*electronics*

MDPI

*Review*

# On the Gap between Domestic Robotic Applications and Computational Intelligence

**Junpei Zhong [1], Chaofan Ling [1], Angelo Cangelosi [2,3,4], Ahmad Lotfi [5] and Xiaofeng Liu [4,*]**

1 S.M. Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 511442, China; joni.zhong@ieee.org (J.Z.); wichaofan@mail.scut.edu.cn (C.L.)
2 AIRC, National Institute of Advanced Industrial Science and Technology, Aomi 2-3-26, Tokyo 135-0064, Japan; angelo.cangelosi@manchester.ac.uk
3 Department of Computer Science, University of Manchester, Oxford Rd, Manchester M13 9PL, UK
4 College of IoT Engineering, Hohai University, Changzhou 213022, China
5 Department of Computer Science, Nottingham Trent University, Nottingham NG11 8NS, UK; ahmad.lotfi@ntu.ac.uk
* Correspondence: xfliu@hhu.edu.cn

**Abstract:** Aspired to build intelligent agents that can assist humans in daily life, researchers and engineers, both from academia and industry, have kept advancing the state-of-the-art in domestic robotics. With the rapid advancement of both hardware (e.g., high performance computing, smaller and cheaper sensors) and software (e.g., deep learning techniques and computational intelligence technologies), robotic products have become available to ordinary household users. For instance, domestic robots have assisted humans in various daily life scenarios to provide: (1) physical assistance such as floor vacuuming; (2) social assistance such as chatting; and (3) education and cognitive assistance such as offering partnerships. Crucial to the success of domestic robots is their ability to understand and carry out designated tasks from human users via natural and intuitive human-like interactions, because ordinary users usually have no expertise in robotics. To investigate whether and to what extent existing domestic robots can participate in intuitive and natural interactions, we survey existing domestic robots in terms of their interaction ability, and discuss the state-of-the-art research on multi-modal human–machine interaction from various domains, including natural language processing and multi-modal dialogue systems. We relate domestic robot application scenarios with state-of-the-art computational techniques of human–machine interaction, and discuss promising future directions towards building more reliable, capable and human-like domestic robots.

**Keywords:** domestic robotics; computational intelligence; robotic applications; natural communication



**Citation:** Zhong, J.; Ling, C.; Cangelosi, A.; Lotfi, A.; Liu, X. On the Gap between Domestic Robotic Applications and Computational Intelligence. *Electronics* **2021**, *10*, 793. https://doi.org/10.3390/electronics10070793

Academic Editor: Janos Botzheim, Savvas A. Chatzichristofis and Teresa Orlowska-Kowalska

Received: 28 December 2020
Accepted: 15 March 2021
Published: 27 March 2021



## 1. Introduction

With the rapid advancement in machine learning and computational intelligence models, faster Internet connection and increasingly affordable hardware, robotic products have become affordable household goods. According to a recent domestic robots market report (https://www.marketsandmarkets.com/Market-Reports/household-robot-market-253781130.html) (accessed on 1 January 2021), by 2024, the global domestic robots market will grow around 3 times compared to 2019. Aiming to assist ordinary users in domestic environment, the domestic robots nowadays (as of 2021) are typically designed to: (1) carry out one or several physical tasks (e.g., cleaning, delivery); (2) communicate with users to convey and obtain information (e.g., translating, command interpretation); (3) act as social companions to entertain users through social interaction (e.g., chatting). In order to provide these three main functions, domestic robots must be able to perceive and understand the surrounding environment and users, act accordingly and communicate with users in various scenarios, which should corresponds to different research fields, such as Natural Language Processing, Scenario Classification, Object Recognition, etc.

Commercial domestic robots are usually not equipped with the most advanced computational intelligence models. On one hand, commercial robotic products developed by industrial companies often prioritize the reliability and maintainability of robot performance to maximize the product life cycles of robotic products. On the other hand, the most advanced computational intelligence models are often built and prototyped in academia. Therefore, the techniques are often evaluated with pre-defined tasks or datasets which are the abstraction but cannot cover all the scenarios from real life. Directly applying such models to robotic products risks user experiences and safety, as the models might encounter unexpected interactions with users which will lead to unexpected and unstable system performance. Hence, it is unsurprising that although some robotic projects from academia received positive feedback from users [1], household users still repeatedly report problems in practical applications in daily life, (e.g., the failures from smart speakers) [2,3]. Even when some computational intelligence models shows impressive performances on some challenges in domestic robotics, robotic products on the market are still limited to the old-fashioned techniques due to various reasons from different perspectives. Therefore, a review of existing domestic robotic products from the following viewpoints is presented:

- to give an analysis about the functions of the current state-of-the-art domestic robots and their technologies behind;
- to identify the potential technologies in computational intelligence to be used in the domestic robots and its gap to the state-of-the-art CI models;
- to further foresee the development path of the domestic robotics.

The rest of the paper is organized as follows: Section 2 provides an overview of some examples and commercially available domestic robots, the taxonomy of domestic robots and the tasks and services they are able to provide. Section 3 introduces the computational techniques, or computing technologies in a broader sense, which are relevant to the domestic robotic platforms. We also compare on which levels of integration these techniques are for the domestic robot systems, and future directions to build better domestic robot products. Section 4 discusses the gap between the existing robot platforms and the state-of-the-art computational intelligence used in these representative scenarios. Section 5 discusses promising future directions, including trends and challenges in building next generation domestic robots. Section 6 summarizes the discussions and arguments of the paper.

## 2. Taxonomy of Domestic Robots

Figure 1 shows a few samples of the off-the-shelf commercial domestic robots. To give a full picture, the existing domestic robots and the corresponding taxonomy are summarized in Table 1. In general, domestic robots fall into two main categories: **virtual robots** (also known as *Software robots*, *Bots* or *Chatbots*) and **physical robots**, each of which can be further divided into several sub-categories. We introduce representative commercial products in each category, then discuss what assistance they can provide and how they communicate with human users to provide assistance, (e.g., physical assistance, cognitive assistance and/or social assistance) depending on how they interact with humans.

### *2.1. Virtual Robots*

As the name implies, **virtual robots** do not physically exist but it is used for interaction as a virtual but intelligent agent. They exist as a form of software installed on computing devices such as smart phones, computers and intelligent household devices (e.g., fridge). Well-known products in this category include *Siri* from Apple, *Alexa* from Amazon, *Xiaoice* from Microsoft and *Google Assistant*.

### *2.2. Physical Robots*

**Physical robots** are mechanical agents that exist physically. According to target applications and physical appearances, we categorize them into: **internet of things (IoT) robots**, **interactive robots** and **service robots**.

**Table 1.** The taxonomy and the software–hardware spectrum of domestic robots.

| Spectrum | Category | Examples | Applications | | |
|---|---|---|---|---|---|
| | | | **Physical Assistance** | **Social Assistance** | **Cognitive Assistance** |
| software | Virtual Robots | Google Assistant | ✗ | Assisting Scheduling Calling, etc. | Entertainment |
| | | Google Nest Hub | ✗ | Assisting Scheduling Calling, etc. | Entertainment |
| | | Amazon Echo | ✗ | Q & A Online ordering | Entertainment |
| | | Siri | ✗ | Q & A Online ordering | Entertainment |
| | | XiaoIce | ✗ | Q & A Online ordering | Entertainment |
| | IoT Robots | Nest Thermostat | Adjust heating | ✗ | ✗ |
| | | Samsung Hub Freezer | Watch food storage | Q&A Online ordering | TV Music |
| | | Wemo | Switch of electricity | ✗ | ✗ |
| | | Phyn Plus | Managing water level | ✗ | ✗ |
| | Interactive Robots | Pepper | Limited | Assistant Receptionist Healthcare | Conversing |
| | | Moxie | Q&A Education | ✗ | Entertainment |
| | | Paro [4] | ✗ | ✗ | Entertaining comforting elderly |
| | | Aibo | ✗ | Online ordering, etc | Entertainment, Respond to actions |
| | Service Robots | HSR [5] | Multiple actions | Very limited | ✗ |
| | | Stretch | Cleaning Manipulating Objects | ✗ | ✗ |
| hardware | | iRobot Roomba | Cleaning | ✗ | ✗ |

#### 2.2.1. IoT Robots

**IoT** is widely referred to as a network of interrelated smart devices (https://en.wikipedia.org/wiki/Internet_of_things) (accessed on 15 March 2021). In the commercial domestic robots market, the term of **IoT robots** also includes automatic devices in smart homes from thermostats and automatic lighting to smart household appliances such as smart freezers. They physically exist in the household environment, and can be controlled via computing devices such as smart phones or software installed on other devices.

Similar to virtual robots, existing IoT robots can retrieve information from the internet. Moreover, since they include sensor modules, they are also able to sense the surrounding environment and control other devices via internet connection. Hence, IoT robots are often integrated into home appliances such as the fridge, thermostat. Nevertheless, different

from the service robots we introduce later, IoT robots usually have little or limited mobility, depending on the sensors and motors they are equipped with.

