

State-of-the-Art Elderly Service Robot

Environmental Perception, Compliance Control, Intention Recognition, and Research Challenges

by Xiaofeng Liu^{ID},
Congyu Huang^{ID},
Haoran Zhu,
Ziyang Wang^{ID},
Jie Li^{ID}, and
Angelo Cangelosi^{ID}

This article surveys the literature on environmental perception, compliance control, and intention recognition for elderly service robots. Population aging is an inevitable trend in current society, leading to an urgent need for service robots. A high-performance collaborative robot should be able to construct an environment of harmonious and effective human-robot symbiosis. This survey starts with an overview of the environment perception system (e.g., environment 3D reconstruction and object location) for elderly service robots, as environment perception is the basis of human-robot interaction. Furthermore, we report various learning methods deployed in compliance control, which opens the way for



©SHUTTERSTOCK.COM/MIRIAM DOERR MARTIN FROMMERZ

the generalization of robots. In this case, we structure our review into three primary domains delimited by learning processes. The article also presents several aspects focusing on processing multimodal signals (e.g., vision, speech, and electromyograms) related to intention recognition. Finally, we briefly discuss several challenges that robots need to overcome to reach interactive performance comparable to humans.

Digital Object Identifier 10.1109/MSMC.2023.3238855
Date of current version: 10 January 2024

Introduction

Human–robot collaboration has always been the hot spot in the robot community. The Industry 4.0 project has broadened the horizon of the robot community [1], [2], where robots can flexibly interact with the surrounding environment. In the past, most robots were restricted to a narrow working region and dealt only with simple repetitive tasks, which dramatically limited their utility. Human–robot collaboration promotes the idea that robots can become a part of human life (i.e., in hospitals, restaurants, and homes) and assist the elderly with various tasks like a partner. With effective human–robot collaboration, human needs in work and life can be satisfied by robots, freeing humans from repetitive dangerous tasks and allowing them to devote energy to more sophisticated tasks. The outbreak of COVID-19 has exacerbated this demand [3], [4], such as for nucleic acid testing and physical distribution. Furthermore, population aging has become an inevitable trend worldwide, leading to an urgent need for service robots. However, current collaborative robots are far from reaching this level and cannot work effectively in our homes and offices.

The definition of human–robot interaction is multifaceted, ranging from physical collaboration [9], [10] to cognitive aspects [11], [12]. For elderly service robots, physical collaboration focuses on equipping robots with the ability to meet various needs of the elderly in real life. In comparison, cognitive aspects pay more attention to elements influencing the relationship between robots and the elderly, such as intention recognition, emotional communication, and human–robot trust. To adapt to the dynamic surroundings and choose a suitable policy for the particular task, the service robot should be able to analyze the multimodal data coming from the environment [13], develop strategies, and plan motion, which is critical during human–robot interaction. Fortunately, the rapid development of electronic hardware technology [9] and artificial intelligence (AI) provides a strong guarantee for realizing many prospects in robotics.

Due to the enormous demand in the elderly service industry, most work in human–robot collaboration has focused on elderly service robots [15] (see Figure 1). In terms of the nursing robot, Toshiharu et al. first proposed RI-MAN [5] for elder care, in 2006. Salichs et al. [6] presented a new social robot, Mini, specifically designed to assist and accompany the elderly at home and in a nursing facility. In [8], Hiroyasu et al. proposed Twendy-One, which is a type of human-symbiotic robot that escorts elderly individuals and performs tasks with dexterity while securing safe physical contact, such as attendant care for 24-h and kitchen support. However, the existing robot technology is still immature, single functional, and expensive, which significantly limits the application of escort robots [16], [17].

This work aims to review the critical techniques of elderly service robots from the broader field of robotics

and provide a starting point for researchers who have just moved into this domain. We begin this article by briefly surveying the work on environmental perception systems, in the “Perception System” section. The “Compliance Control” section reveals compliance control to endow robots with the ability to learn new skills. Then, the “Intention Recognition” section focuses on intention recognition in human–robot interaction, analyzing aspects such as language instruction and behavior perception. We respectively discuss the challenges of elderly service robots in human–robot collaboration and conclude this article in the “Challenge” and “Conclusion” sections.

Perception System

Considering the unpredictable and unstructured accompanying environment, the perception system of service robots discussed herein focuses on environment 3D reconstruction and object perception, as environment perception is a prerequisite for achieving effective human–robot interaction. Its main idea is to acquire information about the surrounding environment, extract the significant features from noise, and, finally, understand the environment through robot sensors. Supported by the accurate perception of the surrounding environment, elderly service robots can better assist humans by cleaning the house, cooking, and feeding.

3D Reconstruction of the Environment

Service robots should complete the 3D environment reconstruction based on the elderly’s demands and environmental information. For example, when asked to go to the kitchen to get a cup, robots should know where to go and how to get there. Typically, during the planning process, the robot should constantly perceive the surrounding environment, adjust its position in real time, and calculate the best policy to reach the target.

In robotic navigation, 3D reconstruction is a classic problem related to computer vision, which is mainly divided into two domains: multiview reconstruction and motion reconstruction. The former includes the famous multiview stereo vision (MVS) algorithm, and the latter includes the simultaneous localization and mapping (SLAM) algorithm, which was first introduced by Smith, Self, and Cheeseman, in 1986, and has been expanded for more than 30 years. 3D reconstruction methods based on SLAM can be further divided into filter-based and graph optimization-based methods.



Figure 1. Elderly service robots: (a) RI-MAN [5], (b) Mini [6], (c) Pepper [7], and (d) Twendy-One [8].

The MVS algorithm proposed using two or more cameras to capture images from different viewpoints and obtain depth information to reconstruct 3D models, as shown in Figure 2. Existing research develops the multiview reconstruction method by fusing the data from multiple sensors [18], [19], which significantly improves the reliability and accuracy of environment perception. Based on stereo vision, Wang et al. [20] proposed a 6D iterative closest-point method for 3D reconstruction and state estimation of unknown and noncooperative targets. The 3D reconstruction is accomplished by generating 3D target points through optical flow tracking of Harris image points between consecutive frames. To efficiently explore the correlation between different viewpoints, Wang et al. [21] performed feature extraction and fusion in a complete architecture to redefine the multiview 3D reconstruction task as a sequence-to-sequence prediction problem. This approach allows the correlation among multiple unordered input perspectives to be fully exploited. Due to hardware constraints, one robot cannot get 360° information on the object in real time. However, the multiperspective method does not seem to be the first choice for 3D environment reconstruction in the elderly care context. Although it is possible to compensate for the information loss by setting up multiple cameras on the ceiling, the computation complexity and refresh rate are big problems.

In motion reconstruction, the SLAM algorithm plays an important role. Initially, researchers transformed the 3D reconstruction into a state estimation problem. They solved it with methods such as extended Kalman filtering, particle filtering, and maximum likelihood estimation. The authors of [23] attempted to use the joint state vectors to represent the robot's position and feature maps, which are estimated by extending the Kalman filter. However, the method is not suitable for nonlinear systems. With the unscented transform, Julier et al. [24] applied the unscented Kalman filter to improve the extended Kalman

filter algorithm. However, these methods suffer from sensitive data correlation and great computational effort, which is not widely applicable. In [25], the particle filtering approach was introduced, which avoids the computational complexity of Kalman filtering and error sensitivity of data correlation by using particle distribution to represent the robot's position and waypoints. However, when the number of sampling particles is insufficient, this method tends to perform poorly and incurs high computational cost.

With the advancement of research, SLAM based on graph optimization has become the mainstream approach for visual SLAM. Essentially, the complete motion trajectory of the robot is estimated with the smoothing principle, which utilizes the observed information to construct a map of the environment. Klein et al. [26] presented a parallel tracking and mapping system, which implemented an online real-time SLAM by graph optimization for the first time. In another study [27], the oriented fast and rotated brief-SLAM system was proposed. It has good generalizability and works well in unknown environments. The single SLAM method may not be robust enough to provide accurate poses under arbitrary conditions. Shao et al. [28] designed a scheme using VisualSLAM to compensate for lidarSLAM that combined vision and lidar to achieve better localization performance. Yeh et al. [29] developed a red-green-blue plus depth data (RGB-D) SLAM system that used high-dynamic-range image information to improve the geometric accuracy of reconstruction. The algorithm has strong robustness and good performance in low-light environments. In [30], several criteria for analyzing the accuracy of 3D reconstruction were proposed to improve the performance.

Obviously, the SLAM-based 3D reconstruction method is more suitable for elderly service robots, considering the environment and hardware factors. In particular, the great success of recent autonomous mobile robots has proved that SLAM-based approaches can effectively complete various indoor and outdoor 3D reconstruction and navigation tasks. Furthermore, service robots should maintain the balance between task execution and the satisfaction of humans in the care environment. A good example is that robots may block the way of the elderly while driving. Therefore, how to design a reasonable priority level for tasks is still a topic worth exploring.

Object Perception

When performing manipulation operations, elderly service robots need to acquire sufficient semantics information about the target object, such as its category, position, and direction. Unlike robots, humans can quickly get these factors through the brain and eyes. Robots understand nothing if the image from the camera is not further processed. In addition, due to noise interference and hardware defects, it is difficult for sensors to provide accurate and complete information in a real environment.

