In: RGB-D Image Analysis and Processing, pp. 285-308, Springer, December 2019.

Chapter 13
Semantic RGB-D Perception for Cognitive Service Robots

Max Schwarz and Sven Behnke

Abstract Cognitive robots need to understand their surroundings not only in terms of geometry, but they also need to categorize surfaces, detect objects, estimate their pose, etc. Due to their nature, RGB-D sensors are ideally suited to many of these problems, which is why we developed efficient RGB-D methods to address these tasks. In this chapter, we outline the continuous development and usage of RGB-D methods, spanning three applications: Our cognitive service robot Cosero, which participated with great success in the international RoboCup@Home competitions, an industrial kitting application, and cluttered bin picking for warehouse automation. We learn semantic segmentation using convolutional neural networks and random forests and aggregate the surface category in 3D by RGB-D SLAM. We use deep learning methods to categorize surfaces, to recognize objects and to estimate their pose. Efficient RGB-D registration methods are the basis for the manipulation of known objects. They have been extended to non-rigid registration, which allows for transferring manipulation skills to novel objects.

13.1 Introduction

The need for truly *cognitive* robots, i.e. robots that can react to and reason about their environment, has been made very clear in recent years. Applications like personal service robots, elderly care, guiding robots, all require higher levels of cognition than what is available today. But also classical domains of robotics, like industrial automation, will benefit greatly from smarter robots which truly relieve the load of their human coworkers.

Autonomous Intelligent Systems, Computer Science Institute VI
University of Bonn, Germany
schwarz@ais.uni-bonn.de, behnke@cs.uni-bonn.de
http://www.ais.uni-bonn.de

A key stepping stone towards higher cognitive function is environment perception. The ready availability of affordable RGB-D sensors, starting with the Microsoft Kinect, now encompassing a multitude of sensors with different properties, has sparked the development of many new perception approaches. Especially in the robotics community, which is not only interested with *perceiving* the environment, but also especially *interacting* with it, the direct combination of color information with geometry offers large advantages over classical sensors which capture the modalities separately.

The interest in our group in RGB-D sensors started with our work in the field of cognitive service robots. An increasing number of research groups worldwide are working on complex robots for domestic service applications. Autonomous service robots require versatile mobile manipulation and human-robot interaction skills in order to really become useful. For example, they should fetch objects, serve drinks and meals, and help with cleaning. The everyday tasks that we perform in our households are highly challenging to achieve with a robotic system, though, because the environment is complex, dynamic, and structured for human rather than robotic needs.

We have developed cognitive service robots since 2008, according to the requirements of the annual international RoboCup@Home competitions [72]. These competitions benchmark integrated robot systems in predefined test procedures and in open demonstrations within which teams can show the best of their research. Benchmarked skills comprise mobility in dynamic indoor environments, object retrieval and placement, person perception, complex speech understanding, and gesture recognition.

Starting from the methods developed for our Cognitive service robot Cosero, described in Section 13.3, we will show how proven RGB-D methods and key ideas were carried over to subsequent robotic systems in other applications, such as industrial kitting (Section 13.4) and cluttered bin picking for warehouse automation (Section 13.5).

13.2 Related Work

Service Robots Prominent examples of service robots include Armar [1], developed at KIT, that has demonstrated mobile manipulation in a kitchen environment [68]. The Personal Robot 2 (PR2 [40]), developed by Willow Garage, popularized the Robot Operating System (ROS [47]) that is used by many research groups. It is equipped with two 7-DOF compliant arms on a liftable torso. For mobility, the robot drives on four individually steerable wheels, similar to our Cosero robot. PR2 perceives its environment using 2D and 3D laser scanners, and a structured light RGB-D sensor in the head. Bohren et al. [9] demonstrated fetching drinks from a refrigerator and delivering them to users with the PR2 platform. Beetz et al. [6] used a PR2 and a custom-built robot to cooperatively prepare pancakes.

Another example is Rollin' Justin [10], developed at DLR. Similarly, it is equipped with two compliant arms and a four-wheeled mobile base. The robot demonstrated several dexterous manipulation skills such as making coffee by operating a pad machine [5] and cleaning windows [36]. Further examples are HoLLie [23], developed at FZI Karlsruhe, and Care-O-Bot 4 [31], recently introduced by Fraunhofer IPA.

The RoboCup Federation holds annual competitions in its @Home league [28], which serve as a general benchmark for service robots. Since research labs usually focus on narrow tasks, this competition is especially important for guiding and evaluating the research on service robotics in a more holistic perspective. Systems competing in the 2017 edition, which was held in Nagoya, Japan, are described in the corresponding team description papers [11, 41, 69]. Most of these custom-designed robots consist of a wheeled mobile base with LiDAR and RGB-D sensors and a single manipulator arm, although humanoid shapes with two arms are becoming more common. Notably, RGB-D sensors play a large role in the competition, since they offer highly semantic environment understanding (see [11, 28]) at very low cost.

Mapping In order to act in complex indoor environments, service robots must perceive the room structure, obstacles, persons, objects, etc. Frequently, they are equipped with 2D or 3D laser scanners to measure distances to surfaces. Registering the laser measurements in a globally consistent way yields environment maps. Graph optimization methods [67] are often used to solve the simultaneous localization and mapping (SLAM) problem. Efficient software libraries are available to minimize the registration error [29, 34]. 2D maps represent walls and obstacles only at the height of a horizontal scan plane [38]. If 3D laser scanners are used [55, 74], the full 3D environment structure can be modeled.

In recent years, RGB-D cameras (see Chapter 1) became available to measure geometry and colored texture of surfaces in smaller indoor environments. Registering these measurements yields colored 3D environment models (see Chapter 5, [15, 30, 70, 71]).

Semantic Perception In addition to modelling the environment geometry and appearance, semantic perception is needed for many tasks. This involves the categorization of surfaces, the detection and recognition of objects and the estimation of their pose. Surface categorization is also known as object-class segmentation. The task is to assign a class label to every pixel or surface element. For example, Hermans et al. [24] train random decision forests to categorize pixels in RGB-D frames. They estimate camera motion and accumulate pixel decisions in a 3D semantic map. Spatial consistency is enforced by a pairwise Conditional Random Field (CRF).

In contrast, Eigen et al. [14] process single frames at multiple resolutions. They train convolutional neural networks (CNN) to predict depth, surface normals, and semantic labels. The network is initialized with pre-trained features [32]. Long et al. [37] combined upsampled predictions from intermediate layers with a final full-resolution layer which leads to more refined results. A whole-image classification network was adapted to a fully convolutional network and finetuned for semantic segmentation. Another example of a convolutional architecture for semantic seg-

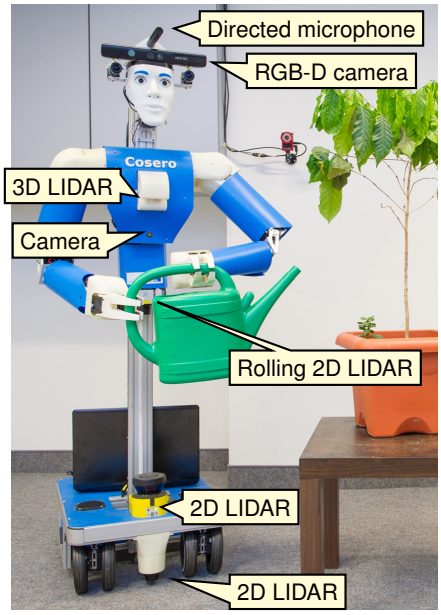
mentation is the work of Badrinarayanan et al. [3]. They use a multi-stage encoder-decoder architecture that first reduces spatial resolution through maximum pooling and later uses the indices of the local pooling maxima for non-linear upsampling to produce class labels at the original resolution.

