Chapter 6

Intelligent Robotic Perception Systems

Cristiano Premebida, Rares Ambrus and
Zoltan-Csaba Marton

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.79742

Abstract

Robotic perception is related to many applications in robotics where sensory data and artificial intelligence/machine learning (AI/ML) techniques are involved. Examples of such applications are object detection, environment representation, scene understanding, human/pedestrian detection, activity recognition, semantic place classification, object modeling, among others. Robotic perception, in the scope of this chapter, encompasses the ML algorithms and techniques that empower robots to learn from sensory data and, based on learned models, to react and take decisions accordingly. The recent developments in machine learning, namely deep-learning approaches, are evident and, consequently, robotic perception systems are evolving in a way that new applications and tasks are becoming a reality. Recent advances in human-robot interaction, complex robotic tasks, intelligent reasoning, and decision-making are, at some extent, the results of the notorious evolution and success of ML algorithms. This chapter will cover recent and emerging topics and use-cases related to intelligent perception systems in robotics.

Keywords: robotic perception, machine learning, advanced robotics, artificial intelligence

1. Introduction

In robotics, perception is understood as a system that endows the robot with the ability to perceive, comprehend, and reason about the surrounding environment. The key components of a perception system are essentially sensory data processing, data representation (environment modeling), and ML-based algorithms, as illustrated in **Figure 1**. Since *strong AI* is still far from being achieved in real-world robotics applications, this chapter is about *weak AI*, i.e., standard machine learning approaches [1].

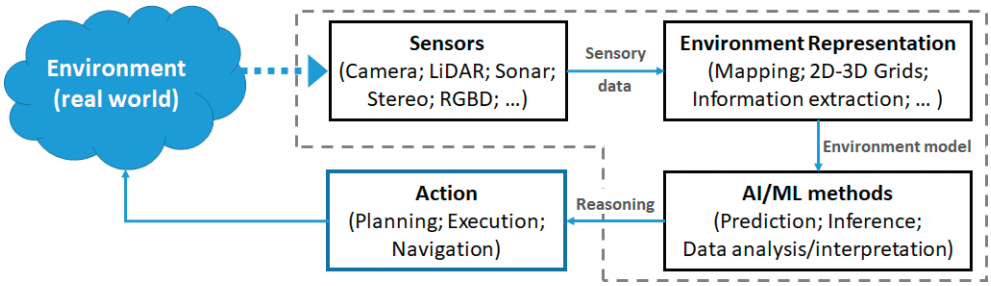

Figure 1. Key modules of a typical robotic perception system: sensory data processing (focusing here on visual and range perception); data representations specific for the tasks at hand; algorithms for data analysis and interpretation (using AI/ML methods); and planning and execution of actions for robot-environment interaction.

Robotic perception is crucial for a robot to make decisions, plan, and operate in real-world environments, by means of numerous functionalities and operations from occupancy grid mapping to object detection. Some examples of robotic perception subareas, including autonomous robot-vehicles, are obstacle detection [2, 3], object recognition [4, 5], semantic place classification [6, 7], 3D environment representation [8], gesture and voice recognition [9], activity classification [10], terrain classification [11], road detection [12], vehicle detection [13], pedestrian detection [14], object tracking [3], human detection [15], and environment change detection [16].

Nowadays, most of robotic perception systems use machine learning (ML) techniques, ranging from classical to deep-learning approaches [17]. Machine learning for robotic perception can be in the form of unsupervised learning, or supervised classifiers using handcrafted features, or deep-learning neural networks (e.g., convolutional neural network (CNN)), or even a combination of multiple methods.

Regardless of the ML approach considered, data from sensor(s) are the key ingredient in robotic perception. Data can come from a single or multiple sensors, usually mounted onboard the robot, but can also come from the infrastructure or from another robot (e.g., cameras mounted on UAVs flying nearby). In multiple-sensors perception, either the same modality or multimodal, an efficient approach is usually necessary to combine and process data from the sensors before an ML method can be employed. Data alignment and calibration steps are necessary depending on the nature of the problem and the type of sensors used.

Sensor-based environment representation/mapping is a very important part of a robotic perception system. Mapping here encompasses both the acquisition of a metric model and its semantic interpretation, and is therefore a synonym of environment/scene representation. This semantic mapping process uses ML at various levels, e.g., reasoning on volumetric occupancy and occlusions, or identifying, describing, and matching optimally the local regions from different time-stamps/models, i.e., not only higher level interpretations. However, in the majority of applications, the primary role of environment mapping is to model data from exteroceptive sensors, mounted onboard the robot, in order to enable reasoning and inference regarding the real-world environment where the robot operates.

Robot perception functions, like localization and navigation, are dependent on the environment where the robot operates. Essentially, a robot is designed to operate in two categories of

environments: indoors or outdoors. Therefore, different assumptions can be incorporated in the mapping (representation) and perception systems considering indoor or outdoor environments. Moreover, the sensors used are different depending on the environment, and therefore, the sensory data to be processed by a perception system will not be the same for indoors and outdoors scenarios. An example to clarify the differences and challenges between a mobile robot navigating in an indoor versus outdoor environment is the ground, or terrain, where the robot operates. Most of indoor robots assume that the ground is regular and flat which, in some manner, facilitates the environment representation models; on the other hand, for field (outdoors) robots, the terrain is quite often far from being regular and, as consequence, the environment modeling is itself a challenge and, without a proper representation, the subsequent perception tasks are negatively affected. Moreover, in outdoors, robotic perception has to deal with weather conditions and variations in light intensities and spectra.

Similar scenario-specific differences exist in virtually all use-cases of robotic vision, as exemplified by the 2016 Amazon Picking Challenge participants' survey [18], requiring complex yet robust solutions, and therefore considered one of the most difficult tasks in the pick-and-place application domain. Moreover, one of the participating teams from 2016 benchmarked a pose estimation method on a warehouse logistics dataset, and found large variations in performance depending on clutter level and object type [2]. Thus, perception systems currently require expert knowledge in order to select, adapt, extend, and fine-tune the various employed components.