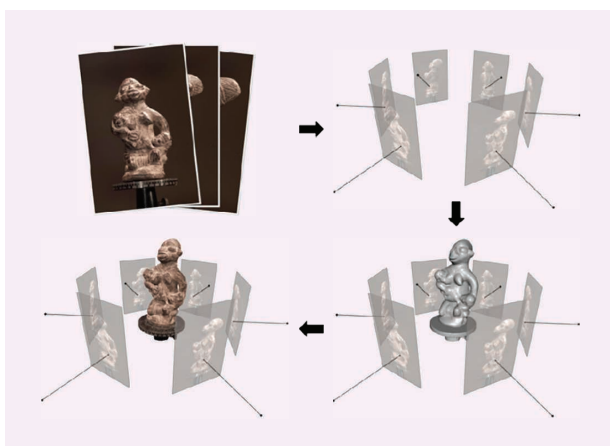


Figure 2. The multiperspective 3D reconstruction [22]. Since it is impossible to obtain 360° information on objects from a single viewpoint, whether 2D images or 3D point clouds, visual data from different perspectives will be complemented and integrated.

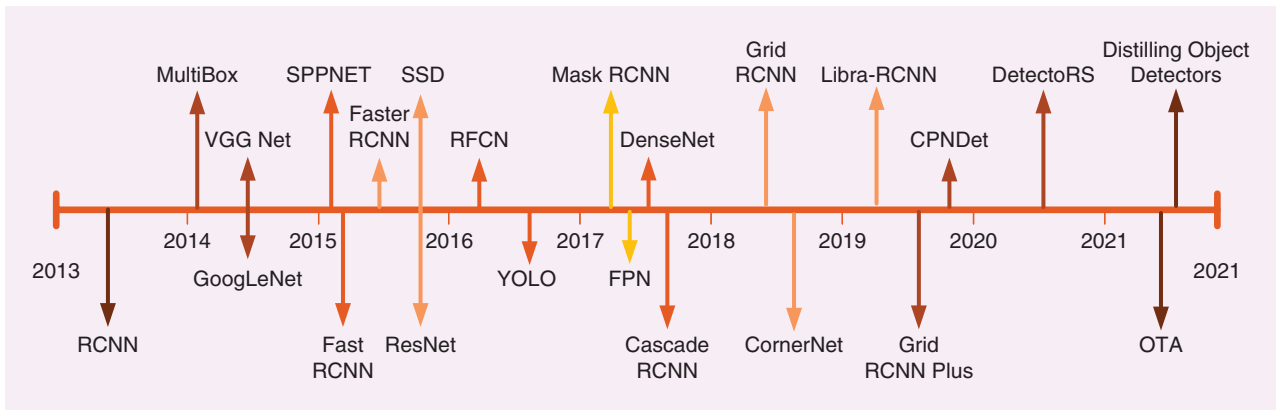


Figure 3. The notable techniques of target detection and instance segmentation in the past decade. RCNN: region-based convolutional neural network; RFCN: region-based fully convolutional network; YOLO: You Only Look Once; FPN: Feature Pyramid Network; CPNDet: Corner Proposal Network for Anchor-Free Two-Stage Object Detection; DetectoRS: DetectoRS: Detecting Objects With Recursive Feature Pyramid and Switchable Atrous Convolution; OTA: optimal target assignment.

Object perception can be broadly classified into three categories depending on the needs of the applicable occasions: localization without recognition, target detection with instance segmentation, and object recognition based on the point cloud.

Given the description, how to locate the proper object from images has always been the most challenging problem in computer vision. The evolution of target detection and instance segmentation is presented in Figure 3. Ahn et al. [43] proposed a method using You Only Look Once (YOLO)-based object recognition for cognitive robots. Subsequently, YOLO algorithms have been continuously optimized and have now evolved to YOLO v8. For example, in [44], the authors showed a YOLO v3-based object recognition system. It allowed robots to perceive surrounding objects as humans do, which reduced network latency and bandwidth requirements and ensured real-time applications. In [45], Babila et al. used YOLO v5 to count inventory. In addition, Kim et al. [46] demonstrated the lightweight and ground-optimized lidar odometry and mapping system, which solved the problem of identifying and localizing multiple objects simultaneously when using 3D SLAM to build maps in indoor environments. However, this method is time-consuming and cannot provide practical applications in the real world.

Service robots should have basic concepts to distinguish different things. Among them, one of the most important things is to detect and segment the human body effectively. Instance segmentation is a combination of target detection and semantic segmentation. Relative to the bounding box brought by target detection, instance segmentation can provide the detailed edge of the object. In contrast to semantic segmentation, instance segmentation requires labeling every individual among the objects on the graph. There are four main methods: the mask-suggested classification method, detection before segmentation method, tagged pixel followed by the

clustering method, and dense sliding window method. The main techniques included in the four ways are listed in Table 1 [47].

Girshick et al. [31] pioneered the region-based convolutional neural network (RCNN) algorithm, which applied a large-capacity CNN as a bottom-up regional proposal to localize and segment objects. Compared to other traditional models, the RCNN algorithm can greatly improve performance. Based on deep CNNs, the fast RCNN [32] was proposed to perform target detection. Subsequently, optimization methods, such as the faster RCNN [33], mask RCNN [34], and mask scoring RCNN [35], have been presented to reduce the running time and improve the accuracy of computational detection. In addition, there are some other methods for object localization and perception. Li et al. [48] proposed the Online Algorithm for Scalable Image Similarity, a single-stage one-shot shape-based instance segmentation algorithm, which can predict the modal mask under the presence of rotation, scale difference, and occlusion. Tomono et al. [49] constructed 2D maps with laser range finder scanning

Table 1. The instance segmentation method.

Group	Technique
Classification of mask proposals	RCNN [31], fast RCNN [32], and faster RCNN [33]
Detection followed by segmentation	Mask RCNN [34], mask scoring RCNN [35], MPN [36], and YOLACT [37]
Labeling pixels followed by clustering	Deep watershed transform [38] and instance cut [39]
Dense sliding window methods	Deep mask [40], instance FCN [41], and tensor mask [42]

MPN: Markovian Policy Network; YOLACT: You Only Look at Coefficients.

and used vision-based target recognition to verify the detected candidate targets. In [50], Park et al. provided specific information about the object through radio-frequency (RF) ID tags. The robot can identify the object based on the detected tag and the downloaded visual descriptor. The disadvantage is that all target objects should be labeled with RFID tags. In another work [51], Yang et al. fused the object's color, texture, and shape features. They proposed a new method for indoor robot perception by combining various features, which improves the matching accuracy compared to single-feature recognition and meets real-time requirements.

Although the preceding algorithms show excellent performance when facing some particular tasks, there is still a long way to go to meet the elderly service robots' requirements. On the one hand, most algorithms are based on 2D images. However, 2D images are often influenced by lighting, occlusion, and other factors, which will significantly cause information loss. Furthermore, 2D images cannot provide spatial information for elderly service robots when performing manipulation tasks. On the other hand, the performance of visual algorithms based on 3D information is sometimes unstable, especially when encountering transparent and textureless objects. Given the demand for elderly service, it is reasonable to combine techniques from 2D and 3D effectively.

Compliance Control

How to accord robots the ability to work across tasks and environments, making the human-robot interaction process smooth and natural, is a significant problem in the robot community. The accompanying scenes are diverse, which is a great challenge to the adaptability and adjustment ability of the service robot. In addition, most robots still rely too much

on the predesigned functions of programmers when executing tasks, which makes the mechanical control too rigid to achieve compliance control. In this case, we focus on the domain of robot skill learning and survey the theme from several aspects, such as method, content, and quality.

Human-Robot Teaching

Limited by technology and other reasons, designers cannot imagine all situations and endow robots with corresponding tips and skills, which significantly restricts the application of collaborative robots in the real world. Presently, the robot's performance is still very rough, and robots often fail to complete manipulation tasks as simple as grasping. To deal with this problem, the theory of learning from demonstration (LfD), which is also called *imitation learning (IL)*, provides a paradigm for robot skill learning [64]. Its core idea is to simplify the functional deployment of robots, where robots can quickly master new skills through simple human demonstration and even carry out self-learning by drawing inferences about other cases from one instance.

A traditional learning method of LfD is human-robot teaching (see Figure 4), in which people directly control the robot's movements to complete some basic actions [52], [53], [65], [66], such as dragging the robot arm to execute a particular trajectory and make correct postures. In this case, each joint of the robot and various sensors will record the relevant data during the movement process and then successfully reproduce the motion when performing the same task next time. Similarly, there are methods of shadowing [67], [68], [69] and teleportation [70], [71], [72]. A common property of these methods is that they allow robots to learn from human experts directly by mapping the robot's action to human behavior [73], as demonstrated in Figure 5. In [53], Sylvain et al. proposed an approach

to gradually teach human gestures to a humanoid robot through the motion sensors worn by a human to record the human's upper-body movement. It also helps the robot reproduce the movement through kinesthetic teaching, which shows that the significant features of a gesture can be efficiently transferred through interaction with the robot. Akgun et al. [52] interacted with PR2 by applying kinesthetic demonstrations and allowing the robot to learn low-level and high-level skills. Wei et al. [54] developed a robot skill learning system leveraging a teleoperation system, which directly transfers the human experience to the robot. This way of obtaining direct experience from a human expert is convenient, enabling robots to learn skills quickly. However, robots can perform only some relatively simple motion and manipulation tasks, which is far from meeting the needs of elderly care. In addition, when the