For the detection of objects, e.g., implicit shape models [35] and Hough forests [18] have been proposed. In recent years, CNNs have also been successfully used for the detection of objects in complex scenes. Girshick et al. [21], for example, use a bottom-up method for generating category-independent region proposals and train a CNN to categorize size-normalized regions. To accelerate detection, all regions are processed with a single forward pass of the CNN [20]. Another line of research is to directly train CNNs to regress object bounding boxes [16, 54]. Ren et al. [48] developed a region proposal network (RPN) that regresses from anchors to regions of interest. More methods are discussed in Chapter 8.

For estimating the pose of objects in 3D data, often voting schemes are used. Drost et al. [13] and Papazov et al. [45] proposed point pair features, defined by two points on surfaces and their normals, which vote for possible object poses. This approach has been recently extended by Choi et al. [12] to incorporate color information from RGB-D sensors. In recent years, CNNs also have been trained to estimate object pose [4, 66]. 3D convolutional neural networks have been used for modeling, detection, and completion of 3D shapes [73]. For an in-depth review of 6D pose estimation methods, we refer to Chapter 11.

13.3 Cognitive Service Robot Cosero

Since 2008, the Autonomous Intelligent Systems group at University of Bonn has been developing cognitive service robots for domestic service tasks [62]. According to the requirements of the RoboCup@Home competitions, we developed the cognitive service robot *Cosero*, shown in Fig. 13.1, that balances the aspects of robust mobility, human-like manipulation, and intuitive human-robot-interaction. The robot is equipped with an anthropomorphic torso and two 7 DoF arms that provide adult-like reach and support a payload of 1.5 kg each. The grippers consist of two pairs of Festo FinGripper fingers on rotary joints, which conform to grasped objects. Cosero's torso can be twisted around and lifted along the vertical axis to extend its workspace, allowing the robot to grasp objects from a wide range of heights—even from the floor. Its narrow base moves on four pairs of steerable wheels that provide omnidirectional driving. For perceiving its environment, Cosero is equipped with multimodal sensors. Four laser range scanners on the ground, on top of the mobile base, and in the torso (rollable and pitchable) measure distances to objects, persons, or obstacles for navigation purposes. The head is mounted on a pan-tilt joint and features a Microsoft Kinect RGB-D camera for object and person perception in 3D and a directed microphone for speech recognition. A camera in the torso provides a lateral view onto objects in typical manipulation height. Cosero is controlled by a



- Environment Perception
  - 2D/3D Laser-based Mapping and Localization
  - RGB-D SLAM
  - Motion Segmentation
  - Semantic Segmentation
- Object Perception
  - Object Segmentation
  - Object Recognition
  - Object Detection and Pose Estimation
  - Primitive-based Object Detection
  - Object Tracking
  - Non-rigid Object Registration
- Perception of Persons
  - Person Detection and Tracking
  - Person Identification
  - Gesture Recognition
  - Speech Recognition

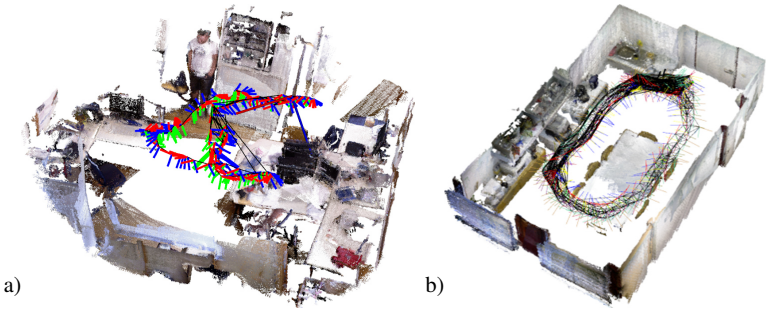
Fig. 13.1: Cognitive service robot *Cosero* with sensors marked and perceptional modules.

high-performance Intel Core-i7 quad-core notebook, located on the rear part of the base.

13.3.1 Environment Perception

RGB-D SLAM For modelling 3D geometry and appearance of objects, we developed an efficient RGB-D-SLAM method, based on Multi-Resolution Surfel Maps (MRSMaps [60]). The key idea is to represent the distribution of points in voxels and their color using a Gaussian. For registering RGB-D views, local multiresolution is used, i.e., the vicinity of the sensor is modeled in more detail than further-away parts of the environment. Graph optimization [34] is used to globally minimize registration error between key views. Fig. 13.2a shows a resulting map of an indoor scene. To reduce the need for sensor motion and to avoid looking only into free space, we constructed a sensor head with four RGB-D cameras that view four orthogonal directions [56]. Fig. 13.2b shows a map of a room that has been created by moving this multi-sensor in a loop.

Motion Segmentation RGB-D SLAM assumes static scenes. By modeling multiple rigid bodies as MRSMap and estimating their relative motion by expectation-maximization (EM), a dense 3D segmentation of the dynamic scene is obtained [61].



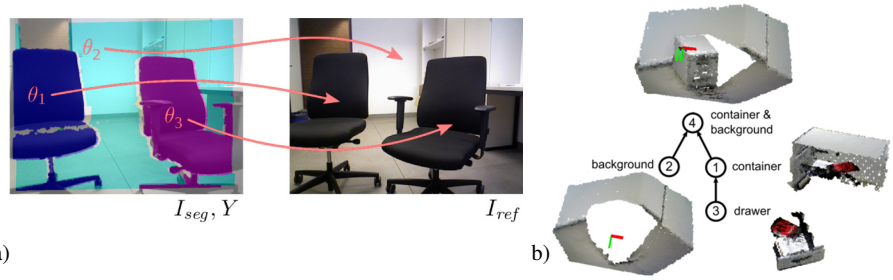
a) b)

Fig. 13.2: RGB-D SLAM: a) Multi-resolution surfel map obtained by registering RGB-D views [60]; b) RGB-D map of a room obtained from four moving RGB-D cameras [56].

Fig. 13.3a shows an example. From common and separate motion, a hierarchy of moving segments can be inferred [57], as shown in Fig. 13.3b.

Semantic Segmentation We developed several approaches for object-class segmentation. One method is using random forests (RF) to label RGB-D pixels [51] based on rectangular image regions that are normalized in size and position by depth and computed efficiently from integral images. Both training and recall have been accelerated by GPU. To obtain a 3D semantic map, we estimate camera motion by RGB-D SLAM and accumulate categorizations in voxels [65].

We developed a method to smooth the noisy RF pixel labels that is illustrated in Fig. 13.4a. It over-segments the scene in RGB-D superpixels and learns relations between them that are modeled as a Conditional Random Field (CRF), based on pair-wise features such as color contrast and normal differences. We also proposed CNN-based methods for semantic segmentation [25, 27, 49], with innovations, such as additional input features derived from depth, like height above ground [50] or distance from wall [27] (Fig. 13.4b)), and size-normalization of covering windows from depth [50].



a) b)

Fig. 13.3: Motion segmentation: a) Three rigid bodies and their motion modeled as MRSMap [61]; b) Motion hierarchy inferred from common/separate motions [57].

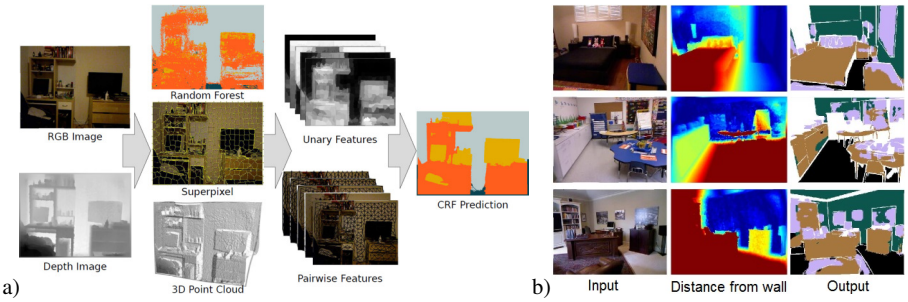


Fig. 13.4: Semantic segmentation: a) Random forest labeling is refined by a superpixel-CRF [42]; b) CNN segmentation based on semantic and geometric features [27].