Apart from the increased training data sizes and robustness, the end-to-end training aspect of deep-learning (DL) approaches made the development of perception systems easier and more accessible for newcomers, as one can obtain the desired results directly from raw data in many cases, by providing a large number of training examples. The method selection often boils down to obtaining the latest pretrained network from an online repository and fine-tuning it to the problem at hand, hiding all the traditional feature detection, description, filtering, matching, optimization steps behind a relatively unified framework. Unfortunately, at the moment an off-the-shelf DL solution for every problem does not exist, or at least no usable pretrained network, making the need for huge amounts of training data apparent. Therefore, large datasets are a valuable asset for modern AI/ML. A large number of datasets exist for perception tasks as well, with a survey of RGB-D datasets presented by Firman [5] (up-to-date list available online: http://www.michaelfirman.co.uk/RGBDdatasets/), and even tools for synthetically generating sensor-based datasets, e.g., the work presented by Handa et al. [4] which is available online: http://robotvault.bitbucket.org/. However, the danger is to overfit to such benchmarks, as the deployment environment of mobile robots is almost sure to differ from the one used in teaching the robot to perceive and understand the surrounding environment. Thus, the suggestions formulated by Wagstaff [19] still hold true today and should be taken to heart by researchers and practitioners.

As pointed out recently by Sünderhauf et al. [17], robotic perception (also designated robotic vision in [17]) differs from traditional computer vision perception in the sense that, in robotics, the outputs of a perception system will result in decisions and actions in the real world. Therefore, perception is a very important part of a complex, embodied, active, and goal-driven robotic system. As exemplified by Sünderhauf et al. [17], robotic perception has to translate images (or scans, or point-clouds) into actions, whereas most computer vision applications take images and translate the outputs into information.

2. Environment representation

Among the numerous approaches used in environment representation for mobile robotics, and for autonomous robotic-vehicles, the most influential approach is the occupancy grid mapping [20]. This 2D mapping is still used in many mobile platforms due to its efficiency, probabilistic framework, and fast implementation. Although many approaches use 2D-based representations to model the real world, presently 2.5D and 3D representation models are becoming more common. The main reasons for using higher dimensional representations are essentially twofold: (1) robots are demanded to navigate and make decisions in higher complex environments where 2D representations are insufficient; (2) current 3D sensor technologies are affordable and reliable, and therefore 3D environment representations became attainable. Moreover, the recent advances in software tools, like ROS and PCL, and also the advent of methods like Octomaps, developed by Hornung et al. [21], have been contributing to the increase in 3D-like environment representations.

The advent and proliferation of RGBD sensors has enabled the construction of larger and evermore detailed 3D maps. In addition, considerable effort has been made in the semantic labeling of these maps, at pixel and voxels levels. Most of the relevant approaches can be split into two main trends: methods designed for online and those designed for offline use. Online methods process data as it is being acquired by the mobile robot, and generate a semantic map incrementally. These methods are usually coupled with a SLAM framework, which ensures the geometric consistency of the map. Building maps of the environment is a crucial part of any robotic system and arguably one of the most researched areas in robotics. Early work coupled mapping with localization as part of the simultaneous localization and mapping (SLAM) problem [22, 23]. More recent work has focused on dealing with or incorporating time-dependencies (short or long term) into the underlying structure, using either grid maps as described in [8, 24], pose-graph representations in [25], and normal distribution transform (NDT) [16, 26].

As presented by Hermans et al. [27], RGBD data are processed by a random forest-based classifier and predict semantic labels; these labels are further regularized through the conditional random field (CRF) method proposed by Krahenbuhl and Koltun [28]. Similarly, McCormac et al. [29] use the elastic fusion SLAM algorithm proposed by Whelan et al. [30] to fuse CNN predictions about the scene in a geometrically consistent map. In the work of Sünderhauf et al. [6], a CNN is used to incrementally build a semantic map, with the aim of extending the number of classes supported by the CNN by complementing it with a series of one-vs-all classifiers which can be trained online. A number of semantic mapping approaches are designed to operate offline, taking as input a complete map of the environment. In the methods described by Ambrus et al. [31, 32] and Armeni et al. [33], large-scale point clouds of indoor buildings are processed, and then, after segmenting the input data, the method's outputs are in the form of a set of "rooms." Ambrus et al. [31, 32] use a 2D cell-complex graph-cut approach to compute the segmentation with the main limitation that only single floor buildings can be processed, while Armeni et al. [33] process multifloor structures by detecting the spaces between the walls, ceilings, etc., with the limitation that the building walls have to be axis-aligned (i.e., the Manhattan world assumption). Similarly, in the work proposed by Mura et al. [34], a large point cloud of an indoor structure is processed by making use of a 3D cell-complex structure

and outputting a mesh containing the semantic segmentation of the input data. However, the main limitation in [34] is that the approach requires knowledge of the positions from which the environment was scanned when the input data were collected.

The recent work presented by Brucker et al. [7] builds on the segmentation of Ambrus et al. [31, 32] and explores ways of fusing different types of information, such as presence of objects and cues of the types of rooms to obtain a semantic segmentation of the environment. The aim of the work presented by Brucker et al. [7] is to obtain an intuitive and human-like labeling of the environment while at the same time preserving as many of the semantic features as possible. Also, Brucker et al. [7] use a conditional random field (CRF) or the fusion of various heterogeneous data sources and inference is done using Gibbs sampling technique.