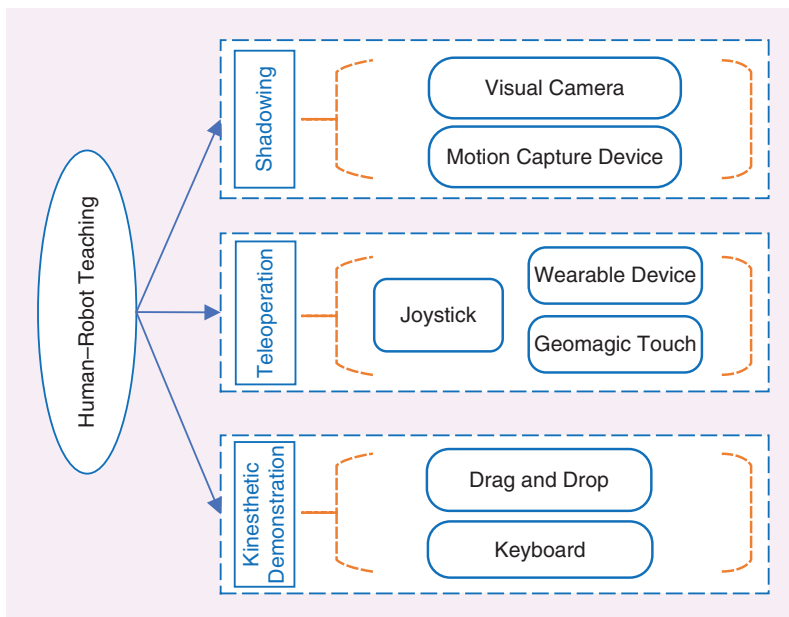


Figure 4. The methods of human-robot teaching.

degrees of freedom of the robot increase, it is almost impossible to construct the motion mapping between humans and robots.

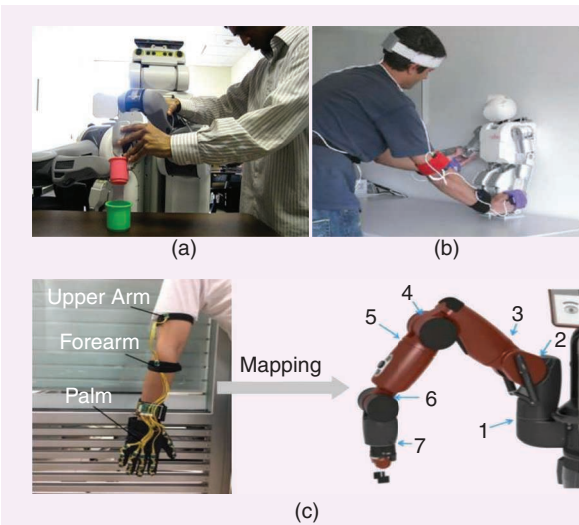


Figure 5. (a) and (b) A kinesthetic demonstration [52], [53]. (c) Teleoperation [54].

Concept Acquisition

Humans are so intelligent that they have invented language and characters and mastered various skills through learning in the long process of evolution [74]. Humans constantly accumulate knowledge and life experiences and acquire concepts by abstracting and generalizing common characteristics from a handful of things. Through concepts, humans are able to reason and think when facing a novel situation. For example, human beings have the concept of shape, so even if they see an object they have never seen before, they can outline it in their mind at first glance.

Current robots just lack the ability to acquire concepts and learn skills. In this case, a large body of work in cognitive robotics has investigated how robots can obtain concepts concerning object categories and attributes (i.e., color, shape, and pose) to complete some complex operations, as detailed in Table 2. In [57], an online robot algorithm is presented to gain knowledge of natural language and learn object concepts by training a language model that uses unsupervised word segmentation and multimodal information from the version. Yang et al. [62] proposed an approach for human-to-robot handovers by learning the

Table 2. An analysis of different work regarding concept acquisition.				
Year	Authors	Human-robot interaction task	Concept learned	Highlights
2014	McColl and Nejat [55]	Perceiving and identifying human social behaviors and responding in a corresponding way	Knowledge of emotional body language	Perception and expression of emotion through body language by a humanlike social robot
2017	Cai et al. [56]	Performing different grasp tasks based on the recognized color	Color concept corresponding to behavioral strategy (e.g., traffic lights)	Color recognition and dynamic decision-making model of sixth axis of industrial robot
2017	Nishihara et al. [57]	Object recognition and classification	Language model that links to object concept	Acquiring the meaning of a word by connecting word sequences to corresponding object concept
2017	Schmidt et al. [58]	6D object pose estimation	Spatial and orientation information of objects	Estimating the 6D pose of objects in cluttered scenes using vision data only
2018	Anderson et al. [59]	Interpreting visually grounded navigation instructions	Spatial relations and orientation perception	Linking natural language to spatial information from the environment
2018	Paul et al. [60]	Understanding natural language instructions based on environment representation	Abstract spatial concepts for natural language interaction	Modeling abstract spatial concepts of objects (e.g., rows, columns, and groups) and interpreting the instruction
2020	Wei et al. [61]	3D shape recognition from multiview features of 2D images	Geometry of the object	Combining multiview features to describe the global 3D shape by view-based graph CNN
2020	Yang et al. [62]	Object transfer between humans and robots	Handover methods based on human grasp classification	Receiving objects in a corresponding manner according to the hand grasp type
2021	Shao et al. [63]	Almost 78 different manipulation tasks (e.g., pull, remove, and plug)	Manipulation concepts that connect language instructions to motor skills	Learning to imitate human from human video demonstration

concept of the human's grasp type, where the robot can accordingly plan a trajectory based on the concept and take the object from the human's hand. Shao et al. [63] drew on the experience of LfD and endowed a robot with the ability to learn manipulation concepts that link natural language instructions to motor skills by leveraging large-scale video datasets of human's manipulation actions.

Undoubtedly, the establishment of the concept will bring significant improvement to the service robot. For example, service robots should construct the health concept for the elderly and generate coping strategies according to a personal situation. The success of deep learning for medical diagnosis and analysis [75] has shown that it is possible for robots, just like human doctors, to diagnose patients based on experience and case libraries. The challenge is how to combine existing AI technology with robotics, which will be further discussed in detail.

Knowledge Transfer and Autonomous Learning

The method of human–robot teaching does not allow robots to learn the advanced skills required to deal with complex tasks, which is unrealistic for robots to achieve absolute compliance control. To construct effective human–robot collaboration, elderly service robots need knowledge transfer and autonomous learning abilities. Research in robotics has focused on combining various types of AI technology to realize some manipulation tasks (i.e., grasping, handing over, and opening the door). The machine learning trend promotes more LfD learning schemes [i.e., reinforcement learning (RL)] so that the robot can learn more complex skills.

RL emphasizes that an agent constantly optimizes the policy according to the feedback of the reward and environmental state to approach the best effect, which coincides with the purpose of autonomous robot learning. RL is a superior technology that has yielded great fruit in the field of robot skill learning [76], [77], [78]. In [76], Chelsea et al. utilized a deep spatial autoencoder to obtain feature points describing the environment state. Then, the robot learned motor skills through RL according to these feature points. Schölkopf et al. [78] introduced a framework using RL for mobile robots. Experimental results show that compared with moderately experienced

programmers, the framework can learn a reasonable control policy faster.

In addition to RL, other aspects of machine learning are also used in robot skill learning [63], [80]. With the help of various open source large-scale datasets, robots can achieve knowledge transfer and quickly master different skills. Mohit et al. combined the strengths of Contrastive Language–Image Pretraining (CLIP) [81] and Transporter [82] to present CLIPort [80], which solved various language-specified tabletop tasks ranging from packing unseen objects to folding clothes. In this case, the robots learn skills from data.

The rapid development of electronic hardware technology makes physical simulation more efficient and convenient. The existing high-performance GPU can perform parallel operations on thousands of simulation algorithms in a few minutes through hardware acceleration. To train the AI for a particular task, years of a single robot's experience can simulate in a little over a day. Currently, researchers in robotics prefer to transfer the optimal data to the motion control of real robots after tens of thousands of training sessions in the simulation environment. In [79], Nikita et al. trained thousands of quadruped robots called ANYmal to walk on uneven ground in parallel on a single GPU, costing fewer than 20 min, as described in Figure 6. Compared to previous results, this method accelerates learning by multiple orders of magnitude.

Inspired by these works, we can train service robots based on the elderly care context and personal preferences through simulation. Therefore, it avoids more time-consuming processing, such as environmental layout and hardware construction. It also reduces the injury and loss caused by accidents. Although it is impossible to completely simulate the physiological and psychological reactions of the elderly in a simulation environment, we believe that the simulation data still play an essential role for elderly service robots.

Intention Recognition

Sensing Mode

To achieve better human–robot interaction, service robots should analyze multimodal information from the elderly to realize intention recognition. In general, intention analysis involves three steps (see Figure 7): sensor information collecting, signal processing, and intention recognition. Specifically, multimodal information comes from different local features of the elderly (i.e., bioelectricity signals, visual signals, and auditory signals). Different sensing methods can complement one another and improve the performance of robots effectively, especially in complex environments with multiagents. For example, for the elderly who have difficulty in action and speaking (e.g., the man with Parkinson's disease), the bioelectrical signal is an effective sensing mode that can directly reflect the elderly's state and demand. More details about the strengths and weaknesses of different sensing signals are summarized in Table 3.