For temporal integration, we directly trained the Neural Abstraction Pyramid [7]—a hierarchical, recurrent, convolutional architecture for learning image interpretation (Fig. 13.5a)—for object class segmentation of RGB-D video sequences [46]. It learns to recursively integrate semantic decisions over time. Fig. 13.5b shows an example result.

13.3.2 Object Perception

When attempting manipulation, our robot captures the scene geometry and appearance with its RGB-D camera. In many situations, objects are located well separated on horizontal support surfaces, such as tables, shelves, or the floor. To ensure good visibility, the camera is placed at an appropriate height above and distance from the surface, pointing downwards with an angle of approximately 45°. To this end, the

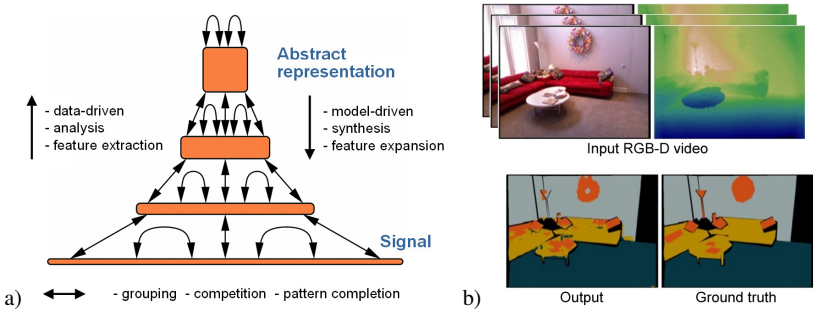


Fig. 13.5: Recurrent temporal integration for semantic segmentation: a) Neural Abstraction Pyramid (NAP) architecture [7]; b) NAP-based semantic segmentation [46].

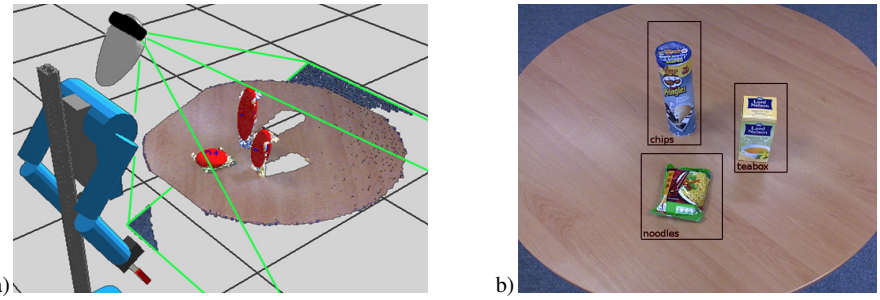


Fig. 13.6: Object perception: a) RGB-D view of a tabletop scene. Detected objects are represented by a fitted red ellipse; b) Recognized objects.

robot aligns itself with tables or shelves using the rollable laser scanner in its hip in its vertical scan plane position. Fig. 13.6a shows a scene.

Object Segmentation An initial step for the perception of objects in these simple scenes is to segment the captured RGB-D images into support planes and objects on these surfaces. Our plane segmentation algorithm rapidly estimates normals from the depth images of the RGB-D camera and fits a horizontal plane through the points with roughly vertical normals by RANSAC [64]. The points above the detected support plane are grouped to object candidates based on Euclidean distance. All points within a range threshold form a segment that is analyzed separately. In Fig. 13.6a, the detected segments are shown.

Object Detection and Pose Estimation For the detection and pose estimation of objects in complex RGB-D scenes, we developed a Hough Forest [18] based approach [2] that is illustrated in Fig. 13.7a. Decision trees do not only learn to categorize pixels, but also vote for object centers in 3D. Each detected object votes for object orientations, which yields detection of objects with the full 3D pose. Fig. 13.7b illustrates an extension of a saliency-based object discovery method [19],

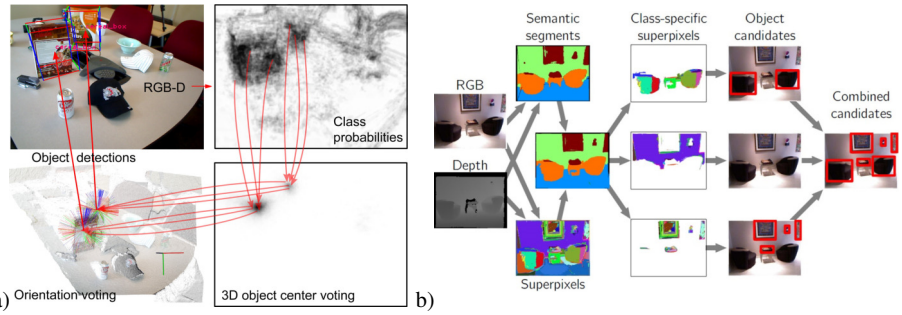


Fig. 13.7: 3D Object detection: a) 6D object detection using Hough forest [2]; b) Generating object proposals separately in semantic channels [19].

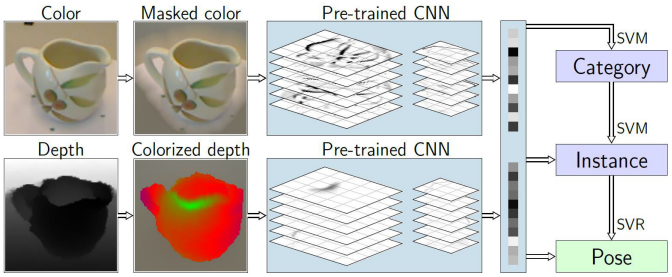
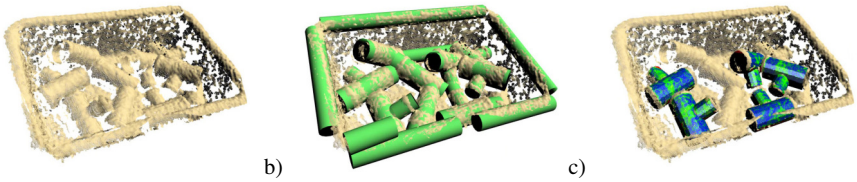


Fig. 13.8: Object categorization, instance recognition, and pose estimation based on features extracted by a pretrained CNN [53]. Depth is converted to a color image by rendering a canonical view and encoding distance from the object vertical axis.

which groups RGB-D superpixels based on semantic segmentation [27] and detects objects per class. This improves the generated object proposals.

For categorizing objects, recognizing known instances, and estimating object pose, we developed an approach that analyzes an object which has been isolated using table-top segmentation. The RGB-D region of interest is preprocessed by fading out the background of the RGB image (see Fig. 13.8 top left). The depth measurements are converted to an RGB image as well by rendering a view from a canonical elevation and encoding distance from the estimated object vertical axis by color, as shown in Fig. 13.8 bottom left. Both RGB images are presented to a convolutional neural network, which has been pretrained on the ImageNet data set for categorization of natural images. This produces semantic higher-layer features, which are concatenated and used to recognize object category, object instance, and to estimate the azimuth viewing angle onto the object using support vector machines and support vector regression, respectively. This transfer learning approach has been evaluated on the Washington RGB-D Object data set and improved the state-of-the-art [53].

Primitive-based Object Detection Objects are not always located on horizontal support surfaces. For a bin picking demonstration, we developed an approach to detect known objects which are on top of a pile, in an arbitrary pose in transport boxes. The objects are described by a graph of shape primitives. Fig. 13.9 illustrates the



a) b) c)

Fig. 13.9: Object detection based on geometric primitives [44]: a) Point cloud captured by Cosero's Kinect camera; b) Detected cylinders; c) Detected objects.