Processing sensory data and storing it in a representation of the environment (i.e., a map of the environment) has been and continues to be an active area in robotics research, including autonomous driving system (or autonomous robotic-vehicles). The approaches covered range from metric representations (2D or 3D) to higher semantic or topological maps, and all serve specific purposes key to the successful operation of a mobile robot, such as localization, navigation, object detection, manipulation, etc. Moreover, the ability to construct a geometrically accurate map further annotated with semantic information also can be used in other applications such as building management or architecture, or can be further fed back into a robotic system, increasing the awareness of its surroundings and thus improving its ability to perform certain tasks in human-populated environments (e.g., finding a cup is more likely to be successful if the robot knows a priori which room is the kitchen and how to get there).

3. Artificial intelligence and machine learning applied on robotics perception

Once a robot is (self) localized, it can proceed with the execution of its task. In the case of autonomous mobile manipulators, this involves localizing the objects of interest in the operating environment and grasping them. In a typical setup, the robot navigates to the region of interest, observes the current scene to build a 3D map for collision-free grasp planning and for localizing target objects. The target could be a table or container where something has to be put down, or an object to be picked up. Especially in the latter case, estimating all 6 degrees of freedom of an object is necessary. Subsequently, a motion and a grasp are computed and executed. There are cases where a tighter integration of perception and manipulation is required, e.g., for high-precision manipulation, where approaches like visual servoing are employed. However, in every application, there is a potential improvement for treating perception and manipulation together.

Perception and manipulation are complementary ways to understand and interact with the environment and according to the common coding theory, as developed and presented by Sperry [35], they are also inextricably linked in the brain. The importance of a tight link between perception and action for artificial agents has been recognized by Turing [36], who suggested to equip computers “with the best sense organs that money can buy” and let them learn from gathered experiences until they pass his famous test as described in [37].

The argument for embodied learning and grounding of new information evolved, considering the works of Steels and Brooks [38] and Vernon [39], and more recently in [40], robot perception involves planning and interactive segmentation. In this regard, perception and action reciprocally inform each other, in order to obtain the best results for locating objects. In this context, the localization problem involves segmenting objects, but also knowing their position and orientation relative to the robot in order to facilitate manipulation. The problem of object pose estimation, an important prerequisite for model-based robotic grasping, uses in most of the cases precomputed grasp points as described by Ferrari and Canny [41]. We can categorize this topic in either template/descriptor-based approaches or alternatively local feature/patch-based approaches. In both cases, an ever-recurring approach is that bottom-up data-driven hypothesis generation is followed and verified by top-down concept-driven models. Such mechanisms are assumed, as addressed by Frisby and Stone [42], to be like our human vision system.

The approaches presented in ([43–45] make use of color histograms, color gradients, depth or normal orientations from discrete object views, i.e., they are examples of vision-/camera-based perception for robots. Vision-based perception systems typically suffer from occlusions, aspect ratio influence, and from problems arising due to the discretization of the 3D or 6D search space. Conversely, in the works of [46–48], they predict the object pose through voting or a PnP algorithm [49]. The performance usually decreases if the considered object lacks texture and if the background is heavily cluttered. In the works listed above, learning algorithms based on classical ML methods and deep-learning (e.g., CNN) have been employed.

The importance of mobile manipulation and perception areas has been signaled by the (not only academic) interest spurred by events like the Amazon Robotics (formerly Picking) Challenge and the workshop series at the recent major computer vision conferences associated with the SIXD Challenge (http://cmp.felk.cvut.cz/sixd/workshop_2018/). However, current solutions are either heavily tailored to a specific application, requiring specific engineering during deployment, or their generality makes them too slow or imprecise to fulfill the tight time-constraints of industrial applications. While deep learning holds the potential to both improve accuracy (i.e., classification or recognition performance) and also to increase execution speed, more work on transfer learning, in the sense of generalization improvement, is required to apply models learned in real-world and also in unseen (new) environment. Domain adaptation and domain randomization (i.e., image augmentations) seem to be important directions to pursue, and should be explored not only for vision/camera cases, but also for LiDAR-based perception cases.

Usually, in traditional mobile robot manipulation use-cases, the navigation and manipulation capabilities of a robot can be exploited to let the robot gather data about objects autonomously. This can involve, for instance, observing an object of interest from multiple viewpoints in order to allow a better object model estimation, or even in-hand modeling. In the case of perception for mobile robots and autonomous (robot) vehicles, such options are not available; thus, its perception systems have to be trained offline. However, besides AI/ML-based algorithms and higher level perception, for autonomous driving applications, environment representation (including multisensor fusion) is of primary concern [50, 51].

The development of advanced perception for (full) autonomous driving has been a subject of interest since the 1980s, having a period of strong development due to the DARPA

Challenges (2004, 2005, and 2007) and the European ELROB challenges (since 2006), and more recently, it has regained considerable interest from automotive and robotics industries and academia. Research in self-driving cars, also referred as autonomous robot-cars, is closely related to mobile robotics and many important works in this field have been published in well-known conferences and journals devoted to robotics. Autonomous driving systems (ADS) comprise, basically, perception (including sensor-fusion and environment modeling/representation), localization, and navigation (path planning, trajectory following, control) and, more recently, cooperation (V2X-based communication technologies). However, the cornerstone of ADS is the perception system because it is involved in most of the essential and necessary tasks for safe driving such as the "segmentation," detection/recognition, of: road, lane-markings, pedestrians, and other vulnerable road users (e.g., cyclists), other vehicles, traffic signals, crosswalks, and the numerous other types of objects and obstacles that can be found on the roads. In addition to the sensors (e.g., cameras, LIDAR, Radar, "new" solid-state LiDAR technology) and the models used in ADS, the common denominator in a perception system consists of AI/ML algorithms, where deep learning is the leading technique for semantic segmentation and object detection [50].

