By Florenz Graf, Jochen Lindermayr, Çağatay Odabaşi, and Marco F. Huber

The long-term vision for robotics is to have fully autonomous mobile robots that perceive the environment as humans do or even better. This article transfers the core ideas from human scene perception onto robot scene perception to contribute toward a holistic scene understanding of robots. The first contribution is to extensively survey and compare state-of-the-art robot scene perception approaches with neuroscience theories and studies of human perception. A step-by-step transfer of the perceptual process reveals similarities and differences between robots and humans. The second contribution represents an analysis of the status quo of holistic robot perception approaches to extract to what extent the perceptual capabilities of humans have been reached. Building on this, the gaps and potentials of robot perception are illustrated to address future research directions.

Toward Holistic Scene Understanding

A Transfer of Human Scene Perception to Mobile Robots

©SHUTTERSTOCK.COM/BLUE PLANET STUDIO

Digital Object Identifier 10.1109/MRA.2022.3210587
Date of current version: 13 October 2022

Introduction

The last few years indicated fast technological improvements in the artificial intelligence of robots. The International Federation of Robotics records a market growth of 12% in 2020 of professional robots used for various applications, such as transport, inspection, cleaning, medical, or hospitality [1]. The report forecasts exponential growth for the upcoming years. Robots will become a part of human society. Acting side by side and collaborating within the same environments motivate robot perception being consistent with human perception.

In the past, robots mainly fulfilled highly customized tasks in industrial applications independently and separated from humans [2]. People adapted the environments to the application requirements. However, adaptations for the robots' needs are especially undesirable in nonindustrial environments. Current products on the market mirror this issue through specialized applications. These products either fulfill a single task autonomously, such as vacuum cleaners and lawnmower robots executing their routine independently from humans, or interact with humans physically with a low autonomy level [1]. Therefore, special attention must be directed to scene perception as enabling technology to break this tradeoff between robot proximity to humans and autonomy. Already in 1993, the psychologist Ulric Neisser mentioned that "without perception there is no knowledge" [3].

The goal of holistic scene perception is to understand the scene by "considering the geometric and semantic context of its contents and the intrinsic relationships between them" [4]. Scene perception is making sense of real-world scenes as a whole by enabling the interaction with the scene [5]. On the one hand, it offers new opportunities for multiuse applications regarding the cost per function and physical assistance systems. On the other hand, it increases autonomous capabilities by overcoming challenges, such as the occlusion of objects; instance-specific handling of dynamics; or spatial–temporal reasoning of complex situations. Holistic scene perception will enable safe and complex behaviors for any collaboration.

Increasing attention has been given to specific areas of perception. The approaches achieve excellent results in semantic extraction, such as object detection or image classification [6]. However, in the real world, it is not sufficient that robots solely understand single pieces of the environment. Much more, various scene information needs to be understood concurrently in the context to reason within dynamics, uncertainties, and incompleteness for high-level control [7, Ch. 23]. Little is known about holistic scene perception, aiming to understand the scene in this integrated manner needed for a large spectrum of real-world applications.

Motivated by the performance of human perception, we distanced ourselves in this research from robotic approaches and studied theories and experiments on human perception. Neuroscience research has been extensively studied since Potter [8] and Biederman [9] in the 1970s. It has reached international bandwidth since the 2000s [10]. The latest research provides a comprehensive overview of human scene perception [5], [10]. This article transfers the core theories of psychological studies on human scene perception to robots and thereby reveals similarities and differences. The comparison of human perception with artificial perception is not new [11], [12], [13]. However, we will focus on the scene perception as a whole using a top-down view of the robots' status quo.

Methodology

Figure 1 describes the methodology of this study. First, we analyzed current mobile robot applications where we identified a trend toward robots in everyday life, covering multiple use cases. For future applications, scene perception will play a

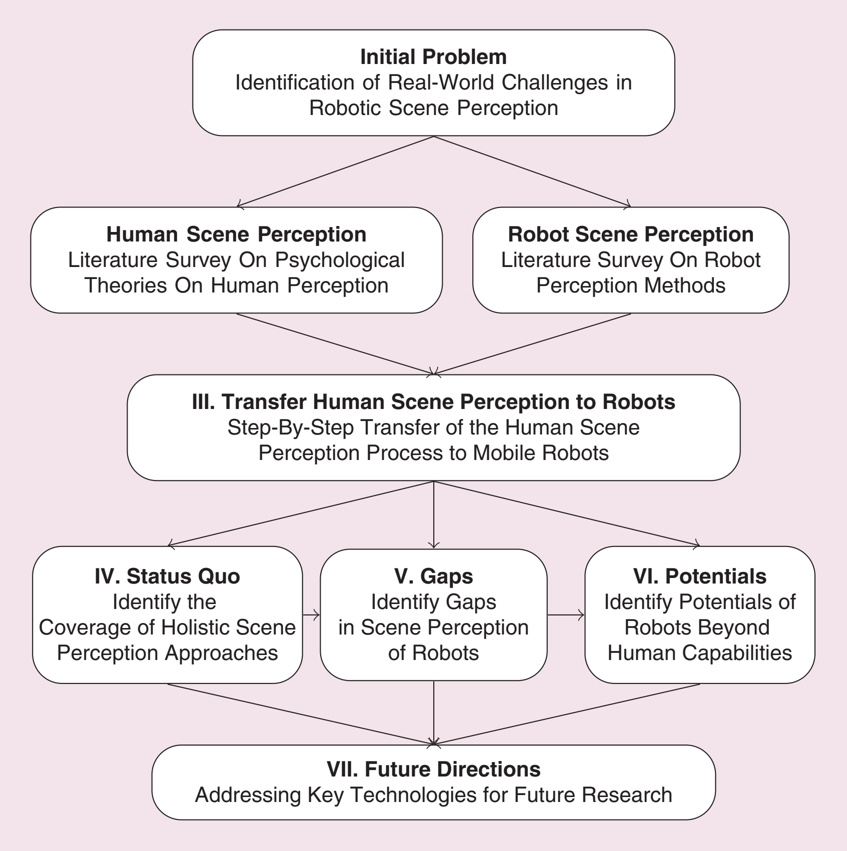
Initial Problem
Identification of Real-World Challenges in Robotic Scene Perception
Human Scene Perception
Literature Survey On Psychological Theories On Human Perception
Robot Scene Perception
Literature Survey On Robot Perception Methods
III. Transfer Human Scene Perception to Robots
Step-By-Step Transfer of the Human Scene Perception Process to Mobile Robots
IV. Status Quo
Identify the Coverage of Holistic Scene Perception Approaches
V. Gaps
Identify Gaps in Scene Perception of Robots
VI. Potentials
Identify Potentials of Robots Beyond Human Capabilities
VII. Future Directions
Addressing Key Technologies for Future Research

Figure 1. A methodological approach with section references. III: The "Transferring Human Scene Perception to Mobile Robots" section. IV: The "The Status Quo of Holistic Scene Perception" section. V: The "The Gap to Human Scene Perception" section. VI: The "The Potentials of Robots" section. VII: The "Future Directions" section.

key role as it provides comprehensive scene knowledge to the cognitive intelligence of robots. Therefore, we analyzed state-of-the-art robot scene perception approaches and theories of human scene perception to elaborate on a transfer. Research in both domains offers a comprehensive overview providing fundamental knowledge as the starting point. The "Transferring Human Scene Perception to Mobile Robots" section summarizes the transfer with a step-by-step analysis of the similarities and differences between human and robot scene perception. Based on this, we addressed the status quo of holistic scene perception approaches of robots within a qualitative evaluation (see the "The Status Quo of Holistic Scene Perception" section) to extract how far robotic scene perception reached the performance of human perception to reveal gaps (see the "The Gap to Human Scene Perception" section) and potentials (see the "The Potentials of Robots" section). Finally, we propose future directions for robotic scene perception (see the "Future Directions" section).

This article transfers the core ideas from human scene perception onto robot scene perception to contribute toward a holistic scene understanding of robots.

Transferring Human Scene Perception to Mobile Robots

Humans perceive the environment by recognizing scene information through different senses, such as sight, hearing, smell, touch, or taste. Herewith, researchers discovered that people primarily rely on their visual perception for perceiving the environment due to the richness of information [14]. Although the process of human perception is unknown in detail, most theories define perception as the process of "recognizing (being aware of), organizing (gathering and storing), and interpreting (binding to knowledge) sensory information" [15, Ch. 3]. We transfer these three process steps to robot scene perception to reveal similarities and differences to the latest robotic research (Figure 2).