Being equipped with certain sensors, IoT robots are able to perceive, communicate and act to a certain degree. For instance, cameras are installed in freezers to observe food conditions. Given the observation, an IoT robot informs users when food goes bad.


**Figure 1.** Samples of domestic robots: (from left) Pepper, iRobot Roomba, Nest Thermostat, Amazon Alexa, an NAO in front of Asimo.

#### 2.2.2. Interactive Robots

**Interactive robots** are often physical robots designed to interact with human users via conversations or other embodied signals with multi-modality interactions, such as speech, gestures and eye gazes.

Interactive robotic platforms such as NAO, Pepper, ELLIQ, Moxie, Buddy and Jibo have been commercially available and used in various occasions to give instruction, therapy or other interactive-based functions without actual physical intervention to the physical world. For example, Pepper and NAO robots have served as receptionists, health caregivers and airport guides. In general, those humanoid robots for interaction are with human-like appearances, so that they can generate human-like behaviors in interactions. The main modality in human-like behaviors is verbal conversation. Therefore, dialogue system is an important component in building competent interactive robots. Furthermore, when equipped with cameras and other sensors, the robots are also able to perceive human emotions, gestures and eye gazes from which they could recognize more complicated interactive signals.

Other products of interactive robots include ELLIQ, Vector and Jibo, which have human-like behaviors but do not have human appearance. Such kinds of design have been shown to give rise to a positive effect in gaze-following and attitudes in human–robot interaction [6]. In comparison, interactive robots, such as AIBO and Paro, appear as friendly robotic pets whose personality and behavior could evolve over time. These pet-like interactive robots are also equipped with sensors. For instance, the AIBO robot is equiped with facial recognition, smile detection and behavior learning modules.

Although the interactive robots do not have the ability to accomplish the physical tasks in domestic environment, targeting at a certain group of users, such as elderly and children, and equipped with educational or healthcare functions seem to be a trend to become a successful interactive robotic product.

#### 2.2.3. Service Robots

**Service robots** are often designed to assist users to complete physical tasks such as cleaning or mowing the lawn. Compared to the aforementioned robots, the hardware configuration of service robots is much more complicated, but it should be subjected to the pre-defined control strategies and behaviors for the particular service the robots should provide.

In general, the design of the perception and action modules are the key components of service robots. For instance, the ASIMO (https://en.wikipedia.org/wiki/ASIMO (accessed on 23 February 2021)) robot, is designed for domestic applications, where it needs to perceive the environment, act accordingly and complete one or a few domestic tasks. For instance, the vacuum robot is a single-function service robot which can only assist in finishing one task (i.e., vacuuming). As a well-developed product in the domestic domain, most vacuum robot products have obtained feedback from users [7]. As such, the only functions that we implement on the vacuum robots are usually obstacle avoidance and path planning.

Service robots nowadays are becoming better in their functions of perception and action, making it possible to complete several tasks such as vacuum cleaning and manipulating simple objects. Compared with single-function service robots, most multi-function service robots in the domestic environment are immature, thus less popular on the market. For example, Asimo was originally designed as a service robot. Although it is flexible in moving around (e.g., walking [8]), it has not been widely accepted by ordinary users due to a lack of communication and understanding abilities, as well as its high price. Other multi-function robot products such as PR2 and HSR face similar dilemmas for similar reasons. Yet, as pioneer models, they are well accepted by academic laboratories for research purposes.

#### 2.2.4. Boundaries between Categories

The previous sections summarize different categories of domestic robots based on which existences, either hardware or software, their main functions rely on. Practically, with the development of high-speed connections, the richness of an eco-system of a robot and the extensibility of both hardware and software, there are no clear boundaries for categorizing domestic robots. We summarized and discussed four sub-categories of the domestic robots: virtual robots, internet of things (IoT) robots, interactive robots and service robots. These four sub-categories form a continuous spectrum of software–hardware robots, within which a specific product may find a place.

From the perspective of users, they also expect a domestic robot could accomplish more and more tasks and become easier to use. Hence, a robot product located in the middle of the spectrum with extensibility in the eco-system, also makes domestic robots even more suitable for the demanding market and are becoming popular for households. For instance, the Pepper robot, targeted for social interaction, is also able to download different functions through the App market. With the possibility of integrating different IoT sensors, it can also obtain the sensory information at home. The Amazon Echo, although designed as a virtual conversation agent, can also connect and has control on some supported house appliances, including the smart thermostat. Nevertheless, in this overview, we examine and analyze the robot as a single product itself without considering the possibilities of connecting with other optional appliances. On the other hand, it also ignites the hope to build a multipurpose domestic robot which can centrally control all the household appliance and IoT devices as well as to do the service and interaction.

### *2.3. Multipurpose Domestic Robots and the Core Functions*

To build robots that are as intelligent, useful and convenient to human users is always the final goal for researchers, engineers and robotic companies. For instance, virtual robots can be integrated into IoT robots to control smart devices via conversations. Google Home can function as a hub of several smart devices. “Alexa”, designed by Amazon, is usually

embedded in the Amazon speaker Echo, but it can also control supported house appliances, such as the smart thermostat. With fast internet connection between home appliances, software robots can easily access the status of other household appliances.

Likewise, when equipped with software agents, physical robots can also provide social and cognitive assistance as virtual robots do. For example, NAO robots can provide users with embodied conversations. It can also chat with users to provide cognitive assistance. Moreover, it is able to deliver its bodily language, the conversation, as well as the screen display to have a better communication with the users. Hence, most of the robotic products lie in the middle of the continuous spectrum, making domestic robots serve multiple purposes.

Note that multi-purpose domestic robots are usually humanoid robots. For example, NAO (https://www.softbankrobotics.com/emea/en/nao (accessed on 1 January 2020)) and Pepper result in more active interactions with human users as they appear human-like, although the interactive NAOs are not designed for manipulation or other service tasks. HSR (https://www.toyota-global.com/innovation/partner_robot/robot/#link02 (accessed on 1 January 2020)) and PR2 both have most of their visual sensors on top of the torso, and the manipulators are assembled in the middle, which are suitable to perceive the environment and manipulate the objects with enough height as humans do.

Although we expect a large difference between the single-function and the multi-purpose robots, they have common built-in functions which require the tailored computational intelligent technologies. When we omit the design of the appearance, and take into consideration the internal computational intelligence functions that are built in the domestic robots, the following four core abilities could be summarized. Conventionally, a robotic system is composed of a **perception** module, an **action** module and a **control** module. Depending on the application scenarios, a domestic robot needs to perform some, if not all, of the following tasks: **perceiving, understanding, communicating and understanding** (Figure 2). Hence, different from the conventional designs of robot systems (e.g., industrial robots), efficient communicating is also one of the key abilities of domestic robots.

- **Perception:** A robot should be able to perceive its surrounding environment, including audio and vision. In addition to being able to recognize audio and vision signals, in the computational intelligence domain, these abilities correspond to automatic speech recognition, face detection and recognition, object detection, action recognition and emotion recognition.
- **Action:** After perceiving and understanding, a robot should be able to react accordingly, including vacuuming floor after receiving commands, making scheduled movements and doing multi-modal conversations including gestures. The action function can be further formulated into low-level actions, such as movements on wheels, median-level actions, such as pre-defined or adaptive behaviors and high-level actions, such as movements under planning.
- **Understanding:** In addition to perceiving, the robots should be able to understand what the signals mean. That is, to recognize what humans talk about via speech signals, to navigate itself via seeing the environment through cameras, to recognize human emotions and so on. These abilities correspond to natural language/commands understanding, scene understanding and so on.
- **Communication:** The communication function is one important element to make the robot "human-like". It can be used two-fold: firstly, the successful communication skills could be fundamental when the robots are involved in interacting with users, especially during the some tasks in the domestic domain can be too complicated to be finished with one or two commands. Therefore, it is often to converse with human users to achieve common ground understanding, especially the robots are involved in providing entertaining and cognitive services. Both verbal and non-verbal communications are often involved in such interactions.

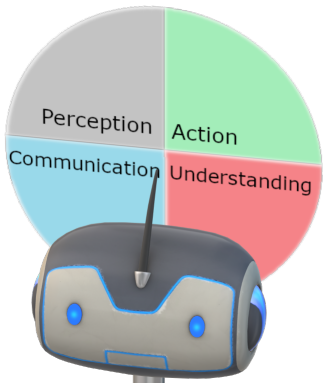


**Figure 2.** Four basic functions of domestic robots.

## 3. Computational Intelligence in Robotics

It can be concluded that action modules of the domestic robots are well-used, especially in terms of motion controls in navigation and other mobility function for domestic robots. Although some problems still exist in robotic navigation, most algorithms are well developed and integrated. On the other hand, there are also many challenges to be solved in their manipulation function in a more dynamic and unconstrained domestic environment than in the industrial one. Thus, in some service robots it is still difficult to achieve their multi-purpose use in terms of their inability to grasp different objects or to assist with other domestic tasks which require physical contact.