One of current trends in autonomous vehicles and robotics is the promising idea of incorporating cooperative information, from connected environment/infrastructure, into the decision loop of the robotic perception system. The rationale is to improve robustness and safety by providing complementary information to the perception system, for example: the position and identification of a given object or obstacle on the road could be reported (e.g., broadcasted through a communication network) in advance to an autonomous car, moments before the object/obstacle are within the onboard sensor's field/range of view.

4. Case studies

4.1. The Strands project

The EU FP7 Strands project [52] is formed by a consortium of six universities and two industrial partners. The aim of the project is to develop the next generation of intelligent mobile robots, capable of operating alongside humans for extended periods of time. While research into mobile robotic technology has been very active over the last few decades, robotic systems that can operate robustly, for extended periods of time, in human-populated environments remain a rarity. Strands aims to fill this gap and to provide robots that are intelligent, robust, and can provide useful functions in real-world security and care scenarios. Importantly, the extended operation times imply that the robotic systems developed have to be able to cope with an ever-increasing amount of data, as well as to be able to deal with the complex and unstructured real world (**Figure 2**).

Figure 3 shows a high level overview of the Strands system (with more details in [52]): the mobile robot navigates autonomously between a number of predefined waypoints. A task scheduling mechanism dictates when the robot should visit which waypoints, depending on the tasks the robot has to accomplish on any given day. The perception system consists, at the lowest level, of a module which builds local metric maps at the waypoints visited by the robot.

Figure 2. The Strands project (image from http://strands.acin.tuwien.ac.at/).

These local maps are updated over time, as the robot revisits the same locations in the environment, and they are further used to segment out the dynamic objects from the static scene. The dynamic segmentations are used as cues for higher level behaviors, such as triggering a data acquisition and object modeling step, whereby the robot navigates around the detected object to collect additional data which are fused into a canonical model of the object [53]. The data can

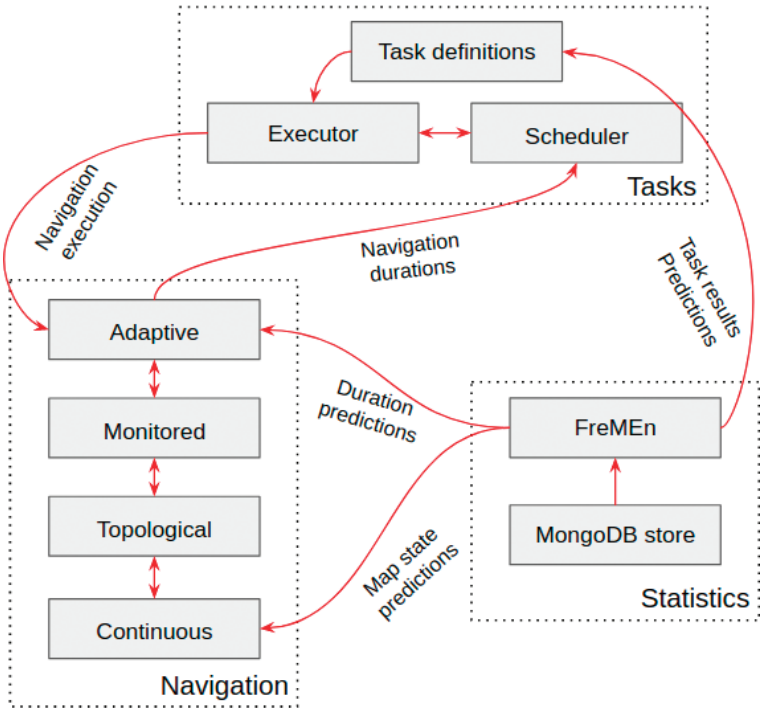

Figure 3. The Strands system—Overview.

further be used to generate a textured mesh through which a convolutional neural network can be trained which can successfully recognize the object in future observations [31, 32]. The dynamics detected in the environment can be used to detect patterns, either through spectral analysis (i.e., by applying a Fourier transform on the raw detection data), as described in [54], or as part of a multitarget tracking system based on a Rao-Blackwellized particle filter.

In addition to the detection and modeling of objects, the Strands perception system also focuses on the detection of people. Beyer et al. [55] present a method to continuously estimate the head-pose of people, while in [15] laser and RGB-D are combined to reliably detect humans and to allow human-aware navigation approaches which make the robot more socially acceptable. Beyer et al. [56] propose a CNN-based system which uses laser scanner data to detect objects; the usefulness of the approach is demonstrated in the case scenario, where it is used to detect wheelchairs and walkers.

Robust perception algorithms that can operate reliably for extended periods of time are one of the cornerstones of the Strands system. However, any algorithm deployed on the robot has to be not only robust, but also able to scale as the robot makes more observations and collects more information about the world. One of the key parts that would enable the successful operation of such a robotic system is a perception stack that is able to continuously integrate observations about the world, extract relevant parts as well as build models that understand and are able to predict what the environment will look like in the future. This spatio-temporal understanding is crucial, as it allows a mobile robot to compress the data acquired during months of autonomous operation into models that can be used to refine the robot's operation over time. Modeling periodicities in the environment and integrating them into a planning pipeline is further investigated by Fentanes et al. [57], while Santos et al. [58] build spatio-temporal models of the environment and use them for exploration through an information-theoretic approach which predicts the potential gain of observing particular areas of the world at different points in time.

4.2. The RobDREAM project

Advanced robots operating in complex and dynamic environments require intelligent perception algorithms to navigate collision-free, analyze scenes, recognize relevant objects, and manipulate them. Nowadays, the perception of mobile manipulation systems often fails if the context changes due to a variation, e.g., in the lightning conditions, the utilized objects, the manipulation area, or the environment. Then, a robotic expert is needed who needs to adjust the parameters of the perception algorithm and the utilized sensor or even select a better method or sensor. Thus, a high-level cognitive ability that is required for operating alongside humans is to continuously improve performance based on introspection. This adaptability to changing situations requires different aspects of machine learning, e.g., storing experiences for life-long learning, generating annotated datasets for supervised learning through user interaction, Bayesian optimization to avoid brute-force search in high-dimensional data, and a unified representation of data and meta-data to facilitate knowledge transfer.