The Recognition of Information

The recognition step processes sensory information to make it understandable for the subsequent perception tasks. Using the sensory input, humans convert the observations from the scene into understandable information. Whereas vision benefits from a high amount of information, other modalities, such as haptics, are robust to noise and independent of light exposure to get properties being insufficiently recognized by other senses [16]. For example, the eye serves as a transducer to convert light into the optic nerve by five layers of cells (particularly photoreceptors) within the retina [17]. The nerve cells transmit information by electric signals to the brain for processing [18]. The processing itself divides into preattentive and postattentive recognition.

In preattentive processing, activities related to low-level vision are usually associated with the extraction of certain physical properties of the visible environment, such as depth, shape, color, object boundaries, or surface material properties [10], [19], [20]. Humans are not capable of processing all sensory information at once. Therefore, the postattentive processing extracts scene information based on attention areas to overcome this issue [15]. It compresses high-level features, such as object classification or redetection [21], [22]. Herewith, humans receive information from hierarchical levels of abstraction [23], [24] corresponding to a region of interest, triggered by the attention that provides the semantic of sensory input.

Similar to humans, the recognition of robots can be divided into pre- and postattentive recognition. Equivalent to the human senses, robots use various sensors to recognize information from the environment. The most common sensor modalities are visual sensors, such as cameras, lidar, ultrasonic, and radar. They are capable of providing the color and/or range information of the surrounding. The recognition of visual features is the most popular modality for robots due to the information richness. However, few approaches investigate other modalities, such as acoustic

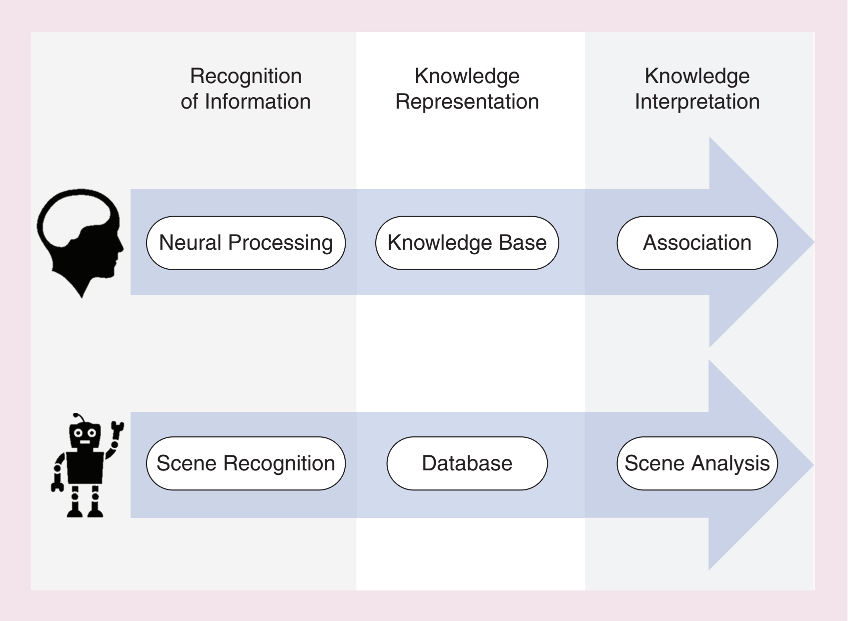

Figure 2. The process of human perception transferred to mobile robots.

perception to recognize objects [25], map the environment [26], [27], or avoid obstacles [28] and haptic perception to actively touch objects [16]. The linking of action to perceive the information of interest is known as *interactive perception*, which is a common human exploration strategy. For instance, if humans cannot recognize an object by vision and touch, they will take different interactions to obtain information from other sensory channels [25].

The usage of smelling and tasting for robots is not popular. One rare example presents a navigation approach using smell for odor source localization [29]. A different niche develops a tasting sense using IR spectroscopic technologies or chemical sensors to discriminate food [30], [31]. Driven by computer vision (CV), 2D image processing is the fundamental pillar for robot recognition tasks. However, 3D methods for scene and object reconstruction gained importance for robots. Various surveys provide an overview of the state-of-the-art robot recognition tasks [32], [33]. Established robotics approaches focus on specific recognition subareas. The preattentive recognition of primitives comprises methods on normal estimation [34], [35], segmentation [36], [37], edge [38], and simple shape (such as planes and cylinders) [39] detection. The methods usually make use of the entire sensory input. However, robots prefilter or scale the data to achieve the appropriate results on their capabilities. For instance, image-based feature recognition can process high-resolution images by making use of image pyramids [40]. Herewith, the extracted features of a lower image resolution deliver discrete regions of interest by keeping the accuracy of the original scale.

Postattentive processing comprises higher-level recognition capabilities based on preattentive processing. Examples are extracting the semantic by (re)-detection of objects [41], [42], humans [43], [44], and places [45]; 3D reconstruction of the scene by Simultaneous Localization and Mapping (SLAM) [32], [46], [47]; or enriching metric information by semantics, known as *semantic segmentation* [33], [48], [49].

Equal to humans, robots use preprocessed attention areas within postattentive processing. For instance, feature-based SLAM methods use sparse feature points to postattentively register a set of sensor data and to detect loop closures to compensate for the mapping drift [46], [47]. Besides, this example shows how humans and robots use different sensory sources to achieve highly accurate localization within the scene. Humans use walked steps combined with eye information for perceiving ego-motion, which improves localization. For illustration, imagine walking a distance with closed eyes. The motion drift sooner or later leads to a loss of localization. Therefore, a series of interconnected spatial–visual features provides the absolute localization within the known scene [50].

Equivalent to humans, the latest robotic SLAM approaches provide similar techniques. Wheel odometry, visual odometry, or an inertial measurement unit (IMU) provides the ego-motion. Laser scanners or camera-based methods simultaneously estimate the pose of landmarks to compensate for the ego-motion drift. Also, some approaches use visual landmarks for place redetection to close the trajectory loop while mapping [46]. The usage of high-level features enables robots to add semantics to metric information. Knowing the semantics of an area in space helps robots to interpret the scene. It allows, e.g., the exclusion of dynamic objects like people while mapping.

A special challenge of robot perception is the recognition of scene information from different locations and times. For instance, a robot recognizes an object in the scene and simultaneously tracks the object in its field of view (FOV) as long as it is visible. While previous research proposed using a Kalman filter [51] or particle filter [52], the latest research utilizes deep learning-based tracking, such as with a convolutional neural network (CNN) [53], [54] or a Vision Transformer (ViT) [55]. ViT, coming from natural language processing (NLP), splits images into fixed-size patches that gained popularity due to their superior performance on continuous data streams, needed for mobile robots [56]. Since the object and the robot could move, occlusions, truncation, or invisibility due to sensor noise (as mentioned previously) must be handled. When the same object appears again, a reidentification (Re-ID) to reallocate the ID is beneficial to better understand the scene. Modern approaches of Re-ID have been proposed that are similar to tracking the usage of a CNN [57], [58], ViT [59], [60], or end-to-end approaches [61].

Knowledge Representation

The second step of the perception process represents environmental information. A knowledge base manages recognized information from different sources and levels of abstraction, times, and places in a centralized and ordered structure. This structure includes understandable information about the scene. The function of the knowledge representation within the human brain has been a controversial topic since the so-called *gestalt theory*. Modern attempts such as Wagemans and Kimchi [24] or Hommel et al. [62] reveal spatial layouts, organized hierarchically, that represent the human perception memory. A relationship-focused multilevel hierarchical structure of parts represents environmental information.

The transfer of the main functionalities of human knowledge representation to robots requires a complex memory structure focused on flexibility. The knowledge representation must be capable of merging observations and interpretations from different sources and times. For instance, recognized information, such as shape, texture, posture, state, probabilities, and trajectory (compare the “The Recognition of Information” section), must be managed in real time within the knowledge base. This issue sets high requirements for the underlying knowledge base as every piece of environmental information needs a known structured representation. Furthermore, the knowledge base includes initial and postprocessed knowledge. Robots usually store and represent the scene knowledge in a database [63], [64], [65], allowing them to deploy a human-understandable ontology that conceptualizes multiple entities within a domain and their relationships [66].