Perception modules in domestic robots, which usually include object and facial recognition functions, are quite useful in all of the four sub-categories. In addition, the state-of-the-art deep learning algorithms have been used in virtual and IoT robots which do not have high mobility. On the other hand, it is still unclear how to solve the problem with limited energy and computational power within limited time with the deep learning based methods.

Although "perception" and "action" functions are imperfect, we discover that the "understanding" and "interacting" functions are even more behind our expectation in most of the domestic robotic applications. For instance, the natural language understanding (NLU) that is used in the domestic robots does not follow the natural conversation manner of ordinary people (https://theconversation.com/ai-theres-a-reason-its-so-bad-at-conversation-103249 (accessed on 1 January 2020)), not mentioning that they are not able to learn new terms in the dynamic domestic environment with natural conversation with inexperienced users. In addition, due to the dynamics of the environment, individual differences between the users and other factors, the understanding and the interactive functions, such as emotion recognition and action recognition, are far from satisfaction. We conjecture that the main reason that the users are not fully convinced to purchase an interactive or service robot with a relatively high cost is that service robots and interactive robots are still not robust enough to understand and communicate efficiently and effectively with the users. As a result, users still consider the domestic robots as machines or toys. Table 2 summarizes the gap between the different types of domestic robots and the techniques. In the following subsections, the detailed analysis of CI techniques in these four areas for domestic robots will be reviewed.

**Table 2.** Four categories of robots and the related techniques. ✗ indicates a technique is inessential; → indicates that the current technique implemented in this category of robots is state-of-the-art and performs well; ↗ indicates there is a large gap between state-of-the-art techniques and available engineering robotic applications.

| | Virtual Robots | IoT Robots | Interactive Robots | Service Robots |
|---|---|---|---|---|
| SLAM and Navigation | ✗ | ✗ | → | → |
| Object Recognition | → | → | ↗ | ↗ |
| Facial recognition | → | ↗ | ↗ | ↗ |
| Action Recognition | ✗ | ↗ | ↗ | ↗ |
| Emotion Recognition | ✗ | ↗ | ↗↗ | ✗ |
| Speech Recognition | → | → | → | → |
| Dialog System | → | → | ↗ | ↗ |

### *3.1. Perception: Speech Recognition, Face Detection and Recognition, Object Detection*

Robots perceive the surrounding environment through sensors integrated on themselves. **Cameras**, **infrared detectors**, **haptics** and **microphones** are some of the important sensors that could be used in existing robotic products. Therefore, existing domestic robots mainly perceive the world via images, video, audio and haptics. That is, domestic robots hear, see and feel the surrounding environment via these signals. Consequentially, to make use of these signals detected from sensors, computational models of speech recognition, object detection, face detection and recognition are most related to enable robots to perceive the surrounding environment.

Next, we discuss state-of-the-art of computational models regarding automatic speech recognition (ASR), object detection, face detection and recognition.

#### 3.1.1. Automatic Speech Recognition (ASR)

Automatic speech recognition models enable users to communicate with robots via speech, which is one of the most natural ways to communicate in our daily life. By transferring raw audio signal of spoken language into readable texts, ASR models could provide input as the understanding of the natural language to the dialogue systems (see Section 4.1), enabling higher level interactions such as conversations.

Taking into consideration over 50 years of research and development, mature ASR products are now available on the market. Google ASR (https://cloud.google.com/speech-to-text (accessed on 1 Jaunary 2007)), CMUSphinx (https://cmusphinx.github.io/ (accessed on 23 October 2019)) and IBM Watson speech to text (https://www.ibm.com/cloud/watson-speech-to-text (accessed on 1 Jaunary 2020)) are among the most popular ones. Based on deep neural network models [9], these products achieve good performances in various applications.

Deep neural networks have been the de facto architecture for automatic speech recognition (ASR) tasks. Modern commercial ASR systems from Google and Apple (as of 2017) are deployed on the cloud and any products could access it via a network connection. While a traditional ASR system is typically composed of five parts: acoustic analysis for feature extraction, acoustic model, language model, pronunciation dictionary and the recognizer for recognition, deep neural network-based systems often adopt an end-to-end structure, which means that a single network can generate the estimated results from the inputs. With well-trained models, they outperform previous models such as Gaussian Mixture Model (GMM) (e.g., [10]) and Hidden Markov Model (HMM) (e.g., [11]) by a large margin. Specifically, the deep neural models usually adopt the CNN (e.g., [12,13]), RNN (e.g., LSTM [14] (Figure 3), Highway-LSTM [15]), attention mechanism (e.g., [16]), encoder-decoder architectures (e.g., [17]). The idea of deep learning methods is to learn the acoustic features as time-series without feature extraction, which can be memorized on different levels of the neurons (e.g., LSTM, Figure 3).

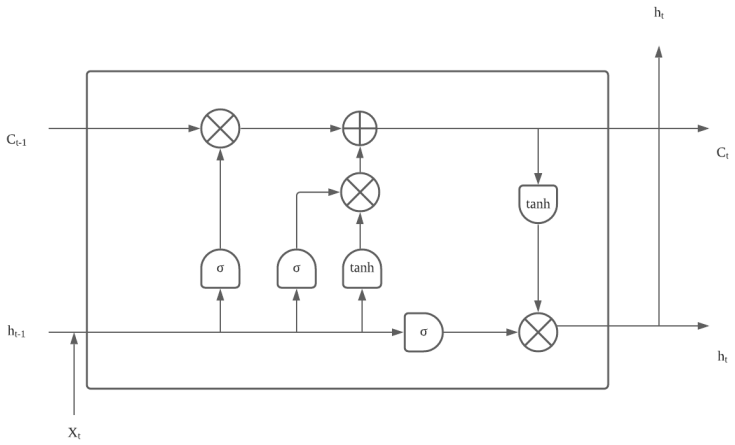


**Figure 3.** The LSTM unit.

#### 3.1.2. Object Detection and Recognition

Object detection models enable robots to recognize physical objects in the environment. Such models take images/video as input, and they output the labels of the learned objects. As one of the active topics in the computer vision and machine learning, recent years have seen significant advancement in object detection algorithms in different applications.

Basically, the object recognition methods need to discriminate the features in the closed training data. Such features can be extracted either in a data-driven way or pre-processed way. Before the data-driven convolutional neural network (CNN)-based methods achieved a great success, most of the classical methods for object detection should first manually extract features which represent the points of interests, either specific objects or human faces. For instance, the scale-invariant feature transform (SIFT) [18] methods use points of interests on the object which are independent of the size and rotation of the image. In addition to collecting the important features and their surrounding information like SIFT, the Viola-Jones object detection framework [19] seeks simple Haar features to detect object features. Such computer vision methods usually need to extract the features, either manually or automatically by evaluating their functions of features, before the recognition stage is processed.

In comparison, the deep learning-based methods (e.g., CNN-based methods) skip the procedure of defining features. Instead, it adopts an end-to-end learning strategy, in which the perceived images or videos can be used as the input of the deep neural networks. These methods are with faster speeds, lower resource consumption and ultimately outperform classical methods on the task of real-time object detection.

When implemented on a robot, the object detection is the first stage of a robotic system to detect if an object exist in its sensors' receptive field. Most of the state-of-the-art object detection is based on the deep CNN methods. These methods can be divided into the two-stage methods and the one-stage methods. They can be distinguished by seeing whether the selective search stage for the objects is required before the object recognition is processed. The selective-search stage in the object detection methods means that it includes the algorithms that will generate a small number of segmentation areas within an image (e.g., bounding boxes) that may belong to one object, based on the colors, textures, sizes and shapes. Moreover, such a process runs in an iterative way in order to combine similar regions to ultimately form the objects. For instance, the R-CNN [20] and Fast R-CNN [21] use such search methods [22] and the Faster R-CNN [23] uses the Region of Interest (RoI) pooling layer to find the bounding boxes with various aspect ratios and sizes for the possible object appearances. After the selective search stage, the two-stage object classification methods use deep learning methods to recognize the objects within bounding regions.

Comparatively, the one-stage methods usually perform faster than the two-stage methods, due to the fact that the search and classification tasks are done with the same network. This is achieved by the pre-trained rich features of this network which can form the bounding-box descriptors. These descriptors are used to recognize the object identities at the detection stage. The most widely used one-stage methods are YOLO [24,25] and SSD [26]. The SSD uses a number of feature pyramids with fixed sizes in its decision stage, while the older version of YOLO only uses a fixed-size of feature detector and a number of predictors to detect objects. The one-stage methods often achieve lower accuracy than the two-stage methods. Nevertheless, the one-stage methods are usually employed in devices with strict requirements in processing speed and low-power.

In terms of the applications of the detection methods, the state-of-the-art object detection methods have achieved satisfactory performances in their accuracy, computational cost and processing speed. Both the one-stage and the two-stage methods have been widely integrated in the perception part of the domestic robotic systems (e.g., [27,28]), such as autonomous navigation [29], pedestrian detection [30], manipulation [31] and other robotic applications [32].