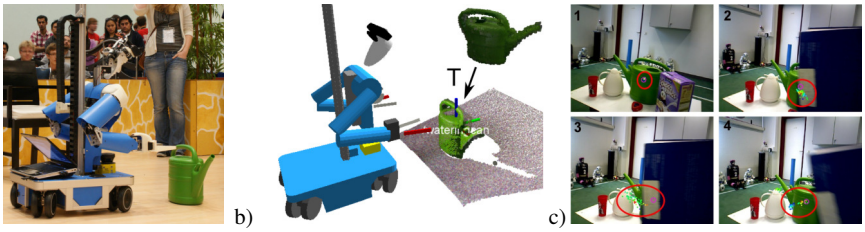


Fig. 13.10: Object tracking: a) Cosero approaching a watering can; b) A multi-view 3D model of the watering can (MRSMap, upper right) is registered with the current RGB-D frame to estimate its relative pose T, which is used to approach and grasp it; c) Joint object detection and tracking using a particle filter, despite occlusion.

object detection process. First, individual primitives, like cylinders of appropriate diameter are detected using RANSAC. The relations between these are checked. If they match the graph describing the object model, an object instance is instantiated, verified and registered to the supporting 3D points. This yields object pose estimates in 6D. Based on this, mobile bin picking has been demonstrated with Cosero [44]. The method has been extended to the detection of object models that combine 2D and 3D shape primitives [8].

Object Tracking Cosero tracks the pose of known objects using models represented as multi-resolution surfel maps (MRSMaps, [60]), which we learn from moving an RGB-D sensor around the object and performing SLAM. Our method estimates the camera poses by efficiently registering RGB-D key frames. After loop closing and globally minimizing the registration error, the RGB-D measurements are represented in a multiresolution surfel grid, stored as an octree. Each volume element represents the local shape of its points as well as their color distribution by a Gaussian. Our MRSMaps also come with an efficient RGB-D registration method which we use for tracking the pose of objects in RGB-D images. The object pose can be initialized using our planar segmentation approach. Fig. 13.10a,b) illustrates the tracking with an example. To handle difficult situations, like occlusions, we extended this approach to joint detection and tracking of objects modeled as MRSMaps using a particle filter [39] (see Fig. 13.10c).

Non-rigid Object Registration To be able to manipulate not only known objects, but also objects of the same category that differ in shape and appearance, we extended the coherent point drift method (CPD) [43] to efficiently perform deformable registration between dense RGB-D point clouds (see Fig. 13.11a). Instead of processing the dense point clouds of the RGB-D images directly with CPD, we utilize MRSMaps to perform deformable registration on a compressed measurement representation [59]. The method recovers a smooth displacement field which maps the surface points between both point clouds. It can be used to establish shape correspondences between a partial view on an object in a current image and a MRSMap object model. From the displacement field, the local frame transformation (i.e., 6D

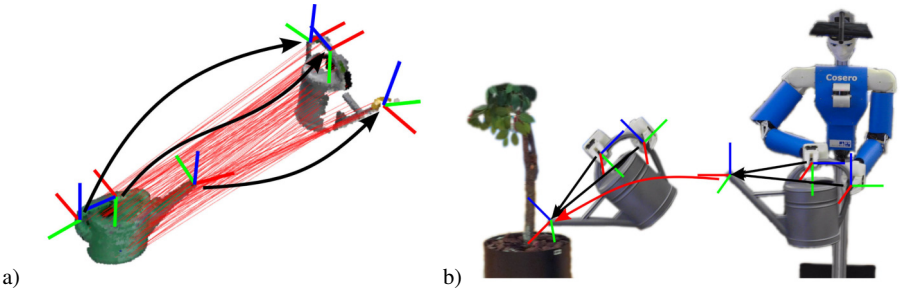


Fig. 13.11: Object manipulation skill transfer: a) An object manipulation skill is described by grasp poses and motions of the tool tip relative to the affected object; b) Once these poses are known for a new instance of the tool, the skill can be transferred.

rotation and translation) at a point on the deformed surface can be estimated. By this, we can determine how poses such as grasps or tool end-effectors change by the deformation between objects (Fig. 13.11b).

13.3.3 Robot Demonstrations at RoboCup Competitions

The developed perceptual components for the robot environment and workspace objects were the basis for many demonstrations of in RoboCup@Home league competitions [72], the top venue for benchmarking domestic service robots.

Mobile Manipulation Several predefined tests in RoboCup@Home include object retrieval and placement. We often used open challenges to demonstrate further object manipulation capabilities. For example, in the RoboCup 2011 *Demo Challenge*, Cosero was instructed where to stow different kinds of laundry, picked white laundry from the floor (Fig. 13.12a), and put it into a basket. In the final round, our robot demonstrated a cooking task. It moved to a cooking plate to switch it on. For this, we applied our real-time object tracking method (Sec. 13.3.2) in order to approach

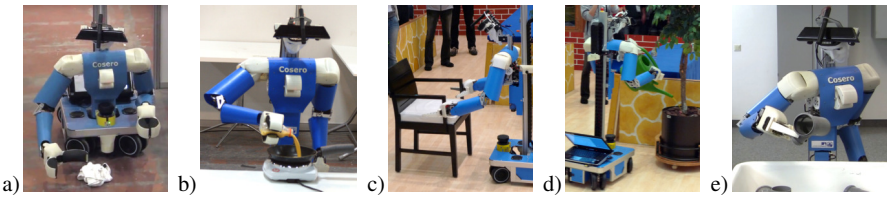


Fig. 13.12: Mobile manipulation demonstrations: a) Picking laundry from the floor; b) Cooking an omelette; c) Pushing a chair; d) Watering a plant; e) Bin picking.

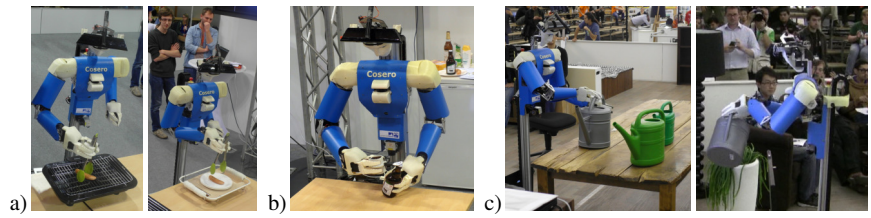


Fig. 13.13: Tool use demonstrations: a) Grasping sausages with a pair of tongs. b) Bottle opening; c) Plant watering skill transfer to unknown watering can.

the cooking plate and to estimate the switch grasping pose. Then, Cosero drove to the location of the dough and grasped it. Back at the cooking plate, it opened the bottle by unscrewing its lid and poured its contents into the pan (Fig. 13.12b).

In the RoboCup 2012 final, Cosero demonstrated the approaching, bi-manual grasping, and moving of a chair to a target pose (Fig. 13.12c). It also approached and grasped a watering can with both hands and watered a plant (Fig. 13.12d). Both were realized through registration of learned 3D models of the objects (Sec. 13.3.2). The robot also demonstrated our bin picking approach, which is based on primitive-based object detection and pose estimation (Fig. 13.12e).

Tool Use In the RoboCup 2013 *Open Challenge*, Cosero demonstrated tool-use skill transfer based on our deformable registration method (Sec. 13.3.2). The jury chose one of two unknown cans. The watering skill was trained for a third instance of cans before. Cosero successfully transferred the tool-use skill and executed it (Fig. 13.13c). In the final, Cosero demonstrated grasping of sausages with a pair of tongs (Fig. 13.13a). The robot received the tongs through object hand-over from a team member. It coarsely drove behind the barbecue that was placed on a table by navigating in the environment map and tracked the 6-DoF pose of the barbecue using MRSMaps (Sec. 13.3.2) to accurately position itself relative to the barbecue. It picked one of two raw sausages from a plate next to the barbecue with the tongs and placed it on the barbecue. While the sausage was grilled, Cosero handed the tongs back to a human and went to fetch and open a beer. It picked the bottle opener from a shelf and the beer bottle with its other hand from a table. Then it executed a bottle opening skill [58] (Fig. 13.13b).