The RobDREAM consortium automated and integrated different aspects of these. Specifically, in the EU's H2020 RobDREAM project, a mobile manipulator was used to showcase the

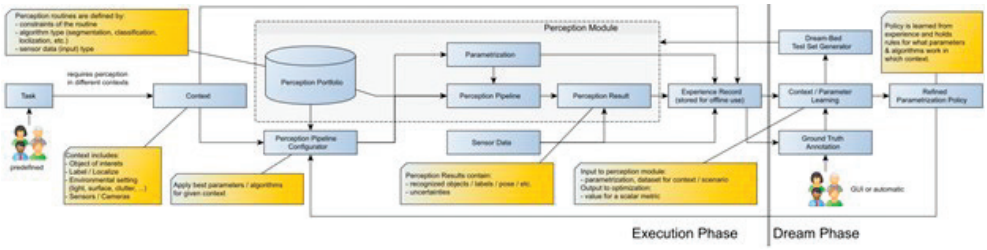

Figure 4. Schematics of the RobDREAM approach (image based on deliverables of http://robdream.eu/).

intuitive programming and simplified setup of robotic applications enabled by automatically tuning task execution pipelines according to user-defined performance criteria.

As illustrated in **Figure 4**, this was achieved by a semantically annotated logging of perceptual episodic memories that can be queried intuitively in order to analyze the performance of the system in different contexts. Then, a ground truth annotation tool can be used by the user to mark satisfying results, or correct unsatisfying ones, where the suggestions and interactive capabilities of the system reduced the cognitive load of this often complicated task (especially when it comes to 6 DoF pose annotations), as shown in user studies involving computer vision expert and nonexpert users alike.

These annotations are then used by a Bayesian optimization framework to tune the off-the-shelf pipeline to the specific scenarios the robot encounters, thereby incrementally improving the performance of the system. The project did not focus only on perception, but on other key technologies for mobile manipulation as well. Bayesian optimization and other techniques were used to adapt the navigation, manipulation, and grasping capabilities independently of each other and the perception ones. However, the combinatorial complexity of the joint parameter space of all the involved steps was too much even for such intelligent meta-learners. The final industrially relevant use-case demo featured the kitting and mounting of electric cabinet board elements, for which a pose-annotated database was built using two RBD-D cameras and released to the public (http://www.dlr.de/rm/thr-dataset).

4.3. The SPENCER project

When deploying robots in scenarios where they need to share the environment and interact with a large number of people, it is increasingly important that their functionalities are "socially aware." This means that they respect the personal space (and also privacy) of encountered persons, does not navigate s.t. to cut up cues or groups, etc. Such functionalities go beyond the usual focus of robotics research groups, while academics focusing on user experience typically do not have the means to develop radically new robots. However, the EU's FP7 program funded such an interdisciplinary project, called SPENCER, driven by an end-user in the aviation industry.

Since around 80% of passenger traffic at different hubs, including Schiphol in Amsterdam, is comprised of passengers who are transferring from one flight to the other, KLM is interested in an efficient management of their movements. For example, when transfer times are short, and finding one's way in a big airport is difficult due to language and alphabet barriers, people are at risk to losing their connection. In such, and similar cases, robotic assistants that can be deployed and booked flexibly can possibly help alleviate some of the problem. This use-case was explored by the SPENCER demonstrator for smart passengers' flow management and mobile information provider, but similar solutions are required in other domains as well (**Figure 5**).

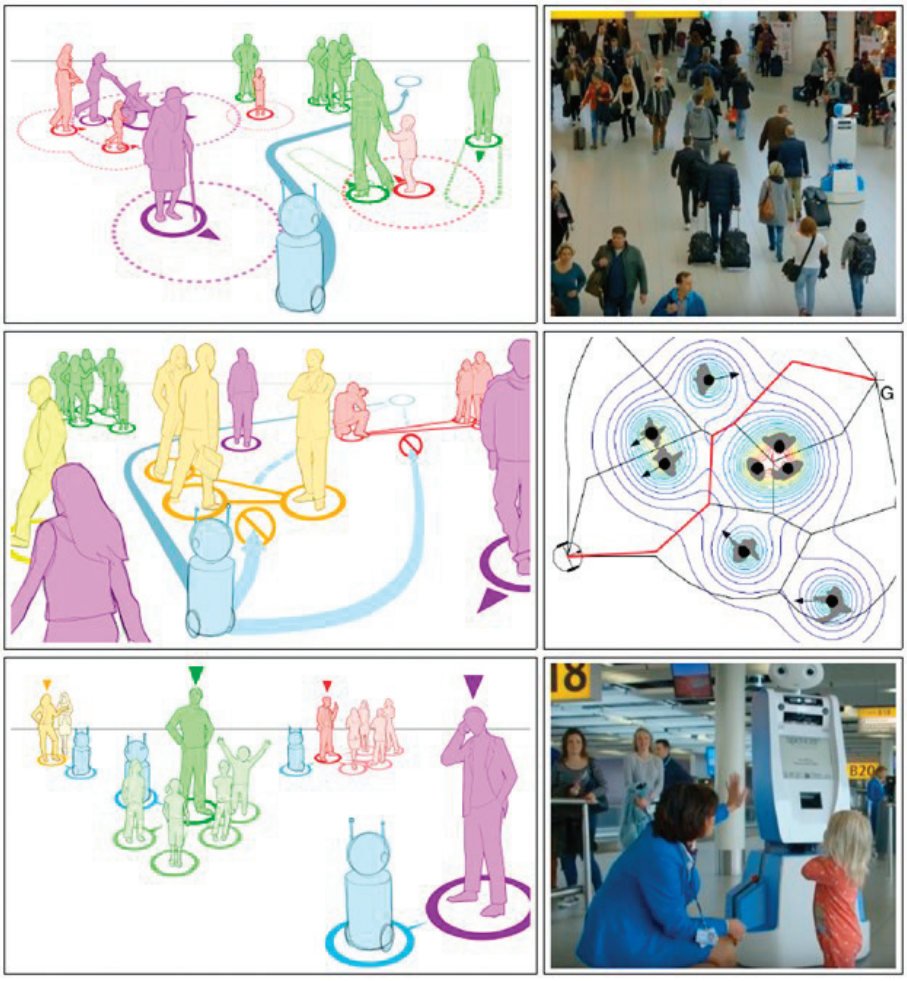

Figure 5. Concept and results of the SPENCER project (images from http://www.spencer.eu/).