#### 3.1.3. Face Detection and Recognition

In general, the tasks of object recognition and facial recognition are different but relevant: both of them aim to recognize objects/faces as an identity in videos or images. The only difference for facial recognition is that the facial features, and their relative placement on the faces should be extracted and understood in an optimized and fine-tuned representation as an open dataset learned in the model, comparing to the object recognition and detection. Compared to object recognition tasks, to recognize a specific person's face, facial recognition is more challenging because of the following reasons:

- in most cases, a person to be detected may not be included in the training set, while most of the objects recognized are within the training set, which we state the facial recognition is an open-set problem.
- since the labelled objects already exist in the training datasets, we usually employ a discriminative model which produces the labels as outputs given the input signals to solve the object recognition problem. This can be achieved by incorporating a discriminative classifier (e.g., soft-max) at the end of the model. In the case of facial detection, since the output labels of facial detection are only binary, the features can either be pre-defined or data-driven.

Furthermore, while in practical use, an important requirement for the facial detection/recognition is to perform the model "in-the-wild", which means that in a dynamic and noisy real-world environment, the human faces can be highly varied. Thus, it is suggested that facial detection is one of the most challenging tasks in computer vision (for more comprehensive reviews, see [33,34]).

Nevertheless, when the data are labeled and the model is learned, facial recognition nowadays using deep learning methods also achieves impressive performance which is on par with human performance. The facial recognition is often evaluated using public standard datasets such as Labeled Faces in the Wild (LFW ) [35], CelebA [36] and MS Celeb (https://github.com/PINTOFSTU/C-MS-Celeb (accessed date 9 October 2020)). As such, the deep learning methods for facial recognition (e.g., DeepFace [37], DeepID [38] and RingLoss [39]) usually use mainstream CNN networks (e.g., AlexNet [40], VGGNet [41], ResNet [42]) to do the recognition task while they also employ assembled networks to match the variances about inputs to solve the "in-the-wild" problem.

In the computer vision (CV) community, higher accuracy of facial recognition methods are usually achieved by improving the CNN architectures (for a comprehensive review, see also [34]). In the field of service robotics, compared with the standardized facial recognition tasks, the facial recognition tasks have additional challenges:

- The state-of-art facial recognition methods are evaluated with the standard datasets. In the practical robotic applications, the perceived information usually contains a lot

of noise and environmental factors vary, which results in the negative effects in the recognition accuracy.
- Most of the facial recognition methods embedded in robotic systems require the users to stand in the receptive fields of the camera (assuming they are also at a reasonable distance from the camera), which might be inconvenient/unfriendly feeling to the users.
- The usages of other sensors, such as RGB-D cameras (e.g., [43,44]), and voice recognition have not been fully used for person identification to help to solve the in-the-wild problem.

### *3.2. Action: SLAM, Obstacle Avoidance, and Path Planning*

Upon perceiving and understanding commands from human users, a domestic robot should react with certain movements to accomplish assigned tasks. In this subsection, a few techniques for controlling the robot to act will be reviewed and discussed. Specifically, we will review the robotic movements (simultaneous localization and mapping (SLAM), navigation) and robotic manipulation.

#### 3.2.1. SLAM

When completing physical tasks, being able to move around in the domestic environment is critical for domestic robots to accomplish some tasks, such as vacuuming. Hence, navigating, simultaneous localization and mapping (SLAM) and path planning are usually basic requirements for domestic robots.

The SLAM algorithms enable a robot to update the map while tracking its locations in an unknown environment. Due to noises and uncertainties from the sensors (perception), the SLAM tasks are often formulated as a probabilistic problem [45,46], which can further be solved by the estimation problems (as shown in Figure 4). SLAM can be theoretically solved by mature algorithms based on statistical estimation methods such as Kalman Filters and Particle Filters [47,48]. Nowadays, most of its existing problems are within the engineering domain, particularly the integration and optimization of different stages of SLAM, such as the data association and dynamic environment problems.

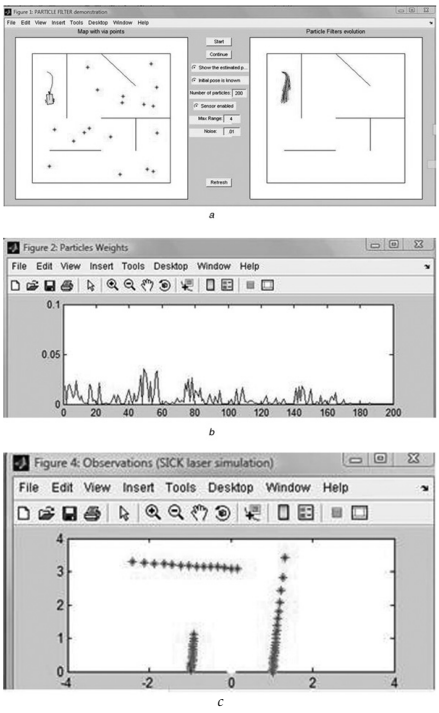


**Figure 4.** An example of particle filtering simultaneous localization and mapping (SLAM) in simulation [49]. (**a**) The true position and the estimated particles of a simulated robot in a maze. (**b**) The weights of distribution of particles. (**c**) The sensor information from a laser sensor.

For instance, one practical problem to be considered in the domestic robotic platforms is the selection and integration of sensors in SLAM. It is a trade-off of usage of expensive sensors and accuracy in SLAM on a robotic system: Conventional SLAM methods used in unmanned vehicles [50] usually require a number of expensive equipment. For instance, the LiDAR sensors are usually necessary in order to provide accurate measurement of the obstacles. They are used in robots such as the PR2 robot [51] and the modified version of NAO robots [52]. On the contrary, the price-tag of the laser sensor as well as its noise, which affects the family members also, become the factors to prevent the usage of it on domestic robot. Therefore, as an alternative method, the SLAM used on the domestic robots are also considered using the low-cost depth camera-based sensors [53]. Alternatively, the camera-based SLAM (e.g., ORB-SLAM [54]) needs more computation and results in less accurate results since they should do the feature matching together with mapping and loop-closure.

They are still enough for some of the domestic products. For example, the vacuum cleaning robot iRobot Roomba 980 [55] is based on vSLAM [56]. Several algorithms have been developed for RGB-D camera-based SLAM as well, such as KinectFusion [57], Kintinuous [58], DynamicFusion [59] and so on. Moreover, the SLAM algorithms employ only vision sensors, such as monocular cameras, and binocular cameras or fish-eye cameras also emerge since the costs of the visual sensors are relatively low. Another advantage of these algorithms is that the results of semantic SLAM can be further used for object recognition and related functions.

#### 3.2.2. Obstacle Avoidance

Based on the results from SLAM, domestic robots still need two sub-functions: obstacle avoidance and path planning to navigate.

Given the local environment map, the obstacle avoidance function endows robots to avoid objects on the way to the target. When the robot has no map, the method of approaching the target is called visual obstacle avoidance technology. The problem that the obstacle avoidance algorithm solves is to avoid static obstacles and dynamic obstacles based on the data of the visual sensor, but still maintain the movement to the target direction and realize real-time autonomous navigation. In academic research, the object avoidance function has also been relatively well-studied.

It is integrated with object recognition, object following or even with fast-moving objects. In applications, since the shapes, colors, sizes and textures of the objects are different, reasonable deployment of different sensors (e.g., ultrasound, tactile and infrared sensors) should be carefully considered depending on the working environment (e.g., [60,61]).

As a well-developed function, there are many obstacle avoidance algorithms, but these methods have strict assumptions. For example, the Virtual Force Field (VFF) algorithm [62], which assumes that the robot is a point and can move in any direction. The Vector Field Histogram (VFH) [63] assumes that the robot is circular and expands through circular obstacles. When considering kinematics, it only assumes that the robot moves in a circular path. However, in practical applications, it is difficult for robots to meet the conditions perfectly which may result in the inaccuracy in obstacle avoidance.

To solve the obstacle avoidance problem in the dynamical world, some of dynamic obstacle avoidance methods have also been developed and used in domestic robots where, in some complicated real scenarios, they should face a lot of moving objects surrounding. For instance, to finish a delivery in a dynamic environment [64].

#### 3.2.3. Path Planning

In contrast to an object avoidance function that reacts immediately, the path planning function aims at finding a reasonably good (but not always most optimal) path to plan ahead and move from one location to another, together with the consideration to minimize time and energy. Using the map and the location information obtained from SLAM, the path planning algorithm outputs appropriate action commands that lead to the desired

target location. According to the environment, there are two categories of path planning algorithms: the complete and the sampling-based approaches. A path planning method is said to be complete if the algorithm can produce a solution or report there exists no solution in finite time. Most of the complete algorithms are geometry-based, which means that the representation of the map is geometric. These algorithms are usually planned directly on the occupant grid map in the field of mobile robots (i.e., a matrix composed by the pixels) with the search methods, the Dijkstra algorithm [65], the A* algorithm [66] and so on. In particular, the Dynamical A* (called D*) algorithm [67,68] has been widely used for mobile robots and autonomous navigation systems and can solve the path planning in a dynamical environment. It also has been proved to be successful in the Mars rovers and the CMU team who participated the DARPA urban challenge. It maintains a cost value, and the search grows outward from the goal by a process called "propagation process". The nodes are being expanded and are denoted the exact cost by the propagation process.