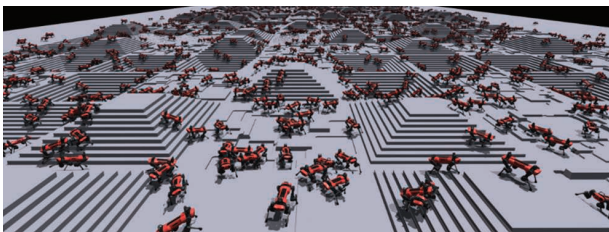


Figure 6. An example of massively parallel deep RL [79], where thousands of robots will simulate in parallel, much faster than in real time.

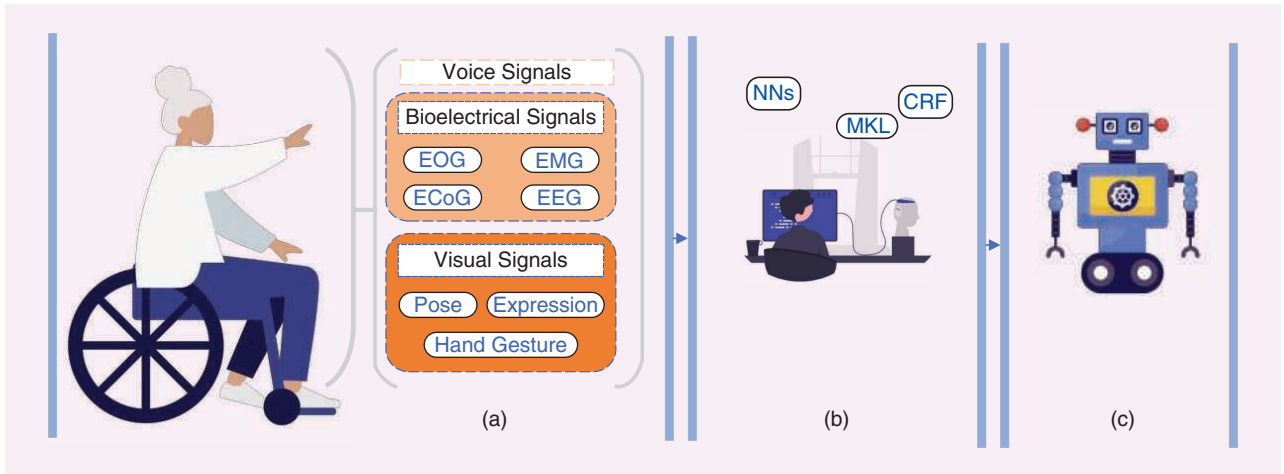


Figure 7. The process of intention analysis using multiple sensing signals: (a) sensor information collection, (b) multimodal signals processing, and (c) intention recognition. EOG: electrooculogram; EMG: electromyography; ECoG: electrocorticography; EEG: electroencephalogram; MKL: multiple kernel learning; CRF: conditional random field.

Bioelectricity Signal

Biological signals directly come from the human body and can be divided into biological electrical and biomechanical signals. The collection of biological signals from human motion is mainly achieved by wearable devices (e.g., data gloves and electronic bracelets). In addition, some special sensors can collect the bioelectric-based signals [e.g., electrooculograms (EOGs), electromyography (EMG), electroencephalograms (EEGs), and electrocorticography (ECoG)]. Biomechanical signals can be detected by traditional physical sensors (e.g., joint position sensors, force-torque sensors, and pressure sensors). Combined with dynamics and kinematics, it is possible to construct a prediction model based on these biological signals to realize intention recognition, especially when the elderly cannot speak flexibly.

Although biomechanical signals can be perceived accurately by traditional sensors, due to the lag in mechanical information in human motion, it is difficult for robots to predict human intentions in advance and respond rapidly. In addition, bioelectrical signals occur before body movement, which provides the advantage in response speed.

However, biochemical sensors are still limited by low frequency, weak amplitude, and a low signal-to-noise ratio. Dong et al. [83] proposed a dual-tree complex wavelet transform method to extract features from an EOG gathered from eye movements. A support vector machine is used to classify the features of the EOG so that the original eye movements can be converted into control commands for computers and rehabilitation aid devices. In [84], Golparvar et al. developed the use of graphene-coated fabric electrodes for the acquisition of different EOGs because graphene textile electrodes display a high degree of flexibility and stretchability, and the possibility of embedding them into various kinds of personal clothing exists. Bhuyain et al. [95] presented an EOG-based system to control an electric wheelchair for people suffering from quadriplegia and quadriparesis, in which four EOG types corresponding to different directions of wheelchair movement are gathered. Tam et al. [85] leveraged a wearable high-density-surface EMG (HD-sEMG) sensor combined with a CNN to classify muscle activity for motorized hand prosthesis control. To address the variability affecting surface EMG (sEMG) accuracy, Zanghieri et al. [90] proposed

Table 3. The strengths and weaknesses of different sensor signals.

Signal type	Sensor	Examples	Strengths	Weaknesses
Bioelectricity signal	Bioelectric sensor	EOG [83], [84], EMG [85], EEG [86], and ECoG [87]	Real time Versatile, and sensitive	Expensive, uncomfortable to wear Difficult data processing caused by complex signals
Visual sense	RGB-D/Stereo/Optical camera	Pose [88], [89], hand gesture [85], [90], and expression [91], [92]	Widely used, low-cost, mature technology, available and well equipped	Information loss due to inappropriate light intensity, occlusion, deformation, and textureless regions
Audio information	Piezoelectric and ceramics/capacitive acoustic sensors	Language command [93], [94], [80]	Wide application Sufficient semantic information	Sensitive to noise Slow response speed and low accuracy

an approach based on the novel temporal convolutional network topology (TEMPONet). Inoue et al. [86] analyzed the characteristics of the recorded EEG by wearing EEG sensor devices, from which existing contaminated artifacts, such as eye blinking, lateral eye movement, and EMG activity, can be detected. Moreover, ECoG has been successfully used for controlling prosthetic limbs by classifying individual finger movements [87].

Visual Signal

Camera sensors have been widely used in human–robot interactions, due to their convenience and visual performance. In general, the service robot uses unique feature descriptors to extract target information from the image. To recognize the intention precisely, service robots should be sensitive to changes in the elderly's local features (e.g., facial expressions and hand gestures). In addition, the combination of different local features can effectively compensate for a deficiency in one aspect. In recent years, the visual sensors have shifted chiefly to depth cameras. Compared with the images generated by traditional cameras, depth images can offer extra spatial information, which plays an indispensable role in human intention recognition. Various depth cameras (e.g., the Kinect series from Microsoft and RealSense [96] from Intel) have been widely applied in robotics, such as human tracking and semantic segmentation. This section introduces several intention recognition methods based on visual information from the perspective of posture, expression, and gesture.

Posture Recognition

The large-scale application of deep NNs in image and video recognition has verified their excellent performance in recent years. In the elderly care context, service robots should collect environmental information and predict human intentions simultaneously. A typical approach is to estimate human posture through video and images. Zhe et al. [88] used partial affinity fields to realize 2D multiperson pose detection through image sequences. The proposed architecture can efficiently predict the position of different body parts and their associated vector representations, becoming a benchmark in 2D posture estimation. The occlusion problem in a 3D environment has always been a focus of posture estimation. Therefore, Casalino et al. [97] proposed a human pose estimation algorithm combined with a particle filter to solve the occlusion problem in human–robot collaborative scenes. In addition to the traditional human body posture recognition methods that widely use RGB/RGB-D information, Fürst et al. [89] added the laser radar to improve the posture estimation performance. Based on the end-to-end architecture using RGB and lidar, a method for forecasting absolute 3D human body posture with high precision was proposed. In addition, they introduced a method of weak supervision simultaneously, in which 3D prediction is generated by 2D pose annotation. In [98], the problem of intention recognition in different

environments is solved by training a visual perception network based on human motion capture, and a conditional probability solution of specific environmental features was calculated through the discriminant sensor network jointly trained with long short-term memory and multilayer perception. Moreover, human–robot interaction datasets should be established based on different scenarios for model training. Luo et al. [99] proposed a human pose estimation method based on a constrained image structure model, which utilized the tracking algorithm to detect the position and size of the target in the images. Then, the GrabCut algorithm was used to extract foreground information as the input of the estimation stage. Finally, the primary segmentation method is applied in the estimation stage so that the proposed method can reduce the network input's search space and the cost of model calculations.

Facial Expression Recognition

Human expressions can intuitively reflect the elderly's current state (e.g., anger, joy, and surprise). The robot can recognize human expressions extracted from facial features, providing a reliable basis for the robot to take corresponding countermeasures. The current mainstream facial expression recognition technology mainly adopts data-driven methods, which consider the recognition process a classification or regression problem, such as obtaining a specific facial expression format from the input image or sequence [100] and rating users' emotions through facial expressions [101].

In general, the change in human expression is a dynamic process. Therefore, in some methods, dynamic information about facial expressions is considered by extracting spatiotemporal features and using dynamic learning methods. Yu et al. [91] proposed a deeper cascaded peak-piloted network that uses a more profound and extensive architecture to increase the discriminative ability of the learned representations. Liu et al. [92] balanced the complex distribution of intraclass and interclass variations by combining the deep metric loss and softmax loss in a unified two fully connected layer branch framework via joint optimization to obtain better facial expression recognition performance.