The SPENCER consortium integrated the developed technologies onto a robot platform whose task consists in picking up short-transfer time passenger groups at their gate of arrival, identifying them with an onboard boarding pass reader, guiding them to the Schengen barrier and instructing them to use the priority track [59]. Additionally, the platform was equipped with a KLM information kiosk and provides services to passengers in need of help.

In crowded environments such as airports, generating short and safe paths for mobile robots is still difficult. Thus, social scene understanding and long-term prediction of human motion in crowds is not sufficiently solved but highly relevant for all robots that need to quickly navigate in human environments, possibly under temporal constraints. Social scene understanding means, in part, that a reliable tracking and prediction of people's motion with low uncertainty is available, and that is particularly hard if there are too many occlusions and too many fast changes of motion direction. Classical path planning approaches often result in an overconstrained or overly cautious robot that either fails to produce a feasible and safe path in the crowd, or plans a large and suboptimal detour to avoid people in the scene.

4.4. The AUTOCITS project

The AUTOCITS (https://www.autocits.eu/) project will carry out a comprehensive assessment of cooperative systems and autonomous driving by deploying real-world Pilots, and will study and review regulations related to automated and autonomous driving. AUTOCITS, cofinanced by the European Union through the Connecting Europe Facility (CEF) Program, aims to facilitate the deployment of autonomous vehicles in European roads, and to use connected/cooperative intelligent transport systems (C-ITS) services to share information between autonomous vehicles and infrastructure, by means of V2V and V2I communication technology, to improve safety and to facilitate the coexistence of autonomous cars in real-world traffic conditions. The AUTOCITS Pilots, involving connected and autonomous vehicles (including autonomous shuttles, i.e., low-speed robot-vehicles), will be deployed in three major European cities in "the Atlantic Corridor of the European Network": Lisbon (Portugal), Madrid (Spain), and Paris (France).

A number of technologies are involved in AUTOCITS, ranging from the onboard and roadside units (OBU, RSU) to the autonomous driving systems that equip the cars. Today, the autonomous and/or automated driving technology we see on the roads belongs to the levels 3 or 4 (with respect to the SAE's levels of automation in vehicles). In AUTOCITS, the Pilot's deployment will be of level 3 to 4. In this context, it is important to say that level 5 cars (i.e., 100% self-driving or full-automated cars: the driving wheels would be unnecessary) operating in real-world roads and streets are still far from reality.

We can say that the perception system is in charge of all tasks related to object and event detection and response (OEDR). Therefore, a perception system—including of course its software modules—is responsible for sensing, understanding, and reasoning about the autonomous car's surroundings. Within a connected and cooperative environment, connected cars would leverage and complement onboard sensor data by using information from vehicular communication systems (i.e., V2X technology): information from other connected vehicles, from infrastructure, and road users (and vice-versa).

5. Conclusions and remarks

So just how capable is current perception and AI, and how close did/can it get to human-level performance? Szeliski [60] in his introductory book to computer vision argued that traditional vision struggled to reach the performance of a 2-year old child, but today's CNNs reach super-human classification performance on restricted domains (e.g., in the ImageNet Large Scale Visual Recognition Challenge: http://www.image-net.org/challenges/LSVRC/).

The recent surge and interest in deep-learning methods for perception has greatly improved performance in a variety of tasks such as object detection, recognition, semantic segmentation, etc. One of the main reasons for these advancements is that working on perception systems lends itself easily to offline experimentation on publicly available datasets, and comparison to other methods via standard benchmarks and competitions.

Machine learning (ML) and deep learning (DL), the latter has been one of the most used keywords in some conferences in robotics recently, are consolidated topics embraced by the robotics community nowadays. While one can interpret the filters of CNNs as Gabor filters and assume to be analogous to functions of the visual cortex, currently, deep learning is a purely nonsymbolic approach to AI/ML, and thus not expected to produce "strong" AI/ML. However, even at the current level, its usefulness is undeniable, and perhaps, the most eloquent example comes from the world of autonomous driving which brings together the robotics and the computer vision community. A number of other robotics-related products are starting to be commercially available for increasingly complex tasks such as visual question and answering systems, video captioning and activity recognition, large-scale human detection and tracking in videos, or anomaly detection in images for factory automation.