The representative methods of sampling-based algorithms are RRT and PRM algorithms. Since, in some scenarios, high-complexity and high-dimensional space exist in the complete path-planning algorithm, the random sampling algorithms [69] can effectively solve this problem by drawing random samples to form a graph (PRM) [70] or a tree (RRT) [71] connecting the start and the end points. Although the sampling-based algorithms are computationally efficient, due to its non-consistence in obtaining the results with the randomly generated instances, it is still difficult to apply them in commercial or industrial applications where the robustness should be taken into consideration. The A* and D* algorithms are still widely used in the scenarios in which path planning functions are needed. For instance, the D* algorithm has been embedded as part of the navigation module (e.g., SLAMWare (https://www.slamtec.com/en/Slamware) (accessed on 1 Jaunary 2020)) and widely used in many domestic robot products which require mobility.

#### 3.2.4. Robotic Action, Behaviors and Their Selections

Traditionally, industrial robots are required to repeat specific actions with relatively high precision. In these scenarios, the motor actions of robots are clearly defined in an industrial environment, where the position, velocity, acceleration and force variables can be precisely calculated by kinematics and forward/inverse dynamics [71]. Based on control theories, they can already be solved mathematically, which belong to the white box systems. On the other hand, while the environments become dynamic, the uncertainty becomes high. The effect of unpredictable perception, including the noise from sensors, to action becomes larger. To deal with the uncertainty of environment has always been a difficult problem for conventional automatic control solutions.

In the dynamical environment, such as domestic households, to react with the noisy perception has always been a difficult problem for traditional control methods. Alternatively, the "robotic behaviors" is often used which initiated a novel robot action and control approach endowing adaptivity and robustness in the dynamic environment.

A classical framework to design the robotic behaviors is the subsumption architecture [72,73], which coordinates the execution of different behaviors. Each behavioral module may be designed by the finite-state machines. As shown in Figure 5, the behavior-based inclusive frame structure divides the robot into several behavioral modules in a vertical manner. The module at the higher position will have the higher priority if two modules conflict. As a result that the designer can structure and add modules of different priorities, it is possible to construct a relatively complex system with this structure.

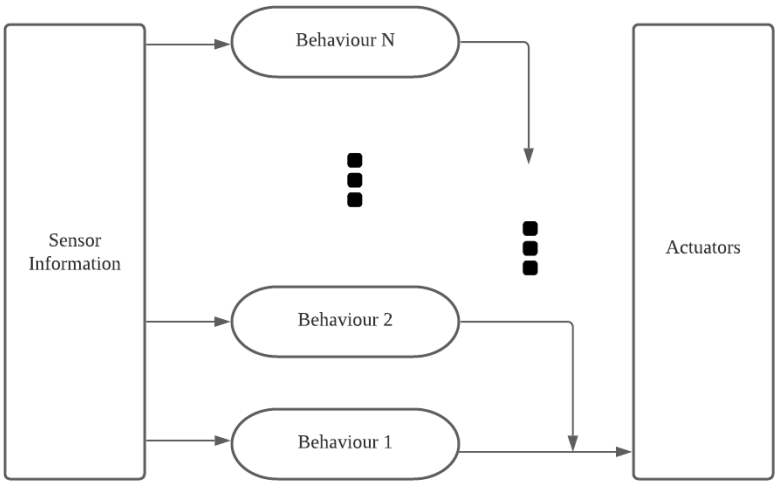


**Figure 5.** The subsumption architecture.

Based on the building blocks of different behaviors, a network-based behavior selection framework can be also used [74,75] to select the most appropriate actions. As an alternative method to the behavior-robotics, the motor scheme [76] is also embedded in architecture to allow the action-centric perception. Therefore, within the structure, each action can be seen as the result of all the perceptual schema using the summation of the potential fields.

Fuzzy logic deals with uncertainty subjectively to integrate different selections from behavioral modules [77–79]. It designs a corresponding sub-fuzzy inference system for each behavior, thereby avoiding the negative impact of the large fuzzy rule base on the real-time performance of the system.

### *3.3. Understanding: Action, Intention and Emotion*

Building the understanding ability in domestic robots usually means that the robot is able to detect or infer the intention, emotion or even thoughts through observations. Such observations rely on cues from the previous or current expressions in behaviors or subtle actions from different parts of the user, such as body, gesture or face. Although electroencephalography (EEG) sensors for brain–computer interface (BCI) driven robots have been found in the research labs [80,81], they are still far from practical use because most of them need calibration and training.

#### 3.3.1. Action Recognition

In the CV community, the action recognition task is also an active topic which aims to identify various actions which are performed throughout or during part of the entire duration based on video sensors. Therefore, the requirements of recognizing the well-trimmed videos and untrimmed videos are different. For instance, there are five different tasks in the well-known action recognition challenge ActivityNet [82], in which the recognition methods for trimmed and untrimmed videos can be quite different. Based on camera signals, the improved Dense Trajectories (iDT) [83,84] method is a conventional method before deep learning became popular. The iDT framework includes dense sampling feature points, feature point tracking and trajectory description. In this framework, the feature points are selected in different spatial scales, which can be further used to obtain three different features to describe the dense optical flows. Once obtained, these bags of features can be further used to do the classification of the actions. However, nowadays, the deep learning-based action recognition methods achieve much better benchmark results. The two popular deep learning methods for action recognition are based on either Two-Stream Convolutional Networks [85] (TSN) or 3D-convNet (C3D) [86] ideas. The TSN method

is inspired by the findings in visual systems, where the ventral and dorsal pathways are separate to deal with the visual information for perception and action [87]. In the TSN model, the two-stream idea is used as two independent recognition streams (spatial and temporal streams). The spatial stream performs motion recognition from still video frames. The temporal stream recognizes behavior from dense motion optical flows. Both of the two streams can be implemented by CNN. It has also been reported that the separation of the temporal stream and the spatial stream improves the generalization ability of the network. Different from the TSN, the C3D-related work uses a different idea, which describes a cube-like convolution kernel to be an effective video descriptor in three-dimension for the temporal actions in videos. Specifically, it is also suggested that the best size of the convolution kernel empirically should be $3 \times 3 \times 3$ . The C3D method is easy to comprehensive because it owns the same principle of general 2D-CNN.

In addition to the action recognition based on visual information, some other sensors could also be used to do action recognition where the visual information could not be fully utilized due to privacy-related issues. In these scenarios, some ubiquitous sensing techniques such as the accelerometer [88], inertial sensors [89], microphones [90] or the combination of more than two modalities [91,92] can also be used for action/activity recognition, especially when the privacy of the users is taken into consideration.

#### 3.3.2. Emotion Recognition

Emotion recognition is also a useful tool when a domestic robot is engaged with the social interaction with humans. During the human–robot interaction, much information about the person can be understood if the robot is able to "read" the emotion. In general, the emotion status can be recognized from facial and bodily features. Facial expression is thus an essential way for the robots to identify the emotion information. Similar as most of the techniques mentioned above, with the great performance of various deep neural networks, most of the facial emotion recognition methods are using deep learning methods. The facial emotion recognition methods usually include a series of pre-processing techniques, such as face-alignment, face normalization, some may need data augmentation to make the method more robust. Then, the main recognition techniques can be implemented by the CNN architectures (e.g., [40,41,93]) plus some fine-tuning tricks. For instance, [94] considers the temporal relation along continuous frames, with which it also focuses some of the peak high-intensity expression and ignores the other with lower intensity in expressions. The main architecture is adapted from GoogleNet [93]. Similar to other specialized techniques of the CV communities, most of the state-of-the-art methods aim at achieving good rankings for specific datasets. In reality, implementation of such techniques should also consider the practical issues such as noise in sensors and the dynamic environment as the facial detection/recognition tasks.

In addition to recognizing emotion through the facial expression, it is also worth mentioning that other forms of expression (e.g., the bodily behavior [95–97] (e.g., Figure 6), voice [98] and conversation [99]) can also be adopted as cues for emotion recognition). It is natural that using multi-modal cues will increase the accuracy of emotion recognition. In this case, using other modalities to do emotion recognition is useful and efficient. Additionally, it is also brain-inspired that the neuroscience finding that our brain recognizes emotion primarily based on the facial expression, but in the mean while, probably because these cues are not as obvious as the face when our brain is recognizing emotions [100]. Therefore, various datasets and competitions for multi-modal emotion recognition are announced [101,102] and it would probably become the next function to be implemented for human–robot interaction. Still, none of the open multi-modal datasets have achieved the popularity of uni-modal datasets for emotion recognition, due to the individual difference among different people and the size of the recorded datasets. Even though the model is well-trained with the datasets, it is still a challenge for them to recognize emotion robustly in the wild [103], not mentioning the existence of noise and sensitivity of the sensors.