Hand Gesture Recognition

Gesture recognition based on vision is already very mature. It often uses visual sensors to extract human gestures from the environment and then employs various algorithms to extract and classify gesture features from the image. Service robots can complete some basic tasks according to the gesture information obtained. A complete vision-based gesture recognition system generally includes the following three modules: detection, tracking, and recognition [102]. The detection module mainly includes gesture segmentation from image frames, while tracking can provide users with real-time recognition in a dynamic environment. Tam et al. [85] realized real-time hand

gesture recognition by leveraging a wearable HD-sEMG sensor with a CNN to classify muscle activity. Zanghieri et al. [90] used TEMPONet to address the temporal variability problem, which can greatly affect the intersession generalization of sEMG-based hand gesture recognition.

Audio Signal

Language is the primary way of conveying instructions between humans and robots. It is almost the most common way to identify an individual's intentions, where the elderly communicate their intentions directly to the robot through voice communication. Therefore, giving robots voice communication capabilities can improve their interaction with the elderly. Language ability can improve the rapport between users and their robot assistants, allowing the elderly to obtain more satisfactory services from their robot companions and give robots more trust. In addition, verbal commands can make up for other signals of intention recognition to some extent. For example, it is difficult for robots to recognize facial expressions and hand gestures when encountering the elderly who have wrinkles on their faces and are unable to act.

Audio signals are often parsed into semantic fragments that robots can understand through speech recognition and natural language processing. In [93], Wang et al. presented a sequence-to-sequence model to translate natural language instructions into action sequences, using Bidirectional Encoder Representations from Transformers as the encoder and the decoder transformer as the language generator to plan for the visual semantic environment. Tiwari et al. [94] developed a domain-agnostic framework, which creates natural language interfaces for computer-controlled systems. The proposed framework employed ontologies built using Web Ontology Language for knowledge representation and machine learning models for language processing tasks. In another work [80], Mohit et al. proposed a language conditional IL agent framework, CLI-Port, to solve assigned tasks that are not explicitly represented by object posture, instance segmentation, symbolic state, and syntactic structure. However, it is important to point out that most speech recognition technologies have strict requirements for input, which may bring several problems for the elderly who cannot speak clearly and have an accent. A practical solution is to use multimodal signal fusion models for service robots, which is discussed in the following section.

Multimodal Signal Fusion Model

To meet various aspects of the needs of elderly users and improve the correctness of decision making, service robots working for elderly people in a specific environment should take multichannel sensor information as input. After the sensors receive the signals from the human body, the robot needs to fuse and analyze the multimodal sensor information to obtain the estimation of human intentions.

Research on multimodal signal fusion methods stems from three main benefits. First, interpreting the same phenomenon from different signal sources can effectively improve the system's robustness. Second, accessing different signals can capture complementary information that would be difficult to obtain in a separate mode. Third, multimodal systems can keep operating when one of these models is missing. For example, the system can alternately identify emotions from visual signals when the person cannot speak [103].

Learning-based methods have presented excellent performance when processing high-dimensional sensing signals. In addition, these methods are more robust for environmental changes and complex scenes, which makes knowledge transfer easy. Deep learning models have been successfully applied to many cases, e.g., speech recognition, image processing, natural language processing, and emotion analysis. Therefore, this article focuses on multimodal signal processing methods based on learning. Baltrušaitis et al. [104] described machine learning methods as model-based approaches, and three approaches for multimodal fusion are categorized, i.e., multiple-kernel learning (MKL) [105], graphical models, and NNs. MKL takes advantage of different kernels for different modalities of the data. The main applications include multimodal emotion recognition and analysis, object classification [106], and behavior recognition [107]. Filippone et al. [108] applied MKL to neural image data processing. Each kernel corresponds to either an imaging modality or an anatomically labeled region. Poria et al. [109] leveraged MKL to conduct emotion recognition and sentiment analysis by classifying features from visual and textual modalities. Due to the ability of graphical models to efficiently exploit the spatial and temporal structure of the data, models, such as the conditional random field (CRF), are also well-known to deal with multimodal information [104]. Chang et al. [110] proposed a CRF model for gesture recognition. In this case, the unary term is formalized by applying a generative probabilistic approach and defined in a non-parametric manner. Consequently, different cues from the multiple features can be probabilistically fused into the unary potential.

To address the ambiguity of instructions in vision and language tasks, Hao et al. presented a pretrained encoder to align language instructions and visual states as joint representations [111]. The model takes image-text-action triplets at each time step as input and outputs the prediction of masked word tokens and following actions. Zeng et al. [112] adequately considered multiple modalities missing in sentiment analysis. They designed a tag that encoded the module to focus on the missing modalities and leveraged a transformer encoder-decoder network to learn the missing modality features. Kumar et al. [112] used a text transformer as the textual encoder and combined the faster RCNN with a ResNet feature extractor as an optical encoder to conduct practical feedback synthesis.

Challenge

Experimental Design

Although we have reviewed various approaches to implementing the perception and learning functions of elderly service robots, integrating these technologies into the care environment is still a considerable challenge. From this study, there are several notes, which are as follows:

- 1) How do we seamlessly integrate subsystems of elderly service robots?
- 2) How can the simulation result be transferred to actual service robots?
- 3) How can benchmarking approaches be designed for elderly service robots?
- 4) How do we design elderly service robots that deliver effective, robust, and ethical care under a specific context, such as a user's health condition and the deployment site and duration?

Undoubtedly, powerful simulation can significantly help researchers when developing robotic applications. It is possible to fuse various algorithms and test comprehensive robotic ability under different conditions. The research of robotic deep RL has proved the great promise of simulated control for robotic skill learning [114]. However, even the perfect simulation is still different from the real-world settings when the robot works. One solution is to increase the task difficulty when training algorithms and evaluate the performance of robots in the real world. Then the closed loop from the real world to the simulation can be realized so that the following virtual world will become more accurate, and our trained robot will perform better. In addition to training algorithms with simulators, another way is to use Wizard of Oz to bypass technical limitations [115]. For example, considering the design requirements and users' reception, human operators can control the robot directly to provide prior knowledge, which can effectively improve the performance of robots during human–robot interaction. As previously discussed, human–robot interaction contains two main parts: physical collaboration and a cognitive aspect. Benchmarking approaches for elderly service robots should comprehensively consider the subjective feedback of the elderly and the quality of task execution.

Trust Issue

The attitudes of the elderly are usually fragile and sometimes even eccentric. To effectively maintain long-term, stable, and harmonious communication between the aged and robots, research regarding trust and other ethical considerations during human–robot collaboration plays an important role [116], as trust comes from the security brought by robots and directly affects the willingness of humans to collaborate. We believe that the elderly's trust in robots is a significant factor in measuring the performance of robots.

Trust and safety issues are multifaced in human–robot interaction, where human safety should be a priority. In

2016, the International Organization for Standardization (ISO) promulgated an industrial standard (ISO/Technical Specification 15066:2016) for collaborative robots [117], which specifies the criteria for safe physical contact during human–robot interactions. During motion planning, collaborative robots need to refer to the human body and detect collisions according to the physical contact position. Some studies [118], [119], [120] have shown that the reasonable consideration of ergonomics and robots cooperating by adapting to human motion can help improve the work quality of cooperative robots, thus enhancing people's trust in robots. Debora et al. [121] studied the influence of anthropomorphic robots on human trust during human–robot interactions and showed that humans are more willing to accept a less humanlike robot. Conversely, Samuele et al. [122] investigated a scene where a robot is a trustor, focusing on a humanoid robot's cognitive architecture using theory of mind. In addition, high-performance collaborative robots are often expensive, so the safety of robots when working also affects human trust to a certain extent.

Just as humans are more likely to believe in workers with solid abilities, collaborative robots' efficiency and quality directly affect human trust during human–robot collaboration. An excellent service provider should be able to pay attention to the elderly's preferences and privacy issues. In addition, robots need to ensure the safety of objects when performing tasks. For example, when handing over a cup of water, robots should avoid a collision between the cup and the external environment and keep the cup as stable as possible to prevent water from spilling. This should be especially considered when robots manipulate valuable, dangerous, and fragile objects.

Learning Capability

A major challenge in the accompanying environment is a reliable and real-time response of the robot according to many possible tasks. In this context, robots must be capable of staying on standby for a long time and performing some more complex tasks. For example, consider a common task where an older man asks the robot to take out the garbage. The robot assistant should be able to clean up the garbage and throw it away at the designated location, which includes several tasks, such as navigation, positioning, and classification. Obviously, a simple function configuration cannot meet such requirements, and a complex task often requires a robot to possess dozens of essential skills and combine them.

Several works related to LfD divide skills into low-level skills and high-level skills, and a high-level skill can be considered a set of low-level skills [14], [113], [123]. However, current research in robot skill learning focuses on respective work (i.e., navigation, walking, and handover). Consequently, there is no unified robot platform to integrate these tasks. In addition, the effect of robot skill learning based on data directly relies on the accuracy and generalization of data, leading to an inability to guarantee the robot's performance in the natural environment.

Conclusion

In this review, different aspects of elderly service robots were discussed. We investigated several components of human–robot interaction, such as environment perception, compliance control, and intention recognition. To accompany and attend to the elderly, service robots should be equipped with sophisticated sensors to perceive the unstructured and dynamic surrounding environment. Furthermore, considering the various needs of the elderly in practice, how to design inbuilt complex algorithms for elderly service robots related to intention recognition and robot skill learning is another critical problem. Only when the robot correctly recognizes the human intention and completes the corresponding task efficiently can the process of human–robot collaboration be considered successful. Finally, we put forward several challenges in maintaining stable and harmonious communication and cooperation between humans and robots (e.g., the trust issue). We believe this review covers the essential elements of elderly service robots and can provide a starting point for researchers interested in robotics.