The scene representation comprises various perception-related information, such as extracted spatial features for navigation, manipulation, and semantically enhanced maps [67]. The requirements for robotic knowledge representation are high. Ideally, it must be real time capable; generic; scalable; flexible in structure; and able to update and extend during the robot's lifetime as well as easy to connect for access and data sharing. Generally, there are two categories of databases: graph based and document based. Both seem suitable for this task as they provide comprehensive features to cover these requirements [68], [69]. Graph-based databases, also known as *relational database management systems*, represent knowledge through relations.

Herewith, it is necessary to explicitly define relations through a common format to link semantics through an ontology [70]. In contrast to relational databases, document-based databases represent data in JavaScript Object Notation-like documents without the need for relations or predefined structures. Nevertheless, these databases provide features for querying or indexing the data to model dependencies and relations implicitly dynamically. For instance, Kunze et al. [63] propose spatial and temporal indexing for query relations within their document-based database. Besides, linking the knowledge base for decision making [71] and providing the represented knowledge to robot actions, such as manipulation or moving, enables reactive behaviors.

The comparison of scene knowledge representation indicates that robots have advantages compared to humans. First, the artificial scene knowledge representation has no memory loss due to an almost unlimited storage capacity. Second, robots can easily share and make use of foreign perception, while humans have to transfer knowledge into an appropriate modality, such as verbal communication. Additionally, humans are limited in the range of information exchange without technical assistance. In contrast, robots can share data in their original format over networks. The sharing of perceptional information enables robots to directly exchange data with the infrastructure or other robots. Another difference between human and robot knowledge representation is the capability of robots to start with a preinstalled environment perception model. Robots can use the prior information of a building information model [72] or a partial or fully premapped environment [64]. Using prior scene knowledge reduces the setup time, especially when using multiple agents.

Knowledge Interpretation

Based on the available scene knowledge, this perception step interprets existing knowledge to make sense using cognitive capabilities. It has been proven that humans interpret the scene; however, it is unclear how this is executed within the brain. The research by Isik et al. [73] found that the human brain starts recognizing view-invariant observations, such as human actions, quickly, in around 200 ms. This suggests that the brain uses the form, as well as the motions, to represent states. Furthermore, previous work [74] proposes that the human brain benefits from causal relations, such as temporal continuity and spatial relations, among objects, humans, and their actions. The prerequisite is that a known structure represents the knowledge (see the "Knowledge Representation" section). A high-level scene analysis encodes spatial and temporal relations between instances [75].

For most examples, the interpretation of scene knowledge is a trivial task for humans due to lifelong learning. The interpretation of perceived information has been trained with preknowledge, dependent on the culture [76, Ch. 14]; context; and situation itself [77]. Therefore, all people have a unique perception system based on and enriched by their environment. Thus, human interpretation is neither predictable nor always the same. The famous duck–rabbit illusion [78] indicates that even the season could differ the interpretation of the scene. Therefore, perception is influenced by environmental factors [77].

For robots, the interpretation of the scene is still an unsolved problem [79]. The challenge is to generate and make use of high-level semantic knowledge to reason about the present scene. There is no commonsense model that could be applied to every environment without the adaptation of the interpretations. The association of scene knowledge across multiple dimensions, such as time and space, with or without relational dependencies between single pieces of information, leads to this new knowledge that improves scene understanding. For instance, a spatial–temporal scene analysis could reveal daily habits, such as when, how, and how often we go to the kitchen to fetch a coffee. This example shows that the kind of required high-level information is environment specific as well as use case specific. The goal is to understand what is involved and what to do when and with which objects. The use of high-level semantics within the scene is fundamental for complex robot behavior tasks. Herewith, we identify two types of interpretations.

On the one hand, there is research aimed at reconstructing and interpreting the structured part of the scene, such as room segmentation, junction detection [32], or occlusion reasoning for simple shapes [80]. On the other hand, there is research focusing on the unstructured part that goes deeper into handling dynamics. The approach presented in [81] is one of the rare examples in which already collected scene information is used to gather new information. Observations of objects are anchored in the scene model, which provides basic tracking functionality. In combination with knowledge about the whole scene, including other objects and their spatial and semantic relations, this is used for reasoning about the state of occluded objects, which improves tracking and hence the whole scene state. The tracking of instances over multiple observations enables further interpretations, such as detecting their action [82]. The actions of the people together with environmental semantics are valuable input for a robot since they usually share an environment. So, they need to understand the actions and fulfill their tasks proactively. For instance, if the person is cooking in a room, it is very likely that this room is a kitchen in which the robot can act accordingly [83].

Previous work [84] focuses on learning human actions by observing humans in their daily life. They merge joint motions and locations concerning the landmark points on the map. Kostavelis et al. [85] propose using object recognition and skeleton-based action recognition to make their social robot understand human actions. They deploy the robot in real home environments to test their system. The previous work [86] employs a long short-term memory [87] network on the robot to greet the user. The background may be misleading for the algorithms that are using appearance features. That is why generating action proposals may improve performance, and this is essential for mobile robots because the background in the robot view is dynamic [88]. In contrast to the spatial–temporal interpretations of a single instance, there are a few approaches that interpret relations within multiple instances. For example, Philipp et al. [89] propose to use Bayesian networks to estimate on which object the user's attention focuses.

The attention-sensitive functionalities as high-level scene interpretations are essential for improving human–robot interaction. More complex interpretations go into the affordance estimation [90]. Herewith, robots know what can be done with objects based on past observations, probabilities, and personal data (such as emotions, preferences, and relations). For instance, knowing that a pod can be used for cooking offers novel capabilities for robots. Affordances link perception into the cognitive capabilities that are fundamental for interpreting scenes.

Implications

Recent research on robot perception displays similarities to fundamental theories on human perception. Figure 3 visualizes the transferred scene perception process. Humans and robots recognize information from the scene by making sensor data understandable. The human brain and the robot storage, respectively, represent the knowledge in multiple layers in a known structure. Compared to humans, robots are capable of using external perceived data within their original format, whereas humans have to exchange perceived data through verbal, visual, or written communication. The interpretation of knowledge over multiple dimensions, such as time, space, and relations, enables a high-level scene understanding that improves the cognitive intelligence of robots. Herewith, humans highly benefit from life-long learning, whereby robots can benefit from shared and initial data.

In recent years, robotic recognition and interpretation use deep learning, such as CNNs and generative adversarial networks [91], based on artificial NNs (ANNs), to solve a very specific task [92] as described previously. Inspired by biological NNs [93], ANNs consist of multiple layers of connected artificial neurons [94]. The weight assigned to these neurons relies on an initial ANN training using a high amount of labeled data. This in turn enables the ANN to compute labels of unknown data via inference. In particular, deep learning-based approaches, such as ViTs, gained popularity in robotics due to their superior performance on continuous data streams. ViTs, coming from NLP, split images into fixed-size patches [56]. Machine learning, such as deep learning, achieves outstanding results for many recognition tasks [95], [96]. However, the division of preattentive and postattentive feature processing can neither be strictly adhered to nor easily extended. The inseparability of ANNs challenges their flexibility and reusability due to the "black box" characteristic. Additionally, they require enhanced tensor computation power that needs special attention when pushing mobile robots to the real world [92].

Recent research on robot perception displays similarities to fundamental theories on human perception.

The Status Quo of Holistic Scene Perception

The transfer reveals similarities and differences in the perception process between humans and robots. But how much does the status quo of robotic scene perception cover holistic capabilities? To answer this question, we first deal with the challenges of fundamental technologies to extract possible boundaries for human-like perception. Afterward, we analyze the scope of most holistic robotic approaches for everyday environments to answer how much they already cover the holistic scope.

From Narrow Toward Holistic Scene Perception

Research on robot scene perception focuses on sensor-close processing by affordable sensors. The availability of cheap

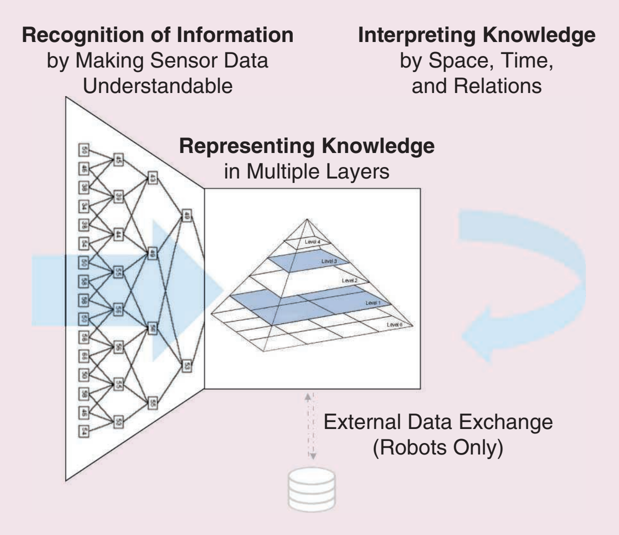

Figure 3. A summary of the three steps of the transferred scene perception.

red-green-blue cameras and 2D laser scanners set the entry barrier in robotic developments, e.g., for student or hobby projects, relatively low. Investigations mainly focus on a deep learning-based detection of common objects in images or on mapping the environment using SLAM. 2D methods in object detection benefit from matured research in CV. It reached high accuracy due to challenges, such as the Pascal Visual Object Classes Challenge [97] and the Large Scale Visual Recognition Challenge [98], by public datasets of labeled data, just like the Common Objects in Context (COCO) dataset [99]. For the application on robots, these 2D image-based approaches must be transferred to time-continuous 3D processing merged with SLAM techniques providing the ego-motion of the robot to reconstruct the environment.

