmund.de

Hazem Youssef, M.Sc., Research Assistant at the Chair of Material Handling and Warehousing, TU Dortmund University. The focus of Mr. Youssef's research lies in object pose estimation.
Phone: + +49 231 755-3450
E-Mail: hazem.youssef@tu-dortmund.de

Anas Gouda, M.Sc., Research Assistant at the Chair of Material Handling and Warehousing, TU Dortmund University. The focus of Mr. Gouda's work lies in unseen object segmentation.
Phone: +49 231 755-2545
E-Mail: anas.gouda@tu-dortmund.de

Article is protected by German copyright law

Dr.-Ing. Christopher Reining, Chief Scientist at the Chair of Material Handling and Warehousing, TU Dortmund University. Dr. Reining was involved in this work as a mentor and counselor.
Phone: +49 231 755-3228
E-Mail: christopher.reining@tu-dortmund.de

Dipl.-Inform. Moritz Roidl, Chief Engineer at the Chair of Material Handling and Warehousing, TU Dortmund University. Mr. Roidl was involved in this work as a mentor and counselor.
Phone: +49 231 755-3092
E-Mail: moritz.roidl@tu-dortmund.de

Address: Chair of Material Handling and Warehousing, TU Dortmund University, Joseph-von-Fraunhofer-Str. 2-4, 44227 Dortmund, Germany,