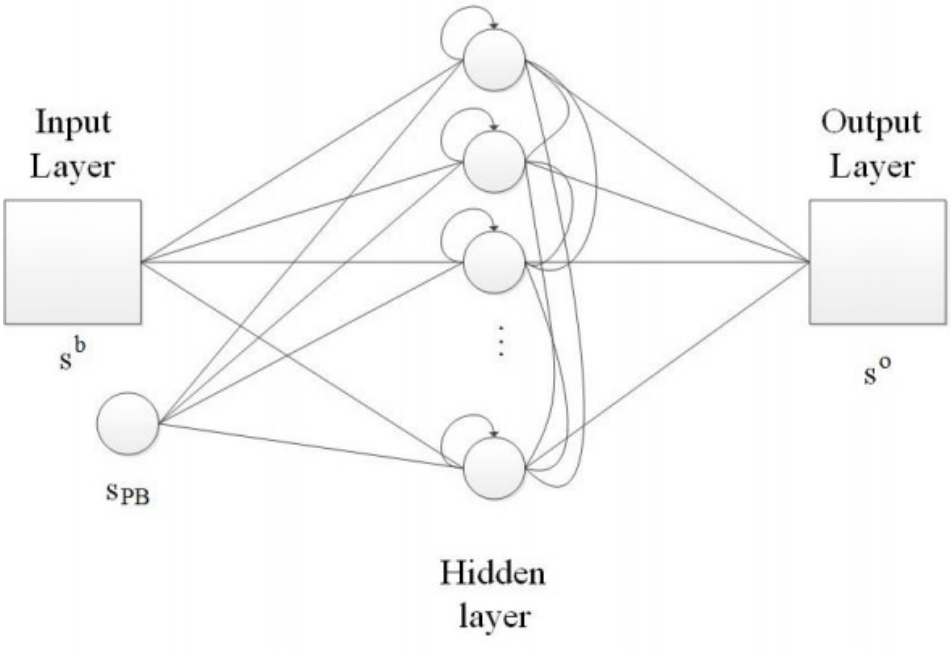


**Figure 6.** An emotion recognition deep-learning based on Kinect [95].

To conclude the techniques for emotion recognition and their implementation on domestic robots, the main challenges for emotion recognition are still open, especially in the human–robot interaction in domestic robots. There are mainly three problems remaining to be solved:

1. Computing-wise: How to efficiently utilize the multi-modal sensor signals to identify the emotional status of users? To solve this, we may consider the multi-modal machine learning methods [104].
2. Interaction-wise: During usual interaction, all the social contexts (e.g., wording in the conversation) and common knowledge can also be considered as cues for recognizing emotion. Can we also utilize such knowledge on a robot?
3. Robot-wise: What kind of sensor signals can we choose to jointly estimate the emotion by reducing the noise and placing them together to work robustly?

### *3.4. Communication: Speech, Dialog and Conversation*

Since the beginning of AI research, researchers have aspired to enable robots to communicate as humans do. That is, communicating with natural language and gestures through conversations. This requires robots to go beyond command understanding. To converse with humans, robots need to understand natural language from humans, build a common ground of understanding and express themselves.

#### 3.4.1. Generation: Speech Synthesis, Inverse Kinematics

As domestic robots are expected to interact with inexperienced users, it is natural to communicate a robots' intention in a natural way as humans do. Therefore, natural language generation has been one of the hot topics in computational linguistics. Speech synthesis software such as text-to-speech (TTS) (https://www.ibm.com/cloud/watson-text-to-speech (accessed on 1 Jaunary 2020)) has been made available as a service. However, the most challenging part is to generate texts in given interaction scenarios. Recently, end-to-end dialogue systems have gained more and more attention [105,106]. Trained with large scale datasets, given an input sentence, these systems generate an output sentence as a reply. Other works have also explored stylistic dialogue generation [107,108]. These systems aim to generate dialogue responses in a certain style such as optimistic and pessimistic.

Therefore, the systems not only generate a response, but conveys certain personalities through the generated texts.

Human–robot interaction not only goes beyond commands, it also goes beyond natural language. Human communication is multi-modal in nature. In addition to natural language, humans often deploy other modalities to vividly convey information [109]. For example, to indicate an object in the situated environment, pointing gestures are often used for a single target object [110]. Researchers have also worked on multi-modal dialogue/virtual agents to enable robots to communicate with speech and gestures [111,112]. Gesture realization in robots is referred to as inverse kinematics. Given a trajectory that a robot wants to gesture, an inverse kinematics algorithm computes all the joint positions that need to realize the gesture, then enables a robot to make the movement. Combing natural language generation and gesture generation brings robots closer to human-like agents.

#### 3.4.2. Language Models and Language Understanding

Language models providing the basis for speech recognition, machine translation, part-of-speech tagging, parsing, optical character recognition and other applications for domestic robots, are the core in their communication function. They are usually statistical models which describe the occurrence of words/phrases in the sequences of corpus. One of the easiest frameworks to form the sequences is to extract $n$ consecutive words, and then to learn them as an $(n-1)$-order Markov Model, where we can predict the next item or the missing items when the $(n-1)$ items are given.

With the extension of this idea in which we can regard the consecutive words as the Markov model, the neural network-based models are recently used widely. One of the great advantages of it is to alleviate the problem of "curse of dimensionality". Ref. [113] used continuous representations or embeddings of words (Figure 7) to make their predictions based on the semantics relations of the words. Usually in the NLP tasks, the models use a large amount of corpus (e.g., the Wikipedia text) to learn those relationships. After proper training, the representation of the trained vectors can be regarded as the word analogy (e.g., "man + king − woman = queen"). The embedding layers are often used for pre-processing instead of the one-hot representation.

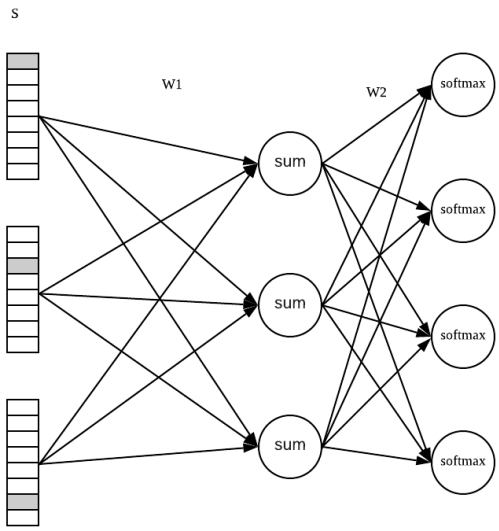


**Figure 7.** An example of embedding: Word2Vec, where the layer in the middle can be used as an embedding representation [113].

In the state-of-the-art language processing tasks based on representation learning, RNN (including LSTM [114]) or the Attention [115,116] mechanisms are the most widely

used. Both of them attempt to solve various language-based tasks such as translation, automatic summarization, relationship extraction, sentiment analysis, speech recognition, topic segmentation, etc.

When implemented on the robotic systems, especially on the categories of interactive robots which emphasize the function of interaction including language, only the language models and the aforementioned language-based tasks will not be enough. When the users can observe various items and scenarios in the domestic household, they imagine the counterparts during interactions will see the similar things, which will be further involved in the language-related tasks. Thereby, the natural language understanding (NLU) is a further step after NLP.

The NLU was proposed in the applications of human–computer interaction. Compared with NLP, it can play a better role in automated reasoning, question answering and large-scale content analysis. While the users do not follow the formal syntax while having an interaction, the NLU could understand the natural human languages. The most common systems of NLU included SHRDLU [117] in the 1970s . Recent applications such as IBM Watsons (https://www.ibm.com/cloud/watson-natural-language-understanding (accessed on 1 January 2020)). Nevertheless, the understanding requirement seems to be far behind the development of other NLP tasks. On the other hand, the current domestic robots even need a higher standard for "understanding": the users may refer to any items in the environment without a formal syntax, while the conventional NLU methods only use semantic parsers to divide the sentences into units and thereby narrowing the scopes of the words. In the domestic robotic applications, we may need a robot to "understand" what the users are talking about by seeing and sensing the same with the users.

#### 3.4.3. Dialogue Systems

Existing dialogue systems fall into two categories: (1) task-oriented dialogue systems, and (2) open-domain dialogue system (also known as chatbots).

Task-oriented dialogue systems aim to assist users to accomplish a certain task. These systems are quite often rule based. They dominate the dialogues and fill a set of slots to obtain useful information related to the task. Prominent examples of such systems are ticket booking systems, and automatic answering machines. To successfully book a ticket for a user, a system usually starts by inquiring a users personal and route information.