Acknowledgment

This work was supported, in part, by the National Key R&D program of China, under grant 2018AAA0100803; National Natural Science Foundation, under grants 62276090 and 62203150; Key Research and Development Program of Jiangsu, under grants BK20192004B and BE2018004-04; China Postdoctoral Science Foundation, under grant 2021M701051; Jiangsu Province Excellent Postdoctoral Program, under grant 2022ZB192; Changzhou Basic Research Program (Application Program), under grant CJ20220051; Guangdong Forestry Science and Technology Innovation Project, under grant 2020KJCX005; Open Research Fund of State Key Laboratory of Bioelectronics; and State Key Laboratory of Integrated Management of Pest Insects and Rodents, under grant IPM1914. Xiaofeng Liu is the corresponding author.

About the Authors

Xiaofeng Liu (xfliu@hhu.edu.cn) earned his B.S. degree in electronic engineering and M.S. degree in computer application from Taiyuan University of Technology, Taiyuan, China, in 1996 and 1999, respectively, and his Ph.D. degree in biomedical engineering from Xi'an Jiaotong University, Xi'an, China, in 2006. In 2006, he joined as an associate professor with the Shandong University of Science and Technology, Qingdao, China. From 2008 to 2011, he was a postdoctoral researcher with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. Since 2010, he has been with Hohai University, Changzhou 213022, China, where he is currently a professor and the vice director of the Jiangsu Key Laboratory of Special Robots. His research interests include the study of nature-inspired navigation and human–robot interaction. He is a Member of IEEE.

Congyu Huang (925668554@qq.com) earned his B.S. degree from Nanjing Forestry University, Nanjing, China, in

2021. He is currently an M.S. degree student at Hohai University, Changzhou 213022, China, where he is a member of the Jiangsu Key Laboratory of Special Robots. His research interests include social robots, learning from demonstration, grasp detection, and human–robot interaction.

Haoran Zhu (871735868@qq.com) earned his B.S. degree from Suzhou University of Science and Technology, Suzhou, China, in 2021. He is currently an M.S. degree student at Hohai University, Changzhou 213022, China. His research interests include human–robot interaction, robotic grasp, and motion planning.

Ziyang Wang (1599201184@qq.com) earned his B.S. degree from Hohai University, Changzhou, China, in 2021. He is currently an M.S. degree student at Hohai University, Changzhou 213022, China. His research interests include manipulator teleoperation, human–machine instruction, and missing data restoration.

Jie Li (jerry1165530693@163.com) earned his M.S. degree from East China Jiaotong University, Nanchang, China, in 2016 and his Ph.D. degree in control theory and control engineering from the School of Control Science and Engineering, Dalian University of Technology, Dalian, China, in 2021. He is currently a postdoctoral research fellow with the College of Internet of Things Engineering, Hohai University, Changzhou 213022, China. His research interests include motion capture, indoor localization, inertial navigation and positioning, and gait analysis.

Angelo Cangelosi (angelo.cangelosi@manchester.ac.uk) earned his Ph.D. degree in psychology and artificial intelligence from Università di Genova, Genova, Italia, in 1997. He is currently a professor of machine learning and robotics at the University of Manchester, M13 9RL Manchester, U.K. Previously, he was the coordinator of the European Union Horizon 2020 Marie Skłodowska-Curie European Industrial Doctorate “Applications of Personal Robotics Through Interaction and Learning.” He was a professor of artificial intelligence and cognition and the founding director of the Center for Robotics and Neural Systems, Plymouth University, Plymouth, U.K. He has been a visiting professor at Waseda University, Japan, and Sassari and Messina Universities, Italy. His research interests include artificial intelligence and cognitive robotics. He was the chair of the IEEE Technical Committee on Autonomous Mental Development from 2012 to 2013. He is an editor of *Interaction Studies* and *IET Cognitive Computation and Systems* and was the editor-in-chief of *IEEE Transactions on Autonomous Development* in 2015. He is a Senior Member of IEEE.

References

- [1] E. H. Østergaard, “The role of cobots in industry 4.0,” Universal Robots, Odense, Denmark, White Paper, 2017. [Online]. Available: <https://info.universal-robots.com/hubs/Enablers/White%20papers/The%20role%20of%20cobots%20in%20industry.pdf>
- [2] A. Billard and D. Kragic, “Trends and challenges in robot manipulation,” *Science*, vol. 364, no. 6446, p. 1149, Jun. 2019, doi: 10.1126/science.aat8414.
- [3] M. Tavakoli et al., “Robotics, smart wearable technologies, and autonomous intelligent systems for healthcare during the COVID-19 pandemic: An analysis of the state

- of the art and future vision," *Adv. Intell. Syst.*, vol. 2, no. 7, Jul. 2020, Art. no. 2000071, doi: 10.1002/aisy.202000071.
- [4] G.-Z. Yang et al., "Combating COVID-19—The role of robotics in managing public health and infectious diseases," *Sci. Robot.*, vol. 5, no. 40, Mar. 2020, Art. no. eabb5589, doi: 10.1126/scirobotics.abb5589.
- [5] T. Mukai et al., "Development of the tactile sensor system of a human-interactive robot 'RI-MAN,'" *IEEE Trans. Robot.*, vol. 24, no. 2, pp. 505–512, Apr. 2008, doi: 10.1109/TRO.2008.917006.
- [6] M. A. Salichs et al., "Mini: A new social robot for the elderly," *Int. J. Social Robot.*, vol. 12, no. 6, pp. 1231–1249, Dec. 2020, doi: 10.1007/s12369-020-00687-0.
- [7] A. Gardecki and M. Podpora, "Experience from the operation of the pepper humanoid robots," in *Proc. Prog. Appl. Electr. Eng. (PAEE)*, Koscielisko, Poland, Jun. 2017, pp. 1–6, doi: 10.1109/PAEE.2017.8008994.
- [8] H. Iwata and S. Sugano, "Design of human symbiotic robot TWENDY-ONE," in *Proc. IEEE Int. Conf. Robot. Automat.*, Kobe, Japan, May 2009, pp. 580–586.
- [9] A. Ajoudani et al., "Progress and prospects of the human–robot collaboration," *Auton. Robots*, vol. 42, no. 5, pp. 957–975, Oct. 2017, doi: 10.1007/s10514-017-9677-2.
- [10] A. Thomaz et al., "Computational human-robot interaction," *Found. Trends Robot.*, vol. 4, nos. 2–3, pp. 104–223, Dec. 2016, doi: 10.1561/23000000049.
- [11] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robot. Auton. Syst.*, vol. 42, nos. 3–4, pp. 143–166, Mar. 2003, doi: 10.1016/S0921-8890(02)00372-X.
- [12] A. Freedy et al., "Measurement of trust in human-robot collaboration," in *Proc. Int. Symp. Collaborative Technol. Syst.*, Orlando, FL, USA, May 2008, pp. 106–114, doi: 10.1109/CTS.2007.4621745.
- [13] A. K. Pandey and R. Alami, "Towards human-level semantics understanding of human-centered object manipulation tasks for HRI: Reasoning about effect, ability, effort and perspective taking," *Int. J. Social Robot.*, vol. 6, no. 4, pp. 593–620, Jul. 2014, doi: 10.1007/s12369-014-0246-y.
- [14] B. Zheng et al., "Imitation learning: Progress, taxonomies and opportunities," 2021, *arXiv:2106.12177*.
- [15] S. Suwa et al., "Home-care professionals' ethical perceptions of the development and use of home-care robots for older adults in Japan," *Int. J. Human-Comput. Interact.*, vol. 36, no. 14, pp. 1295–1303, Aug. 2020, doi: 10.1080/10447318.2020.1736809.
- [16] M. Tsujimura et al., "The essential needs for home-care robots in Japan," *J. Enabling Technol.*, vol. 14, no. 4, pp. 201–220, Nov. 2020, doi: 10.1108/JET-03-2020-0008.
- [17] T. Tanioka et al., *Nursing Robots: Robotic Technology and Human Caring for the Elderly*. Koscielisko, Poland: Fukuro Shuppan Publishing, 2017.
- [18] Y. Yagi, H. Okumura, and M. Yachida, "Multiple visual sensing system for mobile robot," in *Proc. IEEE Int. Conf. Robot. Automat.*, 1994, vol. 2, pp. 1679–1684, doi: 10.1109/ROBOT.1994.351350.
- [19] L. Guo et al., "Environmental perception of mobile robot," in *Proc. IEEE Int. Conf. Inf. Acquisition*, 2007, pp. 348–352, doi: 10.1109/ICIA.2006.306024.
- [20] K. Wang et al., "A 6D-ICP approach for 3D reconstruction and motion estimate of unknown and non-cooperative target," in *Proc. Chinese Contr. Decis. Conf. (CCDC)*, 2016, pp. 6366–6370, doi: 10.1109/CCDC.2016.7532144.
- [21] D. Wang et al., "Multi-view 3D reconstruction with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 5702–5711, doi: 10.1109/ICCV48922.2021.00567.
- [22] Y. Furukawa and C. Hernández, "Multi-view stereo: A tutorial," *Found. Trends Comput. Graph. Vis.*, vol. 9, nos. 1–2, pp. 1–148, Jun. 2015, doi: 10.1561/06000000052.
- [23] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Proc. IEEE Int. Conf. Robot. Automat.*, 1987, vol. 4, pp. 850–850, doi: 10.1109/ROBOT.1987.1087846.
- [24] S. Julier and J. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004, doi: 10.1109/JPROC.2003.823141.
- [25] R. Sim, P. Elinas, and J. Little, "A study of the rao-blackwellised particle filter for efficient and accurate vision-based slam," *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 303–318, Jul. 2007, doi: 10.1007/s11263-006-0021-0.
- [26] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality*, 2007, pp. 225–234, doi: 10.1109/ISMAR.2007.4538852.
- [27] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular slam system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015, doi: 10.1109/TRO.2015.2463671.
- [28] W. Shao et al., "Stereo visual inertial lidar simultaneous localization and mapping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2019, pp. 370–377, doi: 10.1109/IROS40897.2019.8968012.
- [29] C.-H. Yeh and M.-H. Lin, "Robust 3D reconstruction using HDR-based SLAM," *IEEE Access*, vol. 9, pp. 16,568–16,581, Jan. 2021, doi: 10.1109/ACCESS.2021.3051257.
- [30] S. Triputen et al., "Methodology to analyze the accuracy of 3D objects reconstructed with collaborative robot based monocular LSD-slam," in *Proc. Int. Conf. Intell. Auton. Syst. (ICoIAS)*, 2018, pp. 185–190, doi: 10.1109/ICoIAS.2018.8494109.
- [31] R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.
- [32] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.
- [33] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
- [34] K. He et al., "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.
- [35] Z. Huang et al., "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 6402–6411, doi: 10.1109/CVPR.2019.00657.
- [36] S. Zagoruyko et al., "A multipath network for object detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, BMVA Press, Sep. 2016, pp. 15.1–15.12.
- [37] D. Bolya et al., "YOLACT++: Better real-time instance segmentation," Dec. 2019, *arXiv:1912.06218*.
- [38] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2858–2866, doi: 10.1109/CVPR.2017.305.
- [39] A. Kirillov et al., "InstanceCut: From edges to instances with multiCut," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7322–7331, doi: 10.1109/CVPR.2017.774.
- [40] P. O. Pinheiro, R. Collobert, and P. Dollar, *Learning to Segment Object Candidates*. Cambridge, MA, USA: MIT Press, 2015.
- [41] J. Dai et al., "Instance-sensitive fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis. – ECCV*, Cham, Switzerland: Springer, 2016, pp. 534–549, doi: 10.1007/978-3-319-46466-4_32.
- [42] X. Chen et al., "Tensormask: A foundation for dense object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2061–2069, doi: 10.1109/ICCV.2019.00215.
- [43] H. Ahn, "Object handling of cognitive robots using deep learning based object recognition," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*, 2019, pp. 150–153, doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00067.
- [44] M. Fu et al., "Mobile robot object recognition in the internet of things based on fog computing," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, 2019, pp. 1838–1842, doi: 10.1109/APSIPAASC47483.2019.9023187.
- [45] I. F. E. Babila et al., "Object detection for inventory stock counting using yolov5," in *Proc. IEEE 18th Int. Colloq. Signal Process. Appl. (CSPA)*, 2022, pp. 304–309, doi: 10.1109/CSPA55076.2022.9782028.