The first identified gap in robotic scene perception is the missing usage of multiple sensor modalities as the input source.

Only comparatively new research focuses on 3D multiframe scene recognition solving the high computational performance with mobile graphic accelerators [49], [100], [101]. However, both the representation and the high-level interpretation of scene knowledge are not focused. Indeed, as described by Neisser [3], solely the knowledge that was recognized can be represented or interpreted. The focus on sensor-close recognition in combination with visual and mobile robotic challenges, such as changing visual appearance, as well as the limited computation resources, explains this observation. These aspects also explain why research on the representation and interpretation of scene knowledge is comparatively rare. As a result, robotic scene perception in real-world applications is focused on a concrete use case utilizing highly optimized bottom-up perception pipelines with narrow functionality. For instance, industrial environments have adapted to the perception capabilities of robots.

Herewith, robots are enabled to fulfill narrow perception tasks with high accuracy. In particular, deep learning-based recognition systems are often used as monolithic black box systems, being hard to combine efficiently without probing deeply into a technical level. Research articles mostly bury these challenges by describing specific techniques [21]. However, robots in everyday environments need to fulfill multiple recognition tasks simultaneously as a requirement for complex and extensive behaviors. Open source frameworks such as the Robot Operating System (ROS) [102] provide, thanks to its large community, many tutorials as well as open source basic functionalities that are suitable for creating a powerful overall performance based on single software pieces. Standardized communications between various software components enable robots to use, fuse, and analyze multiple data streams.

However, on the one hand, paralyzing multiple narrow perception pipelines is challenging regarding performance. On the other hand, splitting pipelines into multiple steps, to reuse, e.g., preattentive information, is not easy when using monolithic black box models. Nevertheless, few researchers worked on integrated top-down approaches that aim to perceive the scene as a whole by combining multiple methods.

The Holistic Scope of the Latest Integrated Approaches

There are a few approaches in research that integrate multiple scene perception tasks into a single framework. In contrast to the narrow perception techniques, they achieve a more holistic understanding of the scene. These approaches simultaneously recognize environmental information over a long time. Storing these data in a known structure offers a new potential for a more complex spatial and temporal interpretation of the scene. We looked deeper into these approaches to extract how much they cover a holistic scene understanding. In the following, a comparison of the perceptual capabilities should answer this question. As criteria, we select research approaches covering more than a tabletop scene; reconstruct in real time; and represent scene data by multiple types of instances, such as semantics. However, each approach set a different focus starting at sensor-close processing, such as on paralyzing, fusing, and handling of dynamics going to ontology-based reasoning.

Table 1 shows the perceptual properties of these approaches, divided into the three steps of the perception transfer presented in the "Transferring Human Scene Perception to Mobile Robots" section. If we could not find the details to a criterion, we marked it either as not available (N\A) or not specified (NS). The approaches of Table 1 use a 3D camera as sensory input providing the color and depth information of its FOV. Additionally, some approaches use an IMU to support visual odometry for a more precise estimate of the ego-motion, e.g., needed for drones and wheeled-based robots. Based on the sensory input stream, the presented approaches combine several recognition techniques to reconstruct a virtual scene model. However, they cover the recognition differently.

KnowRob [65], a knowledge representation and reasoning framework, solely offers an interface for individual visual recognition systems. Wyatt et al. [67] limit the reconstruction to a metric map without recognizing semantics by vision sensors. Its scene recognition is trained from visual properties by a human tutor using supervised learning. Similarly, the SOMA framework restricts the reconstruction to a metric map but enhances objects by a CNN for color image-based object detection. SOMA aims at understanding changes in everyday environments by perceiving geometries and semantics. Alternatively, the approach of Suchan et al. [103] enhances the metric map by detecting walls, which are used for a clustering algorithm to generate a floor plan. The other approaches investigate further into a fully 3D metric–semantic reconstruction of the scene.

The SLAM++ project of Salas-Moreno et al. [104] is an early approach from 2013 that concentrates on semantic mapping. It consists of an object-based SLAM that uses object recognition trained on a database of scanned object models. It is capable of detecting changes in the environment, such as moving objects. Fusion++ [100] set its focus on semantic mapping. It is similar to SLAM++ but runs a mask region-based CNN (R-CNN) object segmentation to initialize a truncated signed distance field (TSDF) reconstruction for each object. Rosinol et al. [105] recently published Kimera, a multilayer spatial scene perception framework aiming to close the gap between human and robot scene perception. Kimera uses a metric–semantic SLAM to perform a full mesh reconstruction by TSDF volumes with its semantic on top of localization. It recognizes building structures as well as objects from a CAD model match. In addition, human detection and pose estimation extend the dynamic scene information [106]. The recognition techniques feed its scene information into the knowledge base, where they are represented and connected in different ways.

Fusion++ and SLAM++, which concentrate on the semantic mapping, are not providing details of their knowledge representation. In contrast, the approach of Wyatt et al. [67] focuses on the representation of knowledge gaps and uncertainties. A layered structure of proxies, unions, and beliefs represents the spatial scene inside a relational database. They validate by experiments in a lab that a human tutor is capable of helping a robot fill a knowledge gap through verbal conversation. The robot asks for missing visual features, such as the color and shape, to prove an object is believed to close a knowledge gap.

Kimera, which extensively covers the recognition, is also concentrating on spatial knowledge representation by multiple hierarchical layers, separated by the semantic. The spatial layers comprise the metric–semantic mesh, objects, structures, rooms, and buildings. Dynamic scene graphs simultaneously update scene information by linking the spatial scene

Table 1. An overview of integrative robotic scene perception approaches.

	Recognition of Information				**Knowledge Representation**			**Knowledge Interpretation**	
	Sensory Input	**Reconstruction**	**Static Instances**	**Dynamic Instances**	**Database**	**Knowledge Structure**	**Scene Representation**	**Spatial–Temporal**	**Reasoning**
Wyatt et al. [67]	3D camera	Metric	Objects from user input	N/A	Relational	Multilayer spatial representation by proxies, unions, and beliefs	Point map	Place classification	Belief verification by human
SLAM++ (Salas et al. [104])	3D camera	Metric–semantic	N/A	Objects by scan model match	NS	Single-layer spatial object graph	Metric–semantic mesh	N/A	N/A
Suchan et al. [103]	3D camera	Metric–semantic	Walls by planes	Human pose detection	NS	Spatial–temporal representation of entities	Metric map and semantic by primitives	Human activities, spatial relation, and pattern	Human-centered common sense
Fusion++ (McCormac et al. [100])	3D camera	Metric–semantic	Objects by R-CNN	N/A	NS	Single-layer spatial object graph	Metric–semantic mesh	N/A	N/A
SOMA (Kunze et al. [63])	3D camera, 2D laser scanner	Metric	N/A	Objects and people by CNN	Document based	Observation, semantic, and interpretation layer	Point map, objects by pose and bounding box	Human activities	N/A
KnowRob (Beetz et al. [65], Beßler et al. [90])	NS	NS	NS	NS	Relational	Ontology graph, multilevel of metric–semantic, logic, and episodic memories	Mesh and poses	Episodic memory for reasoning	Hypotheses verification, inner world, and motion control
Kimera (Rosinol et al. [105], [106])	3D camera, IMU	Metric–semantic	Building structures and objects by CAD model match	People by pose detection	NS	Hierarchical graph connects spatial layers	Metric–semantic mesh	Place and room classification	N/A

N/A: not available; NS: not specified.

information within the layers. The knowledge representation of SOMA comprises three layers inside a document-based database: the observation layer, semantic layer, and interpretation layer. The scene information is organized similarly to Kimera with a spatial layout of hierarchical structures connected by graphs. Suchan et al. [103] propose using ontologies and formal characterization to represent knowledge by space and motion. The knowledge representation of KnowRob is the most powerful due to the usage of description logic for representing and connecting knowledge within multiple levels by the Web Ontology Language (OWL). The framework links an inner world and virtual, logical, and episodic memories [65].