Open-domain dialogues do not constrain the topics of a conversation. They often serve as chatbots to chitchat with users. In this case, the system does not dominate the conversation, but tries to generate a proper response when users interact. Therefore, an open-domain dialogue system requires broad world knowledge, user intention understanding, dialogue state tracking and other techniques to keep a meaningful dialogue. Deploying such dialogues can be harmful. For instance, XiaoIce from Microsoft had complaints for discrimination and abusive responses.

The choices between task-oriented system and open-domain dialogue system depends on application scenarios. For simple physical assistance, pressing buttons, using a graphical user interface (GUI) or a simple phrase command is sufficient. For more complex tasks such as cognitive assistance (e.g., reduce the symptoms of Alzheimer's via chatting), more natural and human-like communication skills such as interactiveness in dialogues [118] and incrementality in situated dialogue systems [119,120] are essential. To achieve more natural communication, especially in complex tasks such as understanding or giving navigation descriptions [121–123], incorporating hand gestures would be beneficial [109].

## 4. Domestic Robots in Real Life: Where We Can Fill the Gaps

In this section, by examining how the domestic robots perform in several typical real-life scenarios, we will address the gaps between ideal applications and the available computational techniques. Additionally, we will point out the state-of-the-art techniques that are available to improve corresponding performances.

### *4.1. Conversational System*

**Scenario: The user who is in immobility wants to ask the robot to bring something.**
User: Hi robot, I'm hungry.
Robot: I remember you like apples. They are on the table over there.
User: I cannot really see clearly. Could you please hand me the one which is ready to eat?
The robot moves and picks up the apple.

- **The long-term memory and learning:** For the users, the long-term memory of a robot is essential for them to feel like the robot is a continuous being which is co-living with them. Social robots also need long-term memories in order to keep the knowledge acquired from learning to establish long-term relationships with humans and other robots.
  Since the learning world is open and the users have individual differences, household objects and household tasks as well as the human's behaviors differ. In order to endow the long-term memory, the ability of continuous learning [124], active learning [125] or learning via human-in-the-loop [126] could be implemented on the robots.
- **Multi-modal language processing, understanding and grounding:** As we have already discussed, the state-of-the-art language processing is the first step for a robot to possess languages. At present, various data-driven conversation systems have been proposed based on reinforcement learning [127], Attention [128] or the hybrid model of various techniques. If the methods could integrate the continuous learning, they could be possible methods which could avoid the curse of dimensionality in reinforcement learning and result in an open-ended training conversation system. Furthermore, we also anticipate that the language understanding ability of robots is achieved after the language grounding problem for robots, which we will discuss in the next Section 5.1.3 .
- **Communitive gestures:** As we have mentioned, humans naturally communicate with speech and gestures [109]. Despite the continuous effort on building multi-modal interfaces, current robots can only understand a limited and pre-defined set of gestures from humans, which mostly fall into the category of symbolic gestures. However, most common gestures in daily communication are iconic gestures (e.g., drawing in the space to describe the shape of a stone), which have no particular form and bear close semantic and temporal relation to accompanied speech [129,130]. Without understanding iconic gestures, robots rely on natural language to understand humans. Hence, users must articulate themselves via language, making it less convenient and natural than interacting with humans.
- Other techniques involved **object tracking and recognition and object manipulation**.

### *4.2. Affective Communication*

**Scenario: The user is having a conversation with the robot.**
User: I don't think I know this person in the picture.
Robot: This is your grand-daughter Amy. She just saw you last week.
User: Oh fantastic. I even don't remember my family.
Robot: Don't say that. They love you.

- **Affective computing:** Affective computing [131] is a broader field of emotion recognition, which also includes interpret, process and simulate human affects. Personal or domestic robots are nature embedding platforms to implement and test affective computing models since they have human-like appearances. Various robotic platforms,

for instance, Kismet [132,133], have been used to test the affective models as well as tools for human–robot interaction.
The affective computing is still an emerging subject, and its theoretical foundation in cognitive sciences is still open to discussion [134]. Most of its applications used in robotic systems have focused on emotion recognition and interpretation based on speech [135], facial [136] and bodily expression [95]. The simulation of emotion and the its synergy [137] to bodily expression, speech and facial expression.

- **Artificial empathy:** Robots built with artificial empathy are able to detect and respond to human emotions in an empathetic way. The constructing of empathy on a robot should also be included in the affective computing. The level of empathy may be calculated by the theory of simulation [138] in empathy.
  Although these can be also rooted from the emotion recognition techniques, various other cognitive theories—inspired artificially built empathy theories—have be also developed. For instance, from the developmental point of view, the empathy of robots can be developed by the common embodiment to achieve [139], such as artificial pain [140] obtained from the tactile sensors.
- Other techniques involved **language understanding and facial Recognition**.

### *4.3. IoT Robots*

**Scenario: The user who is in immobility wants to ask the robot to bring something.**
User: Hi robot, can you cook a dinner for me tonight?
Robot: Sure. I'll recommend you to have a caesar salad and tomato soup. It's good for you to keep your weight according to your dinner plan.
User: It's okay. Please go ahead.
Robot: Okay. But it looks like we have do some shopping because there is no chicken left in the fridge.

Specifically, some problems are worth investigating:

- **Interconnected with other robots and the internet:** It seems not difficult for a robot to connect with other devices via internet. With the connections, there are still open questions such as the accessibility of different devices, the trade-off between efficiency in interaction and the completeness of information searching.
- **Personalized recommender system:** The commercial recommender system usually uses collaborative filtering which collects information or patterns by the collaboration among multiple agents, viewpoints, data sources, etc. This technique is particularly useful for the recommendation of commercial products or common interests among different groups of people. Nevertheless, when the recommendation is about something not common among different people, for instance, a user A at home likes eating fish, while another user B in another home may follow a diet. This recommender problem should be addressed with other algorithms.
- Other techniques involved **robotic behaviors, e.g., cooking.**

**Scenario: The user is in the room.**
User: Hi robot, this room is too bright.
Robot: I can ask the curtain to close a bit. But the doctor says you need to have a little sunshine to make you happy. Also I see the fridge is out of food. You may want to order some foods today. Should I go ahead to order some eggs and vegetable?
User: I see. Thank you. Please go ahead.

- **Scheduler based on psychological and physiological advice:** Some personal scheduler should also refer to the psychological and physiological advice from the doctors.

These may also need the robot to search the relevant knowledge base for the recommendation for certain individual conditions. A relevant knowledge database should be built and constantly updated online.
- Other techniques involved **interconnected with other robots and internet.**

## 5. Future Directions: Trends, Challenges and Solutions

In addition to the above discussed technical challenges, social issues such as ethics, fairness, privacy, explainability, security and cognitive ability are among the challenges in developing domestic robots in the future. Although these issues have become trending research topics in various communities, corresponding techniques have not yet been widely applied to robotic products yet. Below, we discuss trends and challenges of developing domestic robots. The potential solutions towards these challenges are also presented.

### *5.1. Cognition*

When robots enter our daily life in the domestic environment, the aforementioned stages may bring about robot manipulation, assistive robots, virtual robots, conversational robots across in diverse settings in a safe and an adaptive way. The final problem about the domestic robots probably will be: how do we integrate such functions into one robot and also improve the quality of these functions at the same time? Furthermore, more challenges may include: reasoning, long- and short-time memories, intention, attention, imagination or meta-cognition.

#### 5.1.1. Multi-Modal Learning

Recently, transfer-learning strategy has been proposed to transfer learned knowledge across different tasks. Previous literature has proposed to use a unified model to learn multi-modality information and their association. Here, the "modality" refers to the information depicted in different perspectives of the same object or event via different physical expression, signal perception or different data-formats. The multi-modal learning can be (machine-learning) technically learned via the end-to-end training manner. For example, [141] initially processed sound, text and image with three sub-networks. Then, a higher-level network is added on top to learn a joint representation of the three modalities. Therefore, the learning of a higher-level multi-modal representation is constrained by aligning the three modalities. After training, emergence of a "concept" between un-trained modality pairs can be observed. Focusing on solving multiple machine learning tasks with a unique architecture, [142] extracted the common shared representation hierarchically. The network has been proven to be able to solve a few different problems in different modalities. Deploying transfer-learning in real-life applications would serve the purpose of building multi-purpose robots. It will also alleviate the lack of training data of notorious deep learning methods.

As we summarized from the techniques we introduced, most techniques we are concentrating on domestic robotics are the perception, action, understanding and communication problem where we can observe a trend that the deep learning methods are being widely used in all of the above three areas. Although we believe deep-learning methods would be one important method to achieve the understanding of multi-modal data, we still need other computational methods to achieve a higher-level of intelligence, i.e., the meta-cognition capabilities. It includes a number of aspects of processes, where we just name a few:

- The inference ability which learns the causal relationship;
- the meta-learning ability;
- the self-awareness ability.

In the next subsection, we will continue the discussion about the possible ways to build meta-cognition based on the computational intelligence techniques.