- [46] J. Kim et al., "Implementation of a mobile multi-target search system with 3D slam and object localization in indoor environments," in *Proc. 21st Int. Conf. Control, Automat. Syst. (ICCAS)*, 2021, pp. 2083–2085, doi: 10.23919/ICCAS52745.2021.9650063.
- [47] H. Zhang et al., "A survey on instance segmentation: Recent advances and challenges," *Int. J. Innov. Comput., Inf. Control (IJICIC)*, vol. 17, pp. 1041–1053, Jun. 2021, doi: 10.1007/s13735-020-00195-x.
- [48] A. Li, M. Danielczuk, and K. Goldberg, "One-shot shape-based amodal-to-modal instance segmentation," in *Proc. IEEE 16th Int. Conf. Automat. Sci. Eng. (CASE)*, 2020, pp. 1375–1382, doi: 10.1109/CASE48305.2020.9216733.
- [49] M. Tomono, "Building an object map for mobile robots using LRF scan matching and vision-based object recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2004, vol. 4, pp. 3765–3770, doi: 10.1109/ROBOT.2004.1308855.
- [50] J.-H. Park et al., "A new object recognition system for service robots in the smart environment," in *Proc. Int. Conf. Contr. Automat. Syst.*, 2007, pp. 1083–1087, doi: 10.1109/ICCAS.2007.4407060.
- [51] J. Yang et al., "A method for indoor robots visual perception based on multi-feature fusion," in *Proc. Int. Conf. E-Product E-Service E-Entertainment*, 2010, pp. 1–4, doi: 10.1109/ICEEE.2010.5661339.
- [52] B. Akgun et al., "Learning tasks and skills together from a human teacher," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011, pp. 1868–1869, doi: 10.1609/aaai.v25i1.8002.
- [53] S. Calinon and A. Billard, "Incremental learning of gestures by imitation in a humanoid robot," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2007, pp. 255–262, doi: 10.1145/1228716.1228751.
- [54] X. Wei et al., "Robotic skills learning based on dynamical movement primitives using a wearable device," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, 2017, pp. 756–761, doi: 10.1109/ROBIO.2017.8324508.
- [55] D. McColl and G. Nejat, "Recognizing emotional body language displayed by a human-like social robot," *Int. J. Social Robot.*, vol. 6, no. 2, pp. 261–280, Apr. 2014, doi: 10.1007/s12369-013-0226-7.
- [56] L. Cai et al., "Color recognition and dynamic decision-making model of 6 axis of industrial robot based on embedded system," in *Proc. 29th Chinese Control Decis. Conf. (CCDC)*, May 2017, pp. 2836–2841, doi: 10.1109/CCDC.2017.7978995.
- [57] J. Nishihara et al., "Online algorithm for robots to learn object concepts and language model," *IEEE Trans. Cognit. Develop. Syst.*, vol. 9, no. 3, pp. 255–268, Apr. 2016, doi: 10.1109/TCDS.2016.2552579.
- [58] Y. Xiang et al., "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robot., Sci. Syst. XIV*, H. KressGazit, S. Srinivasa, T. Howard, and N. Atanasov, Eds. Jun. 2018, doi: 10.15607/RSS.2018.XIV.019.
- [59] P. Anderson et al., "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3674–3683.
- [60] R. Paul et al., "Efficient grounding of abstract spatial concepts for natural language interaction with robot platforms," *Int. J. Robot. Res.*, vol. 37, no. 10, pp. 1269–1299, Sep. 2018, doi: 10.1177/0278364918777627.
- [61] X. Wei, R. Yu, and J. Sun, "View-GCN: View-based graph convolutional network for 3D shape analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1850–1859, doi: 10.1109/CVPR42600.2020.00192.
- [62] W. Yang et al., "Human grasp classification for reactive human-to-robot handovers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, 2020, pp. 11,123–11,130, doi: 10.1109/IROS45743.2020.9341004.
- [63] L. Shao et al., "Concept2robot: Learning manipulation concepts from instructions and human demonstrations," *Int. J. Robot. Res.*, vol. 40, nos. 12–14, pp. 1419–1434, Oct. 2021, doi: 10.1177/02783649211046285.
- [64] Y. Guan, N. Wang, and C. Yang, "Review of the techniques used in motor-cognitive human-robot skill transfer," *Cogn. Comput. Syst.*, vol. 3, no. 3, pp. 229–252, Sep. 2021, doi: 10.1049/ccs2.12025.
- [65] N. Jaquier et al., "Learning from demonstration with model-based gaussian process," in *Proc. Conf. Robot Learn.*, Oct. 2020, pp. 247–257.
- [66] M. Hersch et al., "Dynamical system modulation for robot learning via Kinesthetic demonstrations," *IEEE Trans. Robot.*, vol. 24, no. 6, pp. 1463–1467, Dec. 2008, doi: 10.1109/TRO.2008.2006703.
- [67] Y. Derimis and G. Hayes, "Imitations as a dual-route process featuring predictive and learning components: A biologically plausible computational model," in *Imitation Animals Artifacts*. Cambridge, MA, USA: MIT Press, 2002, pp. 327–361.
- [68] M. Niclescu and M. Mataric, "Experience-based representation construction: Learning from human and robot teachers," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. Expanding Societal Role Robot. Next Millennium*, 2001, vol. 2, pp. 740–745, doi: 10.1109/IROS.2001.976257.
- [69] U. Nehmzow et al., "Robot programming by demonstration through system identification," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2007, pp. 801–806, doi: 10.1109/IROS.2007.4399087.
- [70] A. Y. Ng et al., "Autonomous inverted helicopter flight via reinforcement learning," in *Experimental Robotics IX*, M. H. Ang and O. Khatib, Eds. Berlin, Germany: Springer-Verlag, 2006, pp. 363–372.
- [71] B. Browning et al., "Skill acquisition and use for a dynamically-balancing soccer robot," in *Proc. 19th Nat. Conf. Artif. Intell. (AAAI)*, 2004, pp. 599–604.
- [72] P. K. Pook and D. H. Ballard, "Recognizing teleoperated manipulations," in *Proc. IEEE Int. Conf. Robot. Automat.*, 1993, pp. 578–585, doi: 10.1109/ROBOT.1993.291896.
- [73] C. Yang et al., *Robot Learning Human Skills and Intelligent Control Design*. Boca Raton, FL, USA: CRC, 2021.
- [74] V. M. Sloutsky, "From perceptual categories to concepts: What develops?" *Cogn. Sci.*, vol. 34, no. 7, pp. 1244–1286, Aug. 2010, doi: 10.1111/j.1551-6709.2010.01129.x.
- [75] D. Ahmedt-Aristizabal et al., "Graph-based deep learning for medical diagnosis and analysis: Past, present and future," *Sensors*, vol. 21, no. 14, Jul. 2021, Art. no. 4758, doi: 10.3390/s21144758.
- [76] C. Finn et al., "Deep spatial autoencoders for visuomotor learning," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, 2016, pp. 512–519, doi: 10.1109/ICRA.2016.7487173.
- [77] W. D. Smart and L. P. Kaelbling, "Effective reinforcement learning for mobile robots," in *Proc. IEEE Int. Conf. Robot. Automat. (Cat. No. 02CH37292)*, 2002, vol. 4, pp. 3404–3410, doi: 10.1109/ROBOT.2002.1014237.
- [78] B. Schölkopf et al., *An Application of Reinforcement Learning to Aerobatic Helicopter Flight*. Cambridge, MA, USA: MIT Press, 2007, pp. 1–8.
- [79] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Proc. Conf. Robot Learn.*, PMLR, 2022, pp. 91–100.
- [80] M. Shridhar et al., "CLIPort: What and where pathways for robotic manipulation," in *Proc. Conf. Robot Learn.*, PMLR, 2022, pp. 894–906.
- [81] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8748–8763.
- [82] A. Zeng et al., "Transporter networks: Rearranging the visual world for robotic manipulation," 2020, *arXiv:2010.14406*.
- [83] E. Dong, C. Li, and C. Chen, "An EOG signals recognition method based on improved threshold dual tree complex wavelet transform," in *Proc. IEEE Int. Conf. Mechatron. Automat.*, 2016, pp. 954–959, doi: 10.1109/ICMA.2016.7558691.
- [84] A. J. Golparvar and M. Kaya Yapici, "Wearable graphene textile-enabled EOG sensing," in *Proc. IEEE Sens.*, 2017, pp. 1–3, doi: 10.1109/ICSENS.2017.8234242.
- [85] S. Tam et al., "A fully embedded adaptive real-time hand gesture classifier leveraging HD-sEMG and deep learning," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 2, pp. 232–243, Apr. 2020, doi: 10.1109/TBCAS.2019.2955641.