Based on the represented knowledge, a few approaches look deeper into the interpretation. For instance, the second pillar of KnowRob is knowledge-based reasoning to learn general knowledge. An object class-related affordance extraction reasons what to do with objects [90]. Suchan et al. [103] show by temporal human detection and object recognition how to use reasoning to enhance the scene understanding. The use of logical spatial relations between instances (e.g., an object is located to the left of another object) and the detection of human activities represent valuable information about the scene.

A fundamental difference between robots and humans is the acquisition of perceptional understanding.

The presented approaches indicate that they already perform more than one perception process step on an advanced level. Due to the different focus of every approach, none sufficiently covers holistic scene perception. The degree of coverage is difficult to quantify due to missing metrics in research. However, the overview (see Table 1) offers a starting point to identify gaps and potentials compared to human perception.

The Gap to Human Scene Perception

How well does robot scene perception mimic human-level performance nowadays? Previous research compared the perceptual performance of robots with children's age, such as Szeliski [107], who claimed that the computer vision reached the level of a two-year-old child. This comparison might not be beneficial since robot perception is not quantifiable in human terms as there is a specific rather than a broad perceptual skill development of robots. Following this hypothesis, this article details major gaps in robot perception in everyday scenes.

The Nonusage of Sensory Modalities

The first identified gap in robotic scene perception is the missing usage of multiple sensor modalities as the input source. Humans use different senses, providing the opportunity to fuse and rely on the optimal sensor since each sensor modality is affected, such as vision by illumination, color, occlusion, and posture of objects [25] (see the "Recognition of Information" section). The hardware design of robots allows them an optimal sensory input due to flexible sensor choice, amount, and alignment. On the one hand, this underlines the statement of Premebida et al. [33] that robotic recognition tasks, such as vision-based object classification, could deliver higher performance than humans. However, on the other hand, all the robotic approaches of the "The Holistic Scope of the Latest Integrated Approaches" section limit the scene perception to visual sensors, except for ego-motion sensors, as a single modality. Although vision sensors provide the highest information content of the scene, using a single modality is insufficient for the situations mentioned previously (the localization of acoustics; haptic feedback of obstacles; and darkroom problem). Therefore, current integrated scene perception approaches are not able to deploy different senses in the way that humans do.

Different Perceptual Learning

A fundamental difference between robots and humans is the acquisition of perceptional understanding. Perceptual learning was first defined by Gibson [108] as "any relatively permanent and consistent change in the perception of a stimulus array, following practice or experience with this array." Thus, humans learn an individual perception without initial knowledge that is constantly adjusted and influenced by society and the environment. The learning enables humans to achieve an optimal perception within the surrounding since, in particular, high-level scene interpretations depend on the subjective impression that is difficult to generalize (see the duck–rabbit illusion [78]). In contrast to humans, popular perceptual techniques of robots use rigid learning strategies, which do not offer adaptation. It would require retraining or parameter adjustment of the initially deployed models. Moreover, these approaches provide neither the flexibility for extension nor adaptation of the perception capabilities over time.

All presented integrated robotic scene perception approaches (see Table 1) build up prelearned skills except for the method of Wyatt et al. [67], which allows modifications of the perception after deployment (see the "The Holistic Scope of the Latest Integrated Approaches" section). Especially in the future, when robots will be deployed for long periods of time, the initial perception will become obsolete unless the perception is adjusted continuously to the environment. Therefore, a gap concerns the learning of perceptual capabilities during runtime. Although research on robot learning, such as domain adaptation [109] and continuous learning (lifelong learning, perceptual learning, and never-ending learning) [110], reaches back almost 30 years [111], it is still rarely used in practical applications. A common strategy is to learn from demonstration [112]. For instance, a human could teach the robot to adjust a generic perception to changes in the environment. The human can approve or decline estimates of the perception system via a user interface, such as classifying new objects or teaching novel actions.

A possible solution for supervised learning by a human tutor has been presented by Wyatt et al. [67]. However, enabling robots to adapt to the environment requires new techniques. On the one hand, as proposed by Wyatt et al. [67], this technique could close knowledge gaps. On the other hand, learning environment-adapted perception skills once with a tutor will not avoid adjusting the perception within the deployed environment.

The Lack of Commonsense Perception

Our third identified gap is the lack of commonsense perception caused by specific recognition capabilities. Fed by a high amount of training data, the artificial system becomes an expert for narrow recognition tasks. The previous arguments reveal several applications that hit the performance for specialized tasks. Although the recognition performance of a specific perception task can outperform humans, the fundamental issue is the lack of flexibility due to the required retraining or missing capability to extend recognition. Therefore, even the latest approaches achieve just low overall recognition performance (see the "The Holistic Scope of the Latest Integrated Approaches" section).

Humans have individual perception skills depending on various influences, such as profession, culture, and age, since humans learn from practice and experience [108]. However, in a society, there are perception skills simplifying a commonsense understanding. This enables humans to execute trivial tasks, such as knowing how to open a door or how to use an elevator. Nowadays, trivial commonsense perception skills, such as detecting a door and using its handle to move through the door or detecting the elevator buttons, are not default functionalities of mobile robots. Thus, robot recognition is not making sense of the whole scene. Similarly, the interpretation of scene data indicates specific capabilities. There are a few advanced approaches for interpreting a complex scene, such as KnowRob [65] (see the "Knowledge Interpretation" section). However, they provide narrow interpretations independent of the available scene information, whereas robots cannot instantly interpret an unseen scene. Therefore, the knowledge interpretation may not retrieve sufficient scene knowledge.

The Potentials of Robots

This section highlights the potentials in robot scene perception going beyond the perceptual capabilities of humans.

Flexible Sensor Design

The first presented gap in robot scene perception (using few or solely a single sensor modality) can be overcome by the usage of flexible sensor design. Human perception is limited to the recognized information of the senses and their range and accuracy, e.g., fog or darkness negatively influences scene recognition. Contrastively, robots benefit from scene-adjusted sensor modalities and their flexible configuration. They can overcome the limitations of human senses, such as through ultrasonic or radar recognition techniques [113]. Robots can be equipped with sensor modalities, such as radar, that enable them to freely adapt to their environment. The flexibility of the robot design allows the adaptation to the application. Thus, robots can use sensors with the desired amount, properties, and alignments. For instance, multiple visual sensors could enable robots to recognize the scene in 360°. Herewith, blind spots can be avoided, which is especially important for safe usage.

Our third identified gap is the lack of commonsense perception caused by specific recognition capabilities.

Initial Perception Capabilities

In contrast to the highlighted gap of missing perceptual learning, robots can be deployed with prelearned perception capabilities and with initial scene knowledge. This trivial fact allows reducing the setup time of the robot to a minimum. For instance, mirroring the perception skills of an existing robot in an environment enables a new robot to perceive the scene equivalently. In contrast to robots, humans would have to learn from scratch.

Cooperative Perception

Speaking the "same language" in terms of understanding and exchanging information enables the sharing of perceptual data independent of the robot, assistive system, or infrastructure. Therefore, artificial systems have fewer restrictions than humans as the number of collaborators and the communication range and bandwidth are not hard restricted. The sharing of scene information in real time with multiple artificial agents enables the fusion of scene observations from various perspectives. The data exchange enhances the scene knowledge of every agent, which offers a comprehensive scene perception from its scene overview. A popular research area presents cooperative approaches that share navigation data, e.g., to distribute the mapping [64] or for the reactive path planning of multiple robots [114]. Moreover, possible decentralized computation, e.g., cloud computing, can save the resources of mobile robots [115]. Therefore, cooperative perception is a key technique for reducing the setup time of large environments and for providing the safe navigation and better economy of larger robot fleets through its intercommunication [114], [116].