In the RoboCup 2014 final, Cosero grasped a dustpan and a swab in order to clean some dirt from the floor (Fig. 13.14c). After pouring out the contents of the dustpan into the dustbin, it placed the tools back on a table and started to make caipirinha. For this, it used a muddler to muddle lime pieces (Fig. 13.14d).

Cosero also demonstrated awareness and interaction with humans (Fig. 13.14). Since the methods for these capabilities mainly use LIDAR tracking and RGB computer vision techniques and are thus out of scope for this chapter, we refer to [63] for details.

Competition Results We participated in four international RoboCup@Home and four RoboCup German Open @Home competitions 2011-2014. Our robot systems

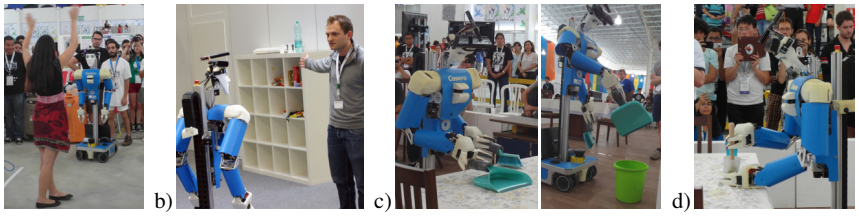


Fig. 13.14: Human-robot interaction and tool use: a) Following a guide through a crowd; b) Recognizing pointing gestures; c) Using a dustpan and a swab; d) Using a muddler.

performed consistently well in the predefined tests and our open demonstrations convinced the juries which consisted of team leaders, members of the executive committee, and representatives of the media, science, and industry. Our team NimbRo won three international competitions 2011-2013 and four German Open competitions 2011-2014 in a row and came in third at RoboCup 2014 in Brazil.

13.4 Kitting-Type Picking in the STAMINA Project

Techniques that were developed for the Cosero system are applicable to a much wider range of problems. As a first application, we investigated industrial bin picking in the STAMINA project [26]. The project targeted shop floor automation, in particular the automation of kitting tasks, where a robotic system needs to collect objects from different sources according to a kitting order. The completed *kit* is then delivered to the manufacturing line.

Fig. 13.15: The STAMINA cognitive robot performing an industrial kitting task in the experimental kitting zone at PSA Peugeot Citroën.

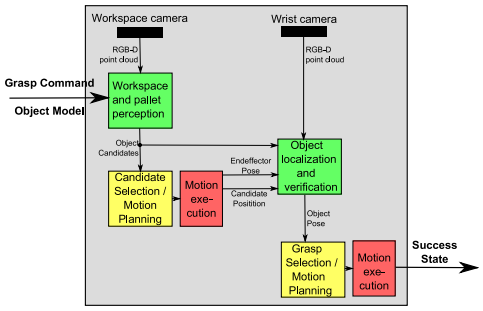


Fig. 13.16: Flow diagram of the two-staged perception pipeline.

13.4.1 System Description

Figure 13.15 shows the STAMINA robot during a typical kitting task. The system consists of a movable base equipped with an industrial arm, carrying a 4-DoF endeffector for grasping a wide variety of items. The system carries three ASUS Xtion Pro RGB-D cameras for perceiving the workspace, and a PrimeSense Carmine RGB-D camera at the wrist for close-range object perception.

The main difficulty lies in detection and pose estimation of the parts to be collected. We employ a two-stage work flow for this purpose (see Fig. 13.16). Here, methods developed for the Cosero system are re-used. In the first stage, a segmentation of the scene into individual parts is performed, following the RGB-D tabletop segmentation method described in Section 13.3.2.

After identifying a possible target part, the wrist camera is positioned above it and the part is recognized and its pose is estimated. Here, we employ the RGB-D registration method described in Section 13.3.2. A key advantage is that we can use the quality of the registration (measured using observation likelihoods for each matched surfel pair) for judging whether we actually a) have identified a part of the correct type and b) the registration was successful. Figure 13.17 shows a typical object perception process.

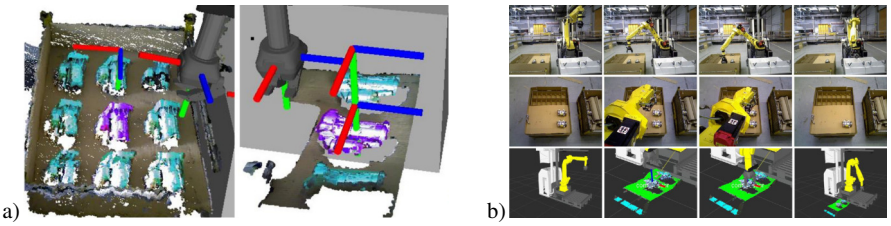


Fig. 13.17: RGB-D registration in a bin picking context: a) Detected objects with selected grasp target and fine registration using wrist-mounted RGB-D camera; b) Pick-and-place process with outside, top, and 3D visualization views.

Table 13.1: Bin picking results. The replanning column gives the number of times replanning of the arm trajectory was necessary. This condition was detected automatically. Taken from [33].

Task	Trials	Replanning	Success rate	Time [s]
5 parts	4	1	4/4	856±105
4 parts	6	3	6/6	723±96
3 parts	3	1	3/3	593±106
2 parts	3	1	3/3	325±16
1 part	14	4	14/14	234±105

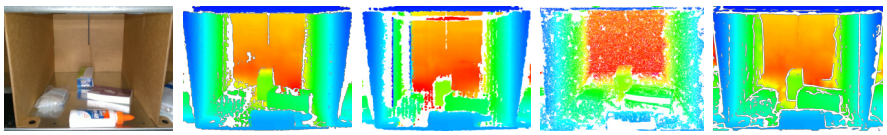
13.4.2 Evaluation

The STAMINA system was evaluated in realistic trials performed at PSA Peugeot Citroën, conducted in a 1,200 m^2 logistics kitting zone. The tests ranged from isolated "baseline" tests showcasing the robustness of the perception and motion planning methods (see Tab. 13.1 for brief results) to larger system-level and integrated tests, which proved overall robustness to a wide variety of possible situations and failures. We refer to [33] for full details on the evaluation.

13.5 Cluttered Bin Picking in the Amazon Robotics Challenge

The Amazon Picking Challenge (APC) 2016 and the subsequent Amazon Robotics Challenge 2017 were further opportunities to continue development of the so-far established object perception methods and to test them in realistic situations. The

Fig. 13.18: Our system at the Amazon Picking Challenge 2016. Left: Full system including robotic arm, endeffector, shelf, and red tote. Right: Custom-built endeffector with linear and rotatory joints, two Intel RealSense SR300 RGB-D cameras, and lighting.



(a) RGB frame (b) Upper depth (c) Lower depth (d) Stereo depth (e) Fused result

Fig. 13.19: RGB-D fusion from two sensors. Note the corruption in the left wall in the lower depth frame, which is corrected in the fused result.

challenge required participants to pick requested items out of highly cluttered, unsorted arrangements in narrow shelf bins or crowded shipment totes.

In contrast to the STAMINA application discussed in Section 13.4, the highly cluttered arrangements of different object require *semantic* segmentation of the scene into single objects, as geometry alone is insufficient for separation. Since a vacuum gripper is used to grasp the objects, requirements on pose estimation can be relaxed, though, since suitable vacuuming spots can be found on the live RGB-D input.