- [86] R. Inoue et al., "Recording and characterization of EEGs by using wearable EEG device," in *Proc. 19th Int. Conf. Contr., Automat. Syst. (ICCAS)*, 2019, pp. 194–197, doi: 10.23919/ICCAS47443.2019.8971564.
- [87] G. Hotson et al., "Individual finger control of a modular prosthetic limb using high-density electrocorticography in a human subject," *J. Neural Eng.*, vol. 13, no. 2, 2016, Art. no. 026017, doi: 10.1088/1741-2560/13/2/026017.
- [88] Z. Cao et al., "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2018, doi: 10.1109/TPAMI.2019.2929257.
- [89] M. Fürst et al., "HPERL: 3D human pose estimation from RGB and lidar," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 7321–7327, doi: 10.1109/ICPR48806.2021.9412785.
- [90] M. Zanghieri et al., "Robust real-time embedded EMG recognition framework using temporal convolutional networks on a multicore IoT processor," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 2, pp. 244–256, Apr. 2020, doi: 10.1109/TBCAS.2019.2959160.
- [91] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *Vis. Comput.*, vol. 34, no. 12, pp. 1691–1699, Dec. 2018, doi: 10.1007/s00371-017-1443-0.
- [92] X. Liu et al., "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 522–531, doi: 10.1109/CVPRW.2017.79.
- [93] K. Wang et al., "Visual semantic planning for service robot via natural language instructions," in *Proc. China Automat. Congr. (CAC)*, 2021, pp. 793–798, doi: 10.1109/CAC53003.2021.9728603.
- [94] S. Tiwari et al., "Domain-agnostic context-aware framework for natural language interface in a task-based environment," in *Proc. IEEE 45th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, 2021, pp. 15–20, doi: 10.1109/COMPSAC51774.2021.00014.
- [95] M. F. Bhuyain et al., "Design and development of an EOG-based system to control electric wheelchair for people suffering from quadriplegia or quadriparesis," in *Proc. Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST)*, 2019, pp. 460–465, doi: 10.1109/ICREST.2019.8644378.
- [96] L. Keselman et al., "Intel(R) realsense(TM) stereoscopic depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 1267–1276, doi: 10.1109/CVPRW.2017.167.
- [97] A. Casalino et al., "Human pose estimation in presence of occlusion using depth camera sensors, in human-robot coexistence scenarios," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2018, pp. 1–7, doi: 10.1109/IROS.2018.8593816.
- [98] D. Wei et al., "A vision-based measure of environmental effects on inferring human intention during human robot interaction," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4246–4256, Mar. 2022, doi: 10.1109/JSEN.2021.3139593.
- [99] W. Luo and Q. Ruan, "Human pose estimation based on constraint pictorial structure model," in *Proc. 12th Int. Conf. Signal Process. (ICSP)*, 2014, pp. 1257–1262, doi: 10.1109/ICOSP.2014.7015201.
- [100] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul./Sep. 2022, doi: 10.1109/TAFFC.2020.2981446.
- [101] T. Hassan et al., "Automatic detection of pain from facial expressions: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1815–1831, Jun. 2021, doi: 10.1109/TPAMI.2019.2958341.
- [102] S. S. Rautaray et al., "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, Jan. 2015, doi: 10.1007/s10462-012-9356-9.
- [103] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, Feb. 2015, doi: 10.1145/2682899.
- [104] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.
- [105] M. Nen, Alpayd and N. Ethem, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, Jan. 2011.
- [106] J. Yang et al., "Group-sensitive multiple kernel learning for object categorization," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 436–443, doi: 10.1109/ICCV.2009.5459172.
- [107] Y. Yi, P. Hu, and X. Deng, "Human action recognition with salient trajectories and multiple kernel learning," *Multimedia Tools Appl.*, vol. 77, no. 14, pp. 17,709–17,730, Jul. 2018, doi: 10.1007/s11042-017-5209-5.
- [108] M. Filippone et al., "Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities," *Ann. Appl. Statist.*, vol. 6, no. 4, pp. 1883–1905, Dec. 2012, doi: 10.1214/12-AOAS562.
- [109] S. Poria et al., "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, 2016, pp. 439–448, doi: 10.1109/ICDM.2016.0055.
- [110] J. Y. Chang, "Nonparametric feature matching based conditional random fields for gesture recognition from multi-modal video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1612–1625, Aug. 2016, doi: 10.1109/TPAMI.2016.2519021.
- [111] W. Hao et al., "Towards learning a generic agent for vision-and-language navigation via pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 13,134–13,143, doi: 10.1109/CVPR42600.2020.01315.
- [112] J. Zeng, T. Liu, and J. Zhou, "Tag-assisted multimodal sentiment analysis under uncertain missing modalities," 2022, *arXiv:2204.13707*.
- [113] B. D. Argall et al., "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, May 2009, doi: 10.1016/j.robot.2008.10.024.
- [114] J. Ibarz et al., "How to train your robot with deep reinforcement learning: Lessons we have learned," *Int. J. Robot. Res.*, vol. 40, nos. 4–5, pp. 698–721, Apr. 2021, doi: 10.1177/0278364920987859.
- [115] L. D. Riek, "Wizard of oz studies in HRI: A systematic review and new reporting guidelines," *J. Human-Robot Interact.*, vol. 1, no. 1, pp. 119–136, Jul. 2012, doi: 10.5898/JHRI.1.1.Riek.
- [116] A. J. C. Sharkey and N. E. Sharkey, "Granny and the robots: Ethical issues in robot care for the elderly," *Ethics Inf. Technol.*, vol. 14, no. 1, pp. 27–40, Mar. 2012, doi: 10.1007/s10676-010-9234-6.
- [117] "Robots and robotic devices: Collaborative robots," ISO-15066, 2016.
- [118] B. Busch et al., "Postural optimization for an ergonomic human-robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017, pp. 2778–2785, doi: 10.1109/IROS.2017.8206107.
- [119] L. Pernel et al., "Towards ergonomic control of human-robot co-manipulation and handover," in *Proc. IEEE-RAS 17th Int. Conf. Humanoid Robot. (Humanoids)*, 2017, pp. 55–60, doi: 10.1109/HUMANOIDS.2017.8239537.
- [120] L. Pernel et al., "A selective muscle fatigue management approach to ergonomic human-robot co-manipulation," *Robot. Comput.-Integr. Manuf.*, vol. 58, pp. 69–79, Aug. 2019, doi: 10.1016/j.rcim.2019.01.013.
- [121] D. Zanatto et al., "Generalisation of anthropomorphic stereotype," *Int. J. Social Robot.*, vol. 12, no. 1, pp. 163–172, Jan. 2020, doi: 10.1007/s12369-019-00549-4.
- [122] S. Vinanzi et al., "Would a robot trust you? Developmental robotics model of trust and theory of mind," *Philos. Trans. Roy. Soc. B Biol. Sci.*, vol. 374, no. 1771, 2019, Art. no. 20180032, doi: 10.1098/rstb.2018.0032.
- [123] S. Ambhore, "A comprehensive study on robot learning from demonstration," in *Proc. 2nd Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, 2020, pp. 291–299, doi: 10.1109/ICIMIA48430.2020.9074946.