Future Directions

We propose two comparatively nonpopular future directions that contribute to deploying robots within our society. First, we propose developing frameworks to combine multiple perception modules. The scope of robotic scene perception is low since fundamental commonsense skills are missing. Commonsense skills can be generated with scene perception frameworks that combine several specific methods to achieve a holistic understanding. Second, although existing research on robot perception achieves good performance on specific tasks, the flexibility for adjusting the perception within the robot usage is missing. It must be possible to extend initially deployed perception skills and to adjust the perception at any time to the scene. Therefore, we address in particular our second and third identified gaps to future research that underlines the importance of perceptual learning being fundamental in everyday life. It is a central technique toward and beyond human-like perceptual performance.

We propose two comparatively nonpopular future directions that contribute to deploying robots within our society.

Conclusion

This article transfers human scene perception to mobile robots for comparison since the performance of human perception is superior in many tasks. Current research in robotics presents specific perception skills evaluated in a predefined and constricted manner, whereby they already achieve promising results for everyday applications. However, for lifelong unsupervised and autonomous usage, multifunctional mobile robots in particular need to perceive the scene holistically to reliably handle unknown and changing situations as well as uncertainties and dynamics. Human perception has been extensively studied since the 1970s, offering mature neuroscience studies and theories that define human perception as the process of recognizing, representing, and interpreting scene information. This article uses this threefold division for a transfer from human to robot scene perception to identify similarities and differences in the process. The transfer revealed that the robotic approaches partly mirror human-like perception.

However, much research investigates specific methods, such as object classification, that outperform human perception. This prerequisite toward and beyond human-like perceptual performance is promising. A new research area integrates multiple state-of-the-art methods in frameworks that aim toward a holistic scene understanding. However, these frameworks lack trivial commonsense perception skills and therefore cannot substantiate a holistic scene understanding. Moreover, only nonpopular research contributes to the perceptual learning of robots, which is needed to learn, adjust, and customize their perception. Therefore, these two major gaps need to be addressed by future research.

Acknowledgment

This research has received funding from the German Ministry of Economic Affairs, Labor and Tourism Baden-Württemberg by the AI Innovation Center Learning Systems and Cognitive Robotics under Grant Agreement 017–180036 (841022).

References

[1] C. Müller, B. Graf, and K. Pfeiffer, "World robotics 2021 – Service robots," International Federation of Robotics, Frankfurt, Germany, Tech. Rep., 2021. [Online]. Available: https://ifr.org/ifr-press-releases/news/service-robots-hit-double-digit-growth-worldwide

[2] A. Gasparetto and L. Scalera, "A brief history of industrial robotics in the 20th century," *Adv. Historical Stud.*, vol. 8, no. 1, pp. 24–35, 2019, doi: 10.4236/ahs.2019.81002.

[3] U. Neisser, "Without perception, there is no knowledge: Implications for artificial intelligence," in *Natural and Artificial Minds*, R. G. Burton, Ed. Albany, NY, USA: SUNY Press, 1993, pp. 174–164.

[4] M. Naseer, S. Khan, and F. Porikli, "Indoor scene understanding in 2.5/3d for autonomous agents: A survey," *IEEE Access*, vol. 7, pp. 1859–1887, 2019, doi: 10.1109/ACCESS.2018.2886133.

[5] G. L. Malcolm, I. I. A. Groen, and C. I. Baker, "Making sense of real-world scenes," *Trends Cogn. Sci.*, vol. 20, no. 11, pp. 843–856, 2016, doi: 10.1016/j.tics.2016.09.003.

[6] S. Garg et al., "Semantics for robotic mapping, perception and interaction: A survey," *Found. Trends® Robot.*, vol. 8, nos. 1–2, pp. 1–224, 2020, doi: 10.1561/2300000059.

[7] H. Levesque and G. Lakemeyer, "Cognitive robotics," in *Handbook of Knowledge Representation* (Foundations of Artificial Intelligence), vol. 3, F. v. Harmelen, V. Lifschitz, and B. Porter, Eds. New York, NY, USA: Elsevier, 2008, pp. 869–886.

[8] M. Potter, "Meaning in visual search," *Science*, vol. 187, no. 4180, pp. 965–966, 1975, doi: 10.1126/science.1145183.

[9] I. Biederman, "Perceiving real-world scenes," *Science*, vol. 177, no. 4043, pp. 77–80, 1972, doi: 10.1126/science.177.4043.77.

[10] R. A. Epstein and C. I. Baker, "Scene perception in the human brain," *Annu. Rev. Vis. Sci.*, vol. 5, no. 1, pp. 373–397, 2019, doi: 10.1146/annurev-vision-091718-014809.

[11] J. Beck, B. Hope, and A. Rosenfeld, *Human and Machine Vision*, 1st ed. New York, NY, USA: Springer-Verlag, 1983.

[12] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001, doi: 10.1023/A:1011139631724.

[13] C. M. Funke, J. Borowski, K. Stosio, W. Brendel, T. S. A. Wallis, and M. Bethge, "Five points to check when comparing visual perception in humans and machines," *J. Vis.*, vol. 21, no. 3, p. 16, 2021, doi: 10.1167/jov.21.3.16.

[14] L. San Roque et al., "Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies," *Cogn. Linguistics*, vol. 26, no. 1, pp. 31–60, 2015, doi: 10.1515/cog-2014-0089.

[15] M. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications*. Natick, MA, USA: A. K. Peters, 2015.

[16] L. Seminara, P. Gastaldo, S. Watt, K. Valyear, F. Zuher, and F. Mastrogiovanni, "Active haptic perception in robots: A review," *Frontiers Neurorobot.*, vol. 13, p. 53, Jul. 2019, doi: 10.3389/fnbot.2019.00053.

[17] S. Deutsch and A. Deutsch, *The Eye as a Transducer*. Piscataway, NJ, USA: IEEE Press, 1993, pp. 227–281.

[18] D. Purves, *Neuroscience*, 6th ed. London, U.K.: Oxford Univ. Press, 2018.

[19] P. Gärdenfors, *Conceptual Spaces - The Geometry of Thought*. Cambridge, MA, USA: MIT Press, 2000.

[20] P. Gärdenfors, *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. Cambridge, MA, USA: MIT Press, 2014.

[21] A. Dix, M. Pohl, and G. Ellis, *Perception and Cognitive Aspects*. Reims, France: Eurographics Association, 2010, ch. 7, pp. 109–130.

[22] J. Wolfe, N. Klempen, and K. Dahlen, "Postattentive vision," *J. Exp. Psychol. Hum. Perception Performance*, vol. 26, no. 2, pp. 693–716, 2000, doi: 10.1037/0096-1523.26.2.693.

[23] S. E. Palmer, "Hierarchical structure in perceptual representation," *Cogn. Psychol.*, vol. 9, no. 4, pp. 441–474, 1977, doi: 10.1016/0010-0285(77)90016-0.

[24] J. Wagemans and R. Kimchi, *The Perception of Hierarchical Structure*. London, U.K.: Oxford Univ. Press, 2014.

[25] S. Jin, H. Liu, B. Wang, and F. Sun, "Open-environment robotic acoustic perception for object recognition," *Frontiers Neurorobot.*, vol. 13, p. 96, Nov. 2019, doi: 10.3389/fnbot.2019.00096.

[26] C. Evers and P. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018, doi: 10.1109/TASLP.2018.2828321.

[27] M. Kreković, I. Dokmanić, and M. Vetterli, "EchoSLAM: Simultaneous localization and mapping with acoustic echoes," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2016, pp. 11–15, doi: 10.1109/ICASSP.2016.7471627.

[28] A. Balachandran, M. Brown, S. Erlien, and J. Gerdes, "Predictive haptic feedback for obstacle avoidance based on model predictive control," *IEEE Trans. Autom. Sci. Eng. (from July 2004)*, vol. 13, no. 1, pp. 26–31, Jan. 2016, doi: 10.1109/TASE.2015.2498924.

[29] C. Lytridis, G. Virk, and E. Kadar, "Co-operative smell-based navigation for mobile robots," in *Climbing and Walking Robots*. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 1107–1117.

[30] H. Shimazu, K. Kobayashi, A. Hashimoto, and T. Kameoka, "Tasting robot with an optical tongue: Real time examining and advice giving on food and drink," in *Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design*, M. J. Smith and G. Salvendy, Eds. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 950–957.

[31] B. Ciui et al., "Chemical sensing at the robot fingertips: Toward automated taste discrimination in food samples," *ACS Sensors*, vol. 3, no. 11, pp. 2375–2384, 2018, doi: 10.1021/acssensors.8b00778.

[32] G. Pintore, C. Mura, F. Ganovelli, L. Fuentes Perez, R. Pajarola, and E. Gobbetti, "State-of-the-art in automatic 3d reconstruction of structured indoor environments," *Comput. Graph. Forum*, vol. 39, no. 2, pp. 667–699, 2020, doi: 10.1111/cgf.14021.