13.5.1 System Description

Figure 13.18 shows an overview of the system at APC 2016. It consists of a Universal Robots UR10 6-DoF robotic arm equipped with a custom 2-DoF endeffector. The endeffector consists of a linear joint for reaching into the narrow shelf bins, and a vacuum suction cup on a rotary joint, which allows to apply suction from above or from the front. The endeffector carries two Intel RealSense SR300 RGB-D cameras and illuminates the scene using own LED lighting to stay independent of outside lighting effects.

The RGB-D streams are interpreted by a separate vision computer. It carries four NVIDIA Titan X GPUs for on-site retraining of the deep learning models.

13.5.1.1 RGB-D Preprocessing

The decision to include two RGB-D cameras was made because of the difficult measurement situation inside the shelf bin. We observed that the nature of the sensors resulted in asymmetric effects, such as corruption of depth measurements on one of the bin walls (see Fig. 13.19). Depth completion alone (e.g. as presented in Chapter 2) did not yield sufficient results, as complete areas were missing. The second camera, mounted with 180° angle with respect to the first camera, had the measurement problems on the other side and thus can be used to correct for these effects. For breaking the tie between the two depth sources, an additional depth stream can

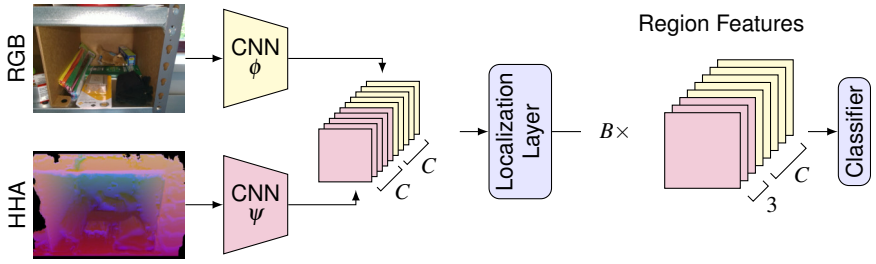


Fig. 13.20: Two-stream architecture for RGB-D object detection [52]. Input images in both modalities are processed individually using CNNs ϕ and ψ. The concatenated feature maps are then used in the classical Fast R-CNN pipeline using RoI pooling and a classification network.

be computed using stereo information from the two RGB cameras. For details on the RGB-D fusion strategy, we refer to [52].

13.5.1.2 Object Perception

For separating the objects in these cluttered situations, we designed an RGB-D object detection method. We followed up on research begun with the depth colorization method described in Section 13.3.2 and further investigated means of leveraging the depth modalities in deep-learning settings. For a modern object detection approach based on Faster R-CNN [48], we benchmarked different methods of incorporating depth in [52], such as a depth-based region proposal generator, a geometry-based encoding called HHA (horizontal disparity, height above ground, angle to gravity) either downsampled and provided to the classifier component, or processed in parallel to the RGB stream in a two-stream architecture. The best-performing method was to learn a separate depth feature extractor using a self-supervised approach called Cross Modal Distillation [22]. Here, the depth CNN is trained to imitate the output of a pre-trained RGB CNN on RGB-D frames. In this way, expensive annotation of RGB-D frames can be avoided. The trained depth CNN is then used in parallel with the pre-trained RGB CNN in a two-stream architecture (see Fig. 13.20). We also obtained small but consistent gains by combining the object detection results with the semantic segmentation approach described in Section 13.3.1.

13.5.2 Evaluation

The system was evaluated during the Amazon Picking Challenge 2016, where it performed highly successfully and reached a second place in the Stow competition (tote $\rightarrow$ shelf) and third place in the Pick competition (shelf $\rightarrow$ tote). Our system

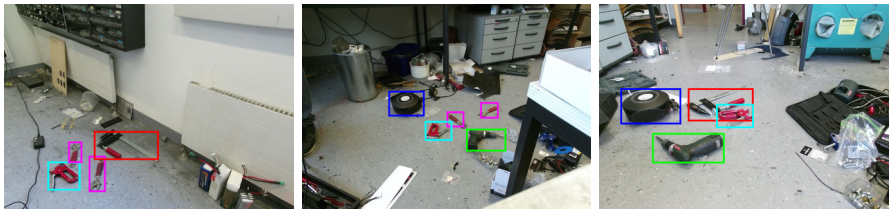


Fig. 13.21: Object detection in scenes with cluttered background. The frames are part of a publicly released RGB-D dataset of 129 frames, captured in a cluttered workshop environment. See `http://centauro-project.eu/data_multimedia/tools_data` for details.

actually performed the highest number of correct grasps during the pick competition (see Table 13.2), highlighting the robustness, speed, and precision of the presented RGB-D perception methods, but dropped three items while moving them, with the subsequent penalties leading to the third place.

In addition to the system-level evaluation during the APC 2016, we also evaluated our methods on in-house datasets. These consists of a 333-frame bin picking dataset, and a 129-frame RGB-D dataset with tools in front of highly cluttered background, captured for the CENTAURO disaster response project[1] (see Fig. 13.21). Here it demonstrated highly robust detection with 97% mAP score (see Tab. 13.3). The combination of the RGB-D object detector with semantic segmentation was also investigated and yielded small but consistent improvements (see Tab. 13.3). We refer to [52] for details.

[1] `https://www.centauro-project.eu`

Table 13.2: Picking Run at APC 2016

Bin	Item	Pick	Drop	Report	Bin	Item	Pick	Drop	Report
A	duct tape	×	×	×	G	scissors	×	×	×
B	bunny book	✓	✓	×[2]	H	plush bear	✓	×	✓
C	squeaky eggs	✓	×	✓	I	curtain	✓	×	✓
D	crayons[1]	✓	×	✓	J	tissue box	✓	×	✓
E	coffee	✓	✓	×[2]	K	sippy cup	✓	×	✓
F	hooks	✓	×	✓	L	pencil cup	✓	✓	×[2]
						Sum	10	3	7

The table shows the individual picks (A-L) executed during the official picking run.
[1] Misrecognized, corrected on second attempt.
[2] Incorrect report, resulting in penalty.

Table 13.3: Object detection results on the APC and CENTAURO tools datasets. *Det+Seg F1* is the semantic segmentation network boosted with object detection results.

	Object Detection		Semantic Segmentation	
Dataset	Mean AP	F1	Seg F1	Det+Seg F1
APC Shelf	0.912	0.798	0.813	0.827
APC Tote	0.887	0.779	0.839	0.853
CENTAURO tools	0.973	0.866	0.805	-

13.6 Conclusion

In this chapter, we described semantic RGB-D perception approaches developed for our cognitive service robot Cosero, industrial kitting in the STAMINA project, and cluttered bin picking for the Amazon Picking Challenge 2016.

We developed several object perception methods to implement the variety of manipulation skills of our robot. We segment scenes at high frame-rate into support surfaces and objects. In order to align to objects for grasping, we register RGB-D measurements on the object with a 3D model using multi-resolution surfel maps (MRSMaps). Through deformable registration of MRSMaps, we transfer object manipulation skills to differently shaped instances of the same object category. Tool-use is one of the most complex manipulation skills for humans and robots in daily life. We implemented several tool-use strategies using our perception and control methods.

The outstanding results achieved at multiple national and international RoboCup@Home competitions clearly demonstrate the versatility and robustness of the introduced methods. The development and benchmarking of the system gave us many insights into the requirements for complex personal service robots in scenarios such as cleaning the home or assisting the elderly. Challenges like RoboCup@Home show that a successful system not only consists of valid solutions to isolated problems—the proper integration of the overall system is equally important.

We also successfully demonstrated applicability of the developed methods for object detection, semantic segmentation, and RGB-D registration on other systems and in other domains, such as bin picking and disaster response.