Author details

Cristiano Premebida[1]*, Rares Ambrus[2] and Zoltan-Csaba Marton[3]

*Address all correspondence to: cpremebida@isr.uc.pt

1 Institute of Systems and Robotics (ISR-UC), Coimbra, Portugal

2 Toyota Research Institute, Los Altos, California, USA

3 German Aerospace Center, Köln, Germany

References

[1] Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-790395-2; 2003

[2] Rennie C, Shome R, Bekris KE, Souza AF. A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place. IEEE Robotics and Automation Letters. July 2016;**1**(2), pp. 1179-1185

[3] Bore N, Jensfelt P, Folkesson J. Multiple object detection, tracking and long-term dynamics learning in large 3D maps. CoRR, https://arxiv.org/abs/1801.09292. 2018

[4] Handa A, Patraucean V, Badrinarayanan V, Stent S, Cipolla R. Understanding real world indoor scenes with synthetic data. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 4077-4085

[5] Firman M. RGBD datasets: Past, present and future. Firman 2016 RGBDDP, In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) on Large Scale 3D Data: Acquisition, Modelling and Analysis; 2016. 661-673

[6] Sünderhauf N, Dayoub F, McMahon S, Talbot B, Schulz R, Corke P, Wyeth G, Upcroft B, Milford M. Place categorization and semantic mapping on a mobile robot. In: IEEE International Conference on Robotics and Automation (ICRA); Stockholm; 2016. pp. 5729-5736

[7] Brucker M, Durner M, Ambrus R, Csaba Marton Z, Wendt A, Jensfelt P, Arras KO, Triebel R. Semantic labeling of indoor environments from 3D RGB maps. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018

[8] Saarinen J, Andreasson H, Lilienthal AJ. Independent Markov chain occupancy grid maps for representation of dynamic environment. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2012. pp. 3489-3495

[9] Fong T, Nourbakhsh I, Dautenhahn K. A survey of socially interactive robots. Robotics and Autonomous Systems. 2003;**42**(3-4):143-166

[10] Diego R. Faria, Mario Vieira, Premebida C, Nunes U. Probabilistic human daily activity recognition towards robot-assisted living. In: Proceedings of the IEEE RO-MAN'15; Japan; 2015

[11] Manduchi R, Castano A, Talukder A, Matthies L. Obstacle detection and terrain classification for autonomous off-road navigation. Autonomous Robots. 2005;**18**(1):81-102

[12] Fernandes R, Premebida C, Peixoto P, Wolf D, Nunes U. Road detection using high resolution LIDAR. In: IEEE Vehicle Power and Propulsion Conference, IEEE-VPPC; 2014

[13] Asvadi A, Garrote L, Premebida C, Peixoto P, Nunes UJ. Multimodal vehicle detection: Fusing 3D LIDAR and color camera data. Pattern Recognition Letters. Elsevier. 2017

[14] Premebida C, Nunes U. Fusing LIDAR, camera and semantic information: A context-based approach for pedestrian detection. The International Journal of Robotics Research. 2013;**32**(3):371-384

[15] Dondrup C, Bellotto N, Jovan F, Hanheide M. Real-time multisensor people tracking for human-robot spatial interaction. IEEE International Conference on Robotics and Automation (ICRA); 2015

[16] Andreasson H, Magnusson M, Lilienthal A. Has something changed here? Autonomous difference detection for security patrol robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2007. pp. 3429-3435

[17] Sünderhauf N, Brock O, Scheirer W, Hadsell R, Fox D, Leitner J, Upcroft B, Abbeel P, Burgard W, Milford M, Corke P. The limits and potentials of deep learning for robotics. The International Journal of Robotics Research. 2018;**37**(4-5):405-420

[18] Correll N, Bekris KE, Berenson D, Brock O, Causo A, Hauser K, Okada K, Rodriguez A, Romano JM, Wurman PR, et al. IEEE Transactions on Automation Science and Engineering. 2016

[19] Wagstaff K. Machine learning that matters. In: Proceedings of the Twenty-Ninth International Conference on Machine Learning (ICML); 2012. pp. 529-536

[20] Moravec H, Elfes A. High resolution maps from wide angle sonar. In: Proceedings of IEEE International Conference on Robotics and Automation; 1985. pp. 116-121

[21] Hornung A, Wurm KM, Bennewitz M, Stachniss C, Burgard W. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. Autonomous Robots. 2013

[22] Thrun S, Burgard W, Fox D. Probabilistic Robotics. The MIT Press; 2005. ISBN:0262201623

[23] Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid I, Leonard JJ. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on Robotics. 2016;**32**(6):1309-1332

[24] Biber P, Ducket T. Experimental analysis of sample-based maps for long-term SLAM. The International Journal of Robotics Research. 2009;**28**:20-33

[25] Walcott-Bryant A, Kaess M, Johannsson H, Leonard JJ. Dynamic pose graph SLAM: Long-term mapping in low dynamic environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2012. pp. 1871-1878

[26] Saarinen J, Stoyanov T, Andreasson H, Lilienthal AJ. Fast 3D mapping in highly dynamic environments using normal distributions transform occupancy maps. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2013. pp. 4694-4701

[27] Hermans A, Floros G, Leibe B. Dense 3D semantic mapping of indoor scenes from RGB-D images. In: IEEE International Conference on Robotics and Automation (ICRA), Hong Kong; 2014. pp. 2631-2638

[28] Krahenbuhl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials. Advances in Neural Information Processing Systems. 2016;**24**:109-117

[29] McCormac J, Handa A, Davison A, Leutenegger S. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In: IEEE International Conference on Robotics and Automation (ICRA), Singapore; 2017. pp. 4628-4635

[30] Whelan T, Leutenegger S, Salas-Moreno RF, Glocker B, Davison AJ. ElasticFusion: Dense SLAM without a pose graph. Robotics: Science and Systems. 2015

[31] Ambrus R, Claici S, Wendt A. Automatic room segmentation from unstructured 3-D data of indoor environments. IEEE Robotics and Automation Letters. 2017;**2**(2):749-756

[32] Ambrus R, Bore N, Folkesson J, Jensfelt P. Autonomous meshing, texturing and recognition of object models with a mobile robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC. 2017. pp. 5071-5078

[33] Armeni I, Sener O, Zamir AR, Jiang H, Brilakis I, Fischer M, Savarese S. 3D semantic parsing of large-scale indoor spaces. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV; 2016. pp. 1534-1543

[34] Mura C, Mattausch O, Pajarola R. Piecewise-planar reconstruction of multi-room interiors with arbitrary wall arrangements. Computer Graphics Forum. 2016;**35**:179-188