[33] C. Premebida, R. Ambrus, and Z. Marton, "Intelligent robotic perception systems," in *Applications of Mobile Robots*, London, U.K.: IntechOpen, 2018, pp. 111–127.

[34] K. Klasing, D. Althoff, D. Wollherr, and M. Buss, "Comparison of surface normal estimation methods for range sensing applications," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 3206–3211, doi: 10.1109/ROBOT.2009.5152493.

[35] R. Bormann, J. Hampp, M. Hägele, and M. Vincze, "Fast and accurate normal estimation by efficient 3d edge detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2015, pp. 3930–3937, doi: 10.1109/IROS.2015.7353930.

[36] A. Nguyen and B. Le, "3d point cloud segmentation: A survey," in *Proc. IEEE Conf. Robot., Autom. Mechatronics (RAM)*, 2013, pp. 225–230, doi: 10.1109/RAM.2013.6758588.

[37] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022, doi: 10.1109/TPAMI.2021.3059968.

[38] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986, doi: 10.1109/TPAMI.1986.4767851.

[39] A. Kaiser, J. A. Y. Zepeda, and T. Boubekeur, "A survey of simple geometric primitives detection methods for captured 3D data," *Comput. Graph. Forum*, vol. 38, no. 1, pp. 167–196, 2019, doi: 10.1111/cgf.13451.

[40] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," in *Readings in Computer Vision: Issues, Problem, Principles, and Paradigms*, M. A. Fischler and O. Firschein, Eds. San Francisco, CA, USA: Morgan Kaufmann, 1987, pp. 671–679.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2012, vol. 1, pp. 1097–1105.

[42] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[43] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: A survey," *Pattern Recognit.*, vol. 51, pp. 148–175, Mar. 2016, doi: 10.1016/j.patcog.2015.08.027.

[44] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[45] S. Lowry et al., "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016, doi: 10.1109/TRO.2015.2496823.

[46] M. Labbé and F. Michaud, "RTAB-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *J. Field Robot.*, vol. 36, no. 2, pp. 416–446, 2018, doi: 10.1002/rob.21831.

[47] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An open-source slam system for monocular, stereo and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017, doi: 10.1109/TRO.2017.2705103.

[48] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robot. Auton. Syst.*, vol. 66, pp. 86–103, Apr. 2015, doi: 10.1016/j.robot.2014.12.006.

[49] F. Poux, "The smart point cloud: Structuring 3d intelligent point data," Ph.D. dissertation, Université de Liège, Liège, Belgium, 2019.

[50] R. Wang and J. Brockmole, "Human navigation in nested environments," *J. Exp. Psychol. Learn., Memory, Cogn.*, vol. 29, no. 3, pp. 398–404, 2003, doi: 10.1037/0278-7393.29.3.398.

[51] G. Welch and G. Bishop, "An introduction to the Kalman filter," Univ. of North Carolina at Chapel Hill, Chapel Hill, USA, 1995. [Online]. Available: https://perso.crans.org/club-krobot/doc/kalman.pdf

[52] F. Gustafsson et al., "Particle filters for positioning, navigation, and tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 425–437, Feb. 2002, doi: 10.1109/78.978396.

[53] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "A simple baseline for multi-object tracking," 2020. [Online]. Available: https://arxiv.org/abs/2004.01888

[54] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," 2017. [Online]. Available: http://arxiv.org/abs/1703.07402

[55] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu, "TransMot: Spatial-temporal graph transformer for multiple object tracking," 2021. [Online]. Available: https://arxiv.org/abs/2104.00194

[56] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[57] X. Li and Z. Zhou, "Object re-identification based on deep learning," in *Visual Object Tracking with Deep Neural Networks*. Rijeka: IntechOpen, 2019, ch. 5.

[58] V. Bansal, S. James, and A. Del Bue, "Re-OBJ: Jointly learning the foreground and background for object instance re-identification," in *Proc. Image Anal. Process. – ICIAP*, Berlin, Heidelberg: Springer-Verlag, 2019, pp. 402–413, doi: 10.1007/978-3-030-30645-8_37.

[59] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14,993–15,002, doi: 10.1109/ICCV48922.2021.01474.

[60] Y. Zhao, S. Zhu, D. Wang, and Z. Liang, "Short range correlation transformer for occluded person re-identification," 2022. [Online]. Available: https://arxiv.org/abs/2201.01090

[61] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017, doi: 10.1109/TIP.2017.2700762.

[62] B. Hommel, J. Gehrke, and L. Knuf, "Hierarchical coding in the perception and memory of spatial layouts," *Psychol. Res.*, vol. 64, no. 1, pp. 1–10, Oct. 2000, doi: 10.1007/s004260000032.

[63] L. Kunze et al., "SOMA: A framework for understanding change in everyday environments using semantic object maps," in *Proc. Conf. Artif. Intell. (AAAI-18)*, 2018, pp. 47–54.

[64] M. Labbé and F. Michaud, "Long-term online multi-session graph-based SPLAM with memory management," *Auton. Robots*, vol. 42, no. 6, pp. 1133–1150, 2018, doi: 10.1007/s10514-017-9682-5.

[65] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. Bozcuoglu, and G. Bartels, "Know rob 2.0 — A 2nd generation knowledge processing framework for cognition-enabled robotic agents," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 512–519, doi: 10.1109/ICRA.2018.8460964.

[66] L. Prieto González, V. Stantchev, and R. Colomo-Palacios, "Applications of ontologies in knowledge representation of human perception," *Int. J. Metadata, Semantics Ontologies*, vol. 9, no. 1, pp. 74–80, Feb. 2014, doi: 10.1504/IJMSO.2014.059128.

[67] J. Wyatt et al., "Self-understanding and self-extension: A systems and representational approach," *IEEE Trans. Auton. Mental Develop.*, vol. 2, no. 4, pp. 282–303, Dec. 2010, doi: 10.1109/TAMD.2010.2090149.

[68] R. Ravichandran, E. Prassler, N. Huebel, and S. Blumenthal, "A workbench for quantitative comparison of databases in multi-robot applications," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2018, pp. 3744–3750, doi: 10.1109/IROS.2018.8594241.

[69] A. Dietrich, S. Mohammad, S. Zug, and J. Kaiser, "ROS meets cassandra: Data management in smart environments with NoSQL," in *Proc. 11th Int. Baltic Conf. DB IS*, 2014, pp. 1–12.

[70] M. Tenorth, U. Klank, D. Pangercic, and M. Beetz, "Web-enabled robots – robots that use the web as an information resource," *IEEE Robot. Autom. Mag.*, vol. 18, no. 2, pp. 56–68, 2011.

[71] P. Philipp, M. Maleshkova, A. Rettinger, and D. Katic, "A semantic framework for sequential decision making," in *Engineering the Web in the Big Data Era*, P. Cimiano, F. Frasincar, G. J. Houben, and D. Schwabe, Eds. Cham, Switzerland: Springer International Publishing, 2015, pp. 392–409.

[72] C. Follini, M. Terzer, C. Marcher, A. Giusti, and D. T. Matt, "Combining the robot operating system with building information modeling for robotic applications in construction logistics," in *Advances in Service and Industrial Robotics*, S. Zeghloul, M. Laribi, and J. Sandoval Arevalo, Eds. Cham, Switzerland: Springer International Publishing, 2020, pp. 245–253.

[73] L. Isik, A. Tacchetti, and T. Poggio, "A fast, invariant representation for human action in the visual system," *J. Neurophysiol.*, vol. 119, no. 2, pp. 631–640, 2018, doi: 10.1152/jn.00642.2017.

[74] Y. Peng, "Causal action: A framework to connect action perception and understanding," Ph.D. dissertation, University of California, Los Angeles, Los Angeles, CA, USA, 2019.

[75] J. Henderson and A. Hollingworth, "High-level scene perception," *Annu. Rev. Psychol.*, vol. 50, no. 1, pp. 243–271, 1999, doi: 10.1146/annurev.psych.50.1.243.

[76] M. H. Segall, D. T. Campbell, and M. J. Herskovit, *The Influence of Culture on Visual Perception*. Indianapolis, IN, USA: Bobbs-Merrill Company, 1966.