Despite a large number of successful demonstrations, our systems are limited to short tasks in partially controlled environments. In order to scale towards real application in domestic service scenarios, we need to address open issues—and many of these are related to RGB-D perception. Object recognition and handling that scales to the large variety of objects in our daily homes is still an open research problem. Significant progress has been made, e.g. through deep learning methods, but occlusions and material properties like transparency or highly reflective surfaces make it still challenging to analyze typical household scenes. Similarly, perceiving people

and understanding their actions in the many situations possible in everyday environments is a challenge.

One promising approach to address these challenges is transfer learning which leverages the feature hierarchies from the large RGB data sets to the small robotic data sets at hand, requiring only few annotated training examples. Another line of research is to instrument the environment with a multitude of sensors in order to track all objects continuously with high accuracy [17].

Acknowledgment The authors thank the numerous people involved in development and operation of the mentioned robotic systems: Nikita Araslanov, Ishrat Badami, David Droeschel, Germán Martín García, Kathrin Gräve, Dirk Holz, Jochen Kläß, Christian Lenz, Manus McElhone, Anton Milan, Aura Munoz, Matthias Nieuwenhuisen, Arul Selvam Periyasamy, Michael Schreiber, Sebastian Schüller, David Schwarz, Ricarda Steffens, Jörg Stückler, and Angeliki Topalidou-Kyniazopoulou.

References

1. Asfour, T., Regenstein, K., Azad, P., Schroder, J., Bierbaum, A., Vahrenkamp, N., Dillmann, R.: Armar-III: An integrated humanoid platform for sensory-motor control. In: IEEE-RAS Int. Conference on Humanoid Robots (Humanoids) (2006)
2. Badami, I., Stückler, J., Behnke, S.: Depth-enhanced Hough forests for object-class detection and continuous pose estimation. In: ICRA Workshop on Semantic Perception, Mapping and Exploration (SPME) (2013)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv:1511.00561 (2015)
4. Bansal, A., Russell, B., Gupta, A.: Marr revisited: 2D-3D alignment via surface normal prediction. arXiv preprint arXiv:1604.01347 (2016)
5. Bäuml, B., Schmidt, F., Wimböck, T., Birbach, O., Dietrich, A., Fuchs, M., Friedl, W., Frese, U., Borst, C., Grebenstein, M., Eiberger, O., Hirzinger, G.: Catching flying balls and preparing coffee: Humanoid Rollin'Justin performs dynamic and sensitive tasks. In: Robotics and Automation (ICRA), IEEE Int. Conf. on (2011)
6. Beetz, M., Klank, U., Kresse, I., Maldonado, A., Mösenlechner, L., Pangercic, D., Rühr, T., Tenorth, M.: Robotic roommates making pancakes. In: Humanoid Robots (Humanoids), IEEE-RAS Int. Conf. on, pp. 529–536 (2011)
7. Behnke, S.: Hierarchical Neural Networks for Image Interpretation. Lecture Notes in Computer Science. Springer (2003)
8. Berner, A., Li, J., Holz, D., Stückler, J., Behnke, S., Klein, R.: Combining contour and shape primitives for object detection and pose estimation of prefabricated parts. In: Image Processing (ICIP), IEEE Int. Conf. on (2013)
9. Bohren, J., Rusu, R., Jones, E., Marder-Eppstein, E., Pantofaru, C., Wise, M., Mösenlechner, L., Meeussen, W., Holzer, S.: Towards autonomous robotic butlers: Lessons learned with the PR2. In: Robotics and Automation (ICRA), IEEE International Conference on (2011)
10. Borst, C., Wimböck, T., Schmidt, F., Fuchs, M., Brunner, B., Zacharias, F., Giordano, P.R., Konietschke, R., Sepp, W., Fuchs, S., et al.: Rollin'Justin–Mobile platform with variable base. In: Robotics and Automation (ICRA), IEEE Int. Conf. (2009)
11. van der Burgh, M., Lunenburg, J., Appeldoorn, R., Wijnands, R., Clephas, T., Baeten, M., van Beek, L., Ottervanger, R., van Rooy, H., van de Molengraft, M.: Tech united eindhoven @Home 2017 team description paper. University of Technology Eindhoven (2017)
12. Choi, C., Christensen, H.I.: RGB-D object pose estimation in unstructured environments. Robotics and Autonomous Systems **75**, 595–613 (2016)

13. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: Efficient and robust 3D object recognition. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on (2010)
14. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV (2015)
15. Endres, F., Hess, J., Sturm, J., Cremers, D., Burgard, W.: 3-D mapping with an RGB-D camera. IEEE Trans. Robotics **30**(1), 177–187 (2014)
16. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on (2014)
17. Fox, D.: The 100-100 tracking challenge. Keynote at ICRA conference (2016)
18. Gall, J., Lempitsky, V.S.: Class-specific Hough forests for object detection. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on (2009)
19. Garcia, G.M., Husain, F., Schulz, H., Frintrop, S., Torras, C., Behnke, S.: Semantic segmentation priors for object discovery. In: Pattern Recognition (ICPR), International Conference on (2016)
20. Girshick, R.B.: Fast R-CNN. In: Computer Vision (ICCV), IEEE International Conference on (2015)
21. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 142–158 (2016)
22. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2827–2836 (2016)
23. Hermann, A., Sun, J., Xue, Z., Rühl, S.W., Oberländer, J., Roennau, A., Zöllner, J.M., Dillmann, R.: Hardware and software architecture of the bimanual mobile manipulation robot HoLLiE and its actuated upper body. In: Advanced Intelligent Mechatronics (AIM), IEEE/ASME Int. Conf. on (2013)
24. Hermans, A., Floros, G., Leibe, B.: Dense 3D semantic mapping of indoor scenes from RGB-D images. In: ICRA (2014)
25. Höft, N., Schulz, H., Behnke, S.: Fast semantic segmentation of RGB-D scenes with GPU-accelerated deep neural networks. In: German Conference on AI (2014)
26. Holz, D., Topalidou-Kyniazopoulou, A., Stückler, J., Behnke, S.: Real-time object detection, localization and verification for fast robotic depalletizing. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, pp. 1459–1466. IEEE (2015)
27. Husain, F., Schulz, H., Dellen, B., Torras, C., Behnke, S.: Combining semantic and geometric features for object class segmentation of indoor scenes. IEEE Robotics and Automation Letters (RA-L) **2**(1), 49–55 (2016)
28. Iocchi, L., Holz, D., Ruiz-del Solar, J., Sugiura, K., van der Zant, T.: RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots. Artificial Intelligence **229**, 258–281 (2015)
29. Kaess, M., Johannsson, H., Roberts, R., Ila, V., Leonard, J.J., Dellaert, F.: isam2: Incremental smoothing and mapping using the bayes tree. I. J. Robotic Res. **31**(2), 216–235 (2012)
30. Kerl, C., Sturm, J., Cremers, D.: Robust odometry estimation for RGB-D cameras. In: Robotics and Automation (ICRA), IEEE International Conference on (2013)
31. Kittmann, R., Fröhlich, T., Schäfer, J., Reiser, U., Weißhardt, F., Haug, A.: Let me introduce myself: I am Care-O-bot 4. In: Mensch und Computer (2015)
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
33. Krueger, V., Rovida, F., Grossmann, B., Petrick, R., Crosby, M., Charzoule, A., Garcia, G.M., Behnke, S., Toscano, C., Veiga, G.: Testing the vertical and cyber-physical integration of cognitive robots in manufacturing. Robotics and Computer-Integrated Manufacturing **57**, 213–229 (2018)
34. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: G^2o: A general framework for graph optimization. In: Robotics and Automation (ICRA), IEEE International Conference on, pp. 3607–3613 (2011)

35. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. I. J. of Computer Vision **77**(1-3), 259–289 (2008)
36. Leidner, D., Dietrich, A., Schmidt, F., Borst, C., Albu-Schäffer, A.: Object-centered hybrid reasoning for whole-body mobile manipulation. In: Robotics and Automation (ICRA), IEEE International Conference on (2014)
37. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
38. Mazuran, M., Burgard, W., Tipaldi, G.D.: Nonlinear factor recovery for long-term SLAM. I. J. Robotic Res. **35**(1-3), 50–72 (2016)
39. McElhone, M., Stückler, J., Behnke, S.: Joint detection and pose tracking of multi-resolution surfel models in RGB-D. In: Europ. Conf. on Mobile Robots (2013)
40. Meeussen, W., Wise, M., Glaser, S., Chitta, S., McGann, C., Mihelich, P., Marder-Eppstein, E., Muja, M., Eruhimov, V., Foote, T., Hsu, J., Rusu, R.B., Marthi, B., Bradski, G., Konolige, K., Gerkey, B.P., Berger, E.: Autonomous door opening and plugging in with a personal robot. In: Robotics and Automation (ICRA), IEEE International Conference on, pp. 729–736 (2010)
41. Memmesheimer, R., Seib, V., Paulus, D.: homer@UniKoblenz: Winning team of the RoboCup@Home open platform league 2017. In: Robot World Cup, pp. 509–520. Springer (2017)
42. Müller, A.C., Behnke, S.: Learning depth-sensitive conditional random fields for semantic segmentation of RGB-D images. In: ICRA, pp. 6232–6237 (2014)
43. Myronenko, A., Song, X.: Point set registration: Coherent point drift. IEEE T. on Pattern Analysis and Machine Intelligence (PAMI) **32**(12), 2262–2275 (2010)
44. Nieuwenhuisen, M., Droeschel, D., Holz, D., Stückler, J., Berner, A., Li, J., Klein, R., Behnke, S.: Mobile bin picking with an anthropomorphic service robot. In: Robotics and Automation (ICRA), IEEE International Conference on (2013)
45. Papazov, C., Haddadin, S., Parusel, S., Krieger, K., Burschka, D.: Rigid 3D geometry matching for grasping of known objects in cluttered scenes. I. J. Robotic Res. **31**(4), 538–553 (2012)
46. Pavel, M.S., Schulz, H., Behnke, S.: Recurrent convolutional neural networks for object-class segmentation of RGB-D video. In: Neural Networks (IJCNN), International Joint Conference on (2015)
47. Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R., Ng, A.: ROS: An open-source Robot Operating System. In: IEEE International Conference on Robotics and Automation (ICRA) (2009)
48. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 91–99 (2015)
49. Schulz, H., Behnke, S.: Learning object-class segmentation with convolutional neural networks. In: European Symposium on Artificial Neural Networks (2012)
50. Schulz, H., Höft, N., Behnke, S.: Depth and height aware semantic RGB-D perception with convolutional neural networks. In: ESANN (2015)
51. Schulz, H., Waldvogel, B., Sheikh, R., Behnke, S.: CURFIL: Random forests for image labeling on GPU. In: International Conference on Computer Vision Theory and Applications (VISAPP), pp. 156–164 (2015)
52. Schwarz, M., Milan, A., Periyasamy, A.S., Behnke, S.: RGB-D object detection and semantic segmentation for autonomous manipulation in clutter. The International Journal of Robotics Research **37**(4-5), 437–451 (2018)
53. Schwarz, M., Schulz, H., Behnke, S.: RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In: Robotics and Automation (ICRA), IEEE Int. Conf. on, pp. 1329–1335 (2015)
54. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: Integrated recognition, localization and detection using convolutional networks. CoRR **abs/1312.6229** (2013)
55. Stoyanov, T., Magnusson, M., Andreasson, H., Lilienthal, A.J.: Fast and accurate scan registration through minimization of the distance between compact 3D NDT representations. I. J. Robotic Res. **31**(12), 1377–1393 (2012)

56. Stroucken, S.: Graph-basierte 3D-Kartierung von Innenräumen mit einem RGBD-Multikamera-System. Diplomarbeit, Universität Bonn, Computer Science VI (2013)
57. Stückler, J., Behnke, S.: Hierarchical object discovery and dense modelling from motion cues in RGB-D video. In: Artificial Intelligence (IJCAI), Int. Conf. (2013)
58. Stückler, J., Behnke, S.: Adaptive tool-use strategies for anthropomorphic service robots. In: Humanoid Robots (Humanoids), IEEE-RAS Int. Conf. on (2014)
59. Stückler, J., Behnke, S.: Efficient deformable registration of multi-resolution surfel maps for object manipulation skill transfer. In: Robotics and Automation (ICRA), IEEE Int. Conf. on (2014)
60. Stückler, J., Behnke, S.: Multi-resolution surfel maps for efficient dense 3D modeling and tracking. Journal of Visual Communication and Image Representation **25**(1), 137–147 (2014)
61. Stückler, J., Behnke, S.: Efficient dense rigid-body motion segmentation and estimation in RGB-D video. Int. J. of Computer Vision **113**(3), 233–245 (2015)
62. Stückler, J., Droeschel, D., Gräve, K., Holz, D., Schreiber, M., Topalidou-Kyniazopoulou, A., Schwarz, M., Behnke, S.: Increasing flexibility of mobile manipulation and intuitive human-robot interaction in RoboCup@Home. In: RoboCup 2013: Robot World Cup XVII, pp. 135–146. Springer (2014)
63. Stückler, J., Schwarz, M., Behnke, S.: Mobile manipulation, tool use, and intuitive interaction for cognitive service robot cosero. Frontiers in Robotics and AI **3**, 58 (2016)
64. Stückler, J., Steffens, R., Holz, D., Behnke, S.: Efficient 3D object perception and grasp planning for mobile manipulation in domestic environments. Robotics and Autonomous Systems **61**(10), 1106–1115 (2013)
65. Stückler, J., Waldvogel, B., Schulz, H., Behnke, S.: Dense real-time mapping of object-class semantics from RGB-D video. Journal of Real-Time Image Processing **10**(4), 599–609 (2015)
66. Su, H., Qi, C.R., Li, Y., Guibas, L.J.: Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In: Computer Vision (ICCV), IEEE International Conference on (2015)
67. Thrun, S., Montemerlo, M.: The graph slam algorithm with applications to large-scale mapping of urban structures. The International Journal of Robotics Research **25**(5-6), 403–429 (2006)
68. Vahrenkamp, N., Asfour, T., Dillmann, R.: Simultaneous grasp and motion planning: Humanoid robot ARMAR-III. Robotics & Automation Magazine (2012)
69. Wachsmuth, S., Lier, F., Meyer zu Borgsen, S., Kummert, J., Lach, L., Sixt, D.: Tobi-team of bielefeld a human-robot interaction system for robocup@ home 2017 (2017)
70. Whelan, T., Kaess, M., Johannsson, H., Fallon, M.F., Leonard, J.J., McDonald, J.: Real-time large-scale dense RGB-D SLAM with volumetric fusion. I. J. Robotic Res. **34**(4-5), 598–626 (2015)
71. Whelan, T., Leutenegger, S., Salas-Moreno, R., Glocker, B., Davison, A.J.: ElasticFusion: Dense SLAM without a pose graph. Robotics: Science & Systems (2015)
72. Wisspeintner, T., van der Zant, T., Iocchi, L., Schiffer, S.: RoboCup@Home: Scientific competition and benchmarking for domestic service robots. Interaction Studies **10**(3), 392–426 (2009)
73. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: A deep representation for volumetric shapes. In: Computer Vision and Pattern Recognition (CVPR), IEEE Conference on (2015)
74. Zhang, J., Singh, S.: Loam: Lidar odometry and mapping in real-time. In: Robotics: Science and Systems Conference (RSS), pp. 109–111 (2014)

Index