[35] Sperry RW. Neurology and the mind-body problem. American Scientist. 1952;**40**:291-312

[36] Turing AM. Intelligent machinery. Journal of Machine Intelligence. 1970;**5**, different sources cite 1947 and 1948 as the time of writing

[37] Turing AM. Computing machinery and intelligence. Mind. 1950;**LIX**:433-460

[38] Steels L, Brooks RA, editors. The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.; 1995

[39] Vernon D. Cognitive vision: The case for embodied perception. In: Image and Vision Computing. Elsevier. 1 January 2008;**26**(1):127-140

[40] Bohg J, Hausman K, Sankaran B, Brock O, Kragic D, Schaal S, Sukhatme G. Interactive perception: Leveraging action in perception and perception in action. IEEE Transactions on Robotics. 2017;**33**:1273-1291

[41] Ferrari C, Canny J. Planning optimal grasps. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA); Vol. 3; 1992. pp. 2290-2295

[42] Frisby JP, Stone JV. Seeing objects, ch. 8. In: Seeing: The Computational Approach to Biological Vision. MIT Press; 2010. pp. 178-179

[43] Ulrich M, Wiedemann C, Steger C. Cad-based recognition of 3d objects in monocular images. In: IEEE International Conference on Robotics and Automation (ICRA); Vol. 9; 2009. pp. 1191-1198

[44] Hinterstoisser S, Lepetit V, Ilic S, Holzer S, Bradski G, Konolige K, Navab N. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian conference on computer vision; Springer; 2012. pp. 548-562

[45] Tjaden H, Schwanecke U, Schomer E. Real-time monocular pose estimation of 3d objects using temporally consistent local color histograms. In: Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE; 2017. pp. 124-132

[46] Brachmann E, Krull A, Michel F, Gumhold S, Shotton J, Rother C. Learning 6d object pose estimation using 3d object coordinates. In: European Conference on Computer Vision. Springer; 2014. pp. 536-551

[47] Krull A, Michel F, Brachmann E, Gumhold S, Ihrke S, Rother C. 6-DOF model based tracking via object coordinate regression. In: Asian Conference on Computer Vision. Springer; 2014. pp. 384-399

[48] Kehl W, Milletari F, Tombari F, Ilic S, Navab N. Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation. In: European Conference on Computer Vision. Springer; 2016. pp. 205-220

[49] Lepetit V, Moreno-Noguer F, Fua P. Epnp: An accurate o(n) solution to the pnp problem. International Journal of Computer Vision. 2009;**81**(2)

[50] Janai J, Guney F, Behl A, Geiger A. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. arXiv:1704.05519v1. 2018

[51] Tas OS, Salscheider NO, Poggenhans F, Wirges S, Bandera C, Zofka MR, Strauss T, Zollner JM, Stiller C. Making Bertha Cooperate—Team AnnieWAY's entry to the 2016 Grand Cooperative Driving Challenge. IEEE Transactions on Intelligent Transportation Systems. 2018;**19**(4):1262-1276

[52] Hawes N, Burbridge C, Jovan F, Kunze L, Lacerda B, Mudrova L, Young J, Wyatt J, Hebesberger D, Kortner T, Ambrus R, Bore N, Folkesson J, Jensfelt P, Beyer L, Hermans A, Leibe B, Aldoma A, Faulhammer T, Zillich M, Vincze M, Chinellato E, Al-Omari M, Duckworth P, Gatsoulis Y, Hogg DC, Cohn AG, Dondrup C, Pulido Fentanes J, Krajnik T, Santos JM, Duckett T, Hanheide M. The STRANDS project: Long-term autonomy in everyday environments. IEEE Robotics and Automation Magazine. 2017;**24**:146-156

[53] Fäulhammer T, Ambruş R, Burbridge C, Zillich M, Folkesson J, Hawes N, Jensfelt P, Vincze M. Autonomous learning of object models on a mobile robot. IEEE Robotics and Automation Letters. 2017;**2**(1):26-33

[54] Krajnik T, Fentanes JP, Cielniak G, Dondrup C, Duckett T. Spectral analysis for long-term robotic mapping. In: IEEE International Conference on Robotics and Automation (ICRA), Hong Kong. 2014. pp. 3706-3711

[55] Beyer L, Hermans A, Leibe B. Biternion nets: Continuous head pose regression from discrete training labels. German Conference on Pattern Recognition. 2014:157-168

[56] Beyer L, Hermans A, Leibe B. DROW: Real-time deep learning-based wheelchair detection in 2-D range data. IEEE Robotics and Automation Letters. 2017;**2**(2):585-592

[57] Fentanes JP, Lacerda B, Krajnik T, Hawes N, Hanheide M. Now or later? Predicting and maximising success of navigation actions from long-term experience. In: IEEE International Conference on Robotics and Automation (ICRA), Seattle, USA. 2015. pp. 1112-1117

[58] Santos JM, Krajník T, Fentanes JP, Duckett T. Lifelong information-driven exploration to complete and refine 4-D spatio-temporal maps. IEEE Robotics and Automation Letters. 2016;**1**(2):684-691

[59] Triebel R, Arras K, Alami R, Beyer L, Breuers S, Chatila R, Chetouani M, Cremers D, Evers V, Fiore M, Hung H, Islas Ramírez OA, Joosse M, Khambhaita H, Kucner T, Leibe B, Lilienthal AJ, Linder T, Lohse M, Magnusson M, Okal B, Palmieri L, Rafi U, van Rooij M, Zhang L. SPENCER: A socially aware service robot for passenger guidance and help in busy airports. Field and Service Robotics: Results of the 10th International Conference. 2016. 607-622

[60] Szeliski R. Computer vision: Algorithms and applications ser. In: Texts in Computer Science. New York, NY, USA: Springer-Verlag New York, Inc.; 2010