[77] S. Torresin, G. Pernigotto, F. Cappelletti, and A. Gasparella, "Combined effects of environmental factors on human perception and objective performance: A review of experimental laboratory works," *Indoor Air*, vol. 28, no. 4, pp. 525–538, 2018, doi: 10.1111/ina.12457.

[78] P. Brugger and S. Brugger, "The easter bunny in October: Is it disguised as a duck?" *Perceptual Motor Skills*, vol. 76, no. 2, pp. 577–578, 1993, doi: 10.2466/pms.1993.76.2.577.

[79] M. Ozturk, M. Ersen, M. Kapotoglu, C. Koc, S. Sariel, and H. Yalçın, "Scene interpretation for self-aware cognitive robots," in *Proc. AAAI Spring Symp.*, 2014, pp. 1–8.

[80] Z. Jiang, B. Liu, S. Schulter, Z. Wang, and M. Chandraker, "Peek-a-Boo: Occlusion reasoning in indoor scenes with plane representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 110–118, doi: 10.1109/CVPR42600.2020.00019.

[81] A. Persson, P. Z. Dos Martires, L. D. Raedt, and A. Loutfi, "Semantic relational object tracking," *IEEE Trans. Cogn. Devel. Syst.*, vol. 12, no. 1, pp. 84–97, Mar. 2020, doi: 10.1109/TCDS.2019.2915763.

[82] X. Zhang, Z. Wu, and Y. G. Jiang, "SAM: Modeling scene, object and action with semantics attention modules for video recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 313–322, 2022, doi: 10.1109/TMM.2021.3050058.

[83] L. Piyathilaka and S. Kodagoda, "Human activity recognition for domestic robots," in *Field and Service Robotics*, L. Mejias, P. Corke, and J. Roberts, Eds. Cham, Switzerland: Springer-Verlag, 2015, pp. 395–408.

[84] P. Duckworth, M. Alomari, Y. Gatsoulis, D. C. Hogg, and A. G. Cohn, "Unsupervised activity recognition using latent semantic analysis on a mobile robot," in *Proc. IOS Press*, 2016, pp. 1062–1070.

[85] I. Kostavelis et al., "Understanding of human behavior with a robotic agent through daily activity analysis," *Int. J. Social Robot.*, vol. 11, no. 3, pp. 437–462, 2019, doi: 10.1007/s12369-019-00513-2.

[86] K. Li, J. Wu, X. Zhao, and M. Tan, "Real-time human-robot interaction for a service robot based on 3D human activity recognition and human-mimicking decision mechanism," in *Proc. IEEE Int. Conf. Cyber Technol. Autom., Control, Intell. Syst.*, 2018, pp. 498–503, doi: 10.1109/CYBER.2018.8688272.

[87] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, 1999, pp. 850–855.

[88] F. Rezazadegan, S. Shirazi, B. Upcrofit, and M. Milford, "Action recognition: From static datasets to moving robots," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 3185–3191, doi: 10.1109/ICRA.2017.7989361.

[89] P. Philipp, M. Bommersheim, S. Robert, and J. Beyerer, "Probabilistic estimation of human interaction needs in context of a robotic assistance in geriatrics," *Current Directions Biomed. Eng.*, vol. 5, no. 1, pp. 433–435, 2019, doi: 10.1515/cdbme-2019-0109.

[90] D. Beßler, R. Porzel, P. Mihai, M. Beetz, R. Malaka, and J. Bateman, "A formal model of affordances for flexible robotic task execution," in *Proc. 24th Eur. Conf. Artif. Intell. (ECAI)*, 2020, pp. 2425–2432, doi: 10.3233/FAIA200374.

[91] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, 2014, vol. 27, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[92] L. Alzubaidi et al., "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[93] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophysics*, vol. 5, no. 4, pp. 115–133, 1943, doi: 10.1007/BF02478259.

[94] D. Purves, G. Augustine, D. Fitzpatrick, W. Hall, A. Lamantia, and L. White, *Neuroscience,* 5th ed. Sunderland: Sinauer, 2011.

[95] L. Jiao and J. Zhao, "A survey on the new generation of deep learning in image processing," *IEEE Access*, vol. 7, pp. 172,231–172,263, Nov. 2019, doi: 10.1109/ACCESS.2019.2956508.

[96] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*.

[97] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2009, doi: 10.1007/s11263-009-0275-4.

[98] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[99] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Computer Vision – ECCV*, Cham, Switzerland: Springer International Publishing, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[100] J. Mccormac, R. Clark, M. Bloesch, A. Davison, and S. Leutenegger, "Fusion++: Volumetric object-level SLAM," in *Proc. 2018 Int. Conf. 3D Vis. (3DV)*, pp. 32–41, doi: 10.1109/3DV.2018.00015.

[101] Y. Lin, J. Tremblay, S. Tyree, P. Vela, and S. Birchfield, "Multi-view fusion for multi-level robotic scene understanding," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2021, pp. 6817–6824, doi: 10.1109/IROS51168.2021.9635994.

[102] A. Koubaa, *Robot Operating System (ROS): The Complete Reference*, vol. 1, 1st ed. Cham, Switzerland: Springer Publishing Company, 2016.

[103] J. Suchan and M. Bhatt, "Commonsense scene semantics for cognitive robotics: Towards grounding embodied visuo-locomotive interactions," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW),* 2017, pp. 742–750, doi: 10.1109/ICCVW.2017.93.

[104] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison, "SLAM++: Simultaneous localisation and mapping at the level of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1352–1359, doi: 10.1109/CVPR.2013.178.

[105] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: An open-source library for real-time metric-semantic localization and mapping," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2020, pp. 1689–1696, doi: 10.1109/ICRA40945.2020.9196885.

[106] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, "3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans," 2020, *arXiv:2002.06289*.

[107] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st ed. Berlin, Heidelberg: Springer-Verlag, 2010.

[108] E. J. Gibson, "Perceptual learning," *Annu. Rev. Psychol.*, vol. 14, no. 1, pp. 29–56, 1963, doi: 10.1146/annurev.ps.14.020163.000333.

[109] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," in *Advances in Data Science and Information Engineering*, R. Stahlbock, G. M. Weiss, M. Abou-Nasr, C. Y. Yang, H. R. Arabnia, and L. Deligiannidis, Eds. Cham, Switzerland: Springer International Publishing, 2021, pp. 877–894.

[110] H. Qin and D. Zhang, "A perpetual learning algorithm that incrementally improves performance with deliberation," *IEEE Access*, vol. 8, pp. 131,425–131,438, Jul. 2020, doi: 10.1109/ACCESS.2020.3009718.

[111] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robot. Auton. Syst.*, vol. 15, nos. 1–2, pp. 25–46, 1995, doi: 10.1016/0921-8890(95)00004-Y.

[112] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009, doi: 10.1016/j.robot.2008.10.024.

[113] C.-C. Carbon, "Understanding human perception by human-made illusions," *Front. Hum. Neurosci.*, vol. 8, p. 566, Jul. 2014, doi: 10.3389/fnhum.2014.00566.

[114] S. Dörr, *Cloud-based Cooperative Long-term SLAM for Mobile Robots in Industrial Applications* (ser. Stuttgarter Beiträge zur Produktionsforschung). Stuttgart, Germany: Fraunhofer Verlag, 2020.

[115] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, "A survey of research on cloud robotics and automation," *IEEE Trans. Autom. Sci. Eng. (from July 2004)*, vol. 12, no. 2, pp. 398–409, Apr. 2015, doi: 10.1109/TASE.2014.2376492.

[116] A. Miller, K. Rim, P. Chopra, P. Kelkar, and M. Likhachev, "Cooperative perception and localization for cooperative driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2020, pp. 1256–1262, doi: 10.1109/ICRA40945.2020.9197463.

Florenz Graf, Department of Robot and Assistive Systems, Fraunhofer IPA, Stuttgart 70569, Germany. E-mail: florenz.graf@ipa.fraunhofer.de.

Jochen Lindermayr, Department of Robot and Assistive Systems, Fraunhofer IPA, Stuttgart 70569, Germany. E-mail: jochen.lindermayr@ipa.fraunhofer.de.

Çağatay Odabaşi, Department of Robot and Assistive Systems, Fraunhofer IPA, Stuttgart 70569, Germany. E-mail: cagatay.odabasi@ipa.fraunhofer.de.

Marco F. Huber, Center for Cyber Cognitive Intelligence, Fraunhofer IPA, and Institute of Industrial Manufacturing and Management IFF, University of Stuttgart, Stuttgart 70569, Germany. E-mail: marco.huber@ieee.org.