#### 5.1.2. Meta-Cognition

Cognitive functions are inspired by the way the human brain works. According to the Moravec's paradox, the higher-level intelligence probably does not need much computational power to be accomplished. Thus, the corresponding aforementioned cognitive functions may need techniques with less computational requirements, such as symbolic planner to accomplish. On the contrary, realizing low-level intelligence such as the coordination between perception and action, and how it connects the the higher-level cognitive functions, is still an open issue. This issue is much related to the topic about language grounding and understanding that we will discuss in the following Section 5.1.3. Specifically, in the context of domestic robots, it is embodied in the behaviors of intelligent agents in the physical world.

If we imagine the cognitive process are information processing activities where the information was ultimately from the objective world. The meta-cognition refers to the the processes about the of concepts, perception, judgment, or imagination and other psychological activities that relate to the aforementioned information activities, That is, the psychological function of individual thinking for information processing. For instance, when someone is reviewing the coursework by memorizing the details of each chapter of the textbook. The meta-cognition processes of self-evaluation helps to evaluate the outcome of the memorizing process. The self-regulation encourages and monitors the agent to stay still and to continue the reading. In addition, the generation of concepts will probably extract and digest the contents. In general, the meta-cognition includes, for example, reasoning about reasoning, reasoning about learning, and learning about reasoning [143,144]. The meta-cognition is also defined [145] as

> Meta-cognitive experiences are any conscious cognitive or affective experiences that accompany and pertain to any intellectual enterprise. An example would be the sudden feeling that you do not understand something another person just said.

Obviously, an implementation of meta-cognition will facilitates the learning of cognitive processes. Some of the meta-cognition abilities are particularly useful for establishing the safety and robustness in human–robot interactions. The most interesting applications include:

- A human–robot interaction (HRI) by implementing theory of mind (ToM). ToM is referred to a meta-cognitive process that an agent could think of and understand other thoughts and decisions made by its counterpart. Therefore, in the domestic robotic scenario, the robots can take into account (monitor) others' mental state and use that knowledge to predict others' behavior.
- A safety-lock based on an implementation of self-regulation. This self-regulation mechanism can be close to immediate awareness and body awareness. Therefore, this mechanism can control any non-safe decision making processes when the sensorimotor imagination of the own body is hurt.

#### 5.1.3. Language Grounding

The research question of symbol grounding problem is to discuss how symbols get their meanings. The language process in robots does not isolate from the environment but it is learned and understood in the physical world in which the robot is situated in. Here the symbols can be referred to in the spoken language, the vocal commands or the conversation. Therefore, the symbol grounding problem can be seen as a way of information interchanging and sharing while the robot and/or the human know the symbol has been grounded to one element/concept in the environment. After the symbol is grounded with the mutual parties, for instance, in human–robot interaction, the grounded set of intentionality can be mutually believed by both speakers. In addition to the essential target of symbol grounding, there are also other benefits for different perspectives of domestic robots.

Firstly, the process of symbol grounding can be useful and practical for the robots to incrementally learn and acquire a language as well as acquire a novel symbol (language). We agree that NLP is one of the core contents for the communication part of robotic applications. Due to the complexity of natural language, it is very difficult for the engineers to implement the whole content of language as well as its corresponding element in the environment for robots to understand and like human beings. It is true that the data-driven method can train both language models and visual classification well. Even the learned model can successfully understand the novel and complicated structures of language and visual objects. How these two structures can be linked and understood together is still a problem. Such links would be useful to link the symbolic structure, which the language model used to follow in the 1990s, and the deep-learning or other statistical methods.

Secondly, the grounding problem seems to be possible to be implemented in a robotic system and make it work [146], but there is a dilemma: the most important problem about the Z-condition is that the agent needs to be intentionality [147] is ultimately the hard problem [148] of consciousness, which is the root requirement that is essential for any autonomous agent to acquire such kind of "understanding". Therefore, it seems that it is difficult to let a robot understand a meaning. Nevertheless, it is believed in a goal-directed robot, such intentionality can be marked from other data to indicate a successful behavior [149]. In addition, with the grounded symbol, we can also solve the easier problem between verbal instructions and the tasks: How can we explain and reproduce the behavioral ability and function of meaning (and other intentional phenomena) in robots?

#### 5.1.4. Solutions

Indeed, researchers in the computational intelligence or artificial intelligence communities have proposed various methods in order to achieve the ultimate goals of "intelligence". Along with the debate about the definition and the goals of "intelligence" [150], it is suggested that the requirements for a robot to be equipped with "cognition" are higher than the "intelligent domestic robot". Therefore, to realize the aforementioned cognition for domestic robots needs more advances in computing techniques, cognitive sciences and neuroscience.

The detailed technical solutions for realizing cognitive domestic robots are not discussed in this section. Nevertheless, to build a holistic mechanism related with cognitive phenomena as well as the aforementioned intelligent functions, based on the concept of situated cognition, three aspects of research directions to build a holistic cognitive mechanism for domestic robotics should be considered:

1. Knowledge acquisition. The structure of knowledge may be acquired by embodiment. Therefore, the aforementioned multi-modal learning and a symbolic grounding may be necessary.
2. Behavioral learning. Such learning can be conducted via learning by curriculum or human demonstration.
3. Social interaction based on mutual understanding and theory of mind (ToM) [151]. The ToM ability, if it is successfully implemented, is able to allow the robot to dynamically switch the roles during the interaction. It also allows the robot to endow abilities of empathy to assist as a more practical robot assistant.

### *5.2. Data Safety and Ethics*

#### 5.2.1. Data Safety and Ethics in NLP

The security and robustness of machine learning models used in real-world environments such as a home is always an issue, especially the robots having close interaction with inexperienced users. As domestic robots have easy access to personal data of users such as personally identifiable information (PII) and personal health information (PHI), researchers have raised concerns of privacy protection in the past few years [152]. Although

the European Union has published the General Data Protection Regulation (GDPR) [153], personal data protection is far from being well protected.

For example, in order to be triggered by speech command, some smart speakers keep listening even when there is no interaction. Therefore, the speaker keeps listening to users all the time. It will harm users if companies leak these private conservation data. Although there have been widespread discussions on this topic, to the best of knowledge, there has not been a perfect solution to protect the privacy of users.

Ethics and fairness have been a trending topic in the natural language processing community. As the deep learning-based language model becomes popular, researchers discovered that these models encode biases from the data they were trained on. For example, gender biases have been observed in language models. Women are often associated with nurses and teachers in terms of employment. Names of black people are represented closer to negative words in word embedding spaces. Applying such NLP models to real-life applications would be extremely dangerous. After XiaoIce was put online for a few hours, users discovered that it uses abusive language in conversations. It had to be taken offline shortly.

To ensure that domestic robots conform to ethical codes as humans do, it is necessary to develop a standard evaluation method of ethics of robots. This evaluation method should be exhaustive enough to test unethical behaviors in daily communications with users from different backgrounds. Up to now, we are not aware of such industrial effort to systematically evaluate ethical behaviors of robots.

#### 5.2.2. Data Safety, Ethics and Explainability in Domestic Robots

As social companions in domestic environment, domestic robots ought to conform to the constraints of ethics in the way that humans do. That is, they should guarantee the safety of the data they access via their sensors, and follow ethical guidelines.

Safety is not a one-way framework while we only control the robot to avoid the physical interaction the human as much as possible. The human users should also need to understand what are robot is doing/going to do, as well as the rationality that behind these choices. This is also related to the explainability in AI research which addresses the task of explaining how machine learning models achieve certain decisions. As machine learning methods become more important in decision making processes, establishing trust, transparency and accountability of decision making process between human users and domestic robots is critical.

Deep learning models have been criticized as "black boxes" as it is difficult to analyze the parameter distribution of a deep learning model and understand how the result comes out. Moreover, adversarial samples also warn that deep learning algorithms are not always robust [154]. To this end, robotics research should also concentrate on the verification of the reliability of the model (e.g., [155,156]). From the perspective of model training, since the changes of data distribution also matter in the model performance, in the case of domestic robotics, the robustness of individual differences also should be tested (e.g., [157])

On the other hand, rather than using the deep learning techniques, it is possible to improve the explainability using explainable machine learning techniques, such as decision trees, symbolic planner, etc. A recent study shows that people feel most trust in human–robot interaction cases when they are able to see the explainable decisions the robot is making [158].

Most importantly, as domestic robots aim to serve ordinary users, it is critical to explain how they achieved certain decisions in the way ordinary users can understand. Researchers in the NLP community have worked towards on generating natural language descriptions to explain decision processes of deep learning models. This is a promising direction to build close relations between users and domestic robots. However, such effort have only been observed in academia, not in industrial communities.

#### 5.2.3. Solution

To build a robotic system which guarantees the data safety and ethical issue, a safety design of the whole system should be emphasized in different aspects of the robotic system:

1. Sensor-wise: sensors that use bio-metric information, such as cameras, sometimes can be replaced with other sensors such as LiDAR and infrared sensors [159–161]. If the bio-metric details of users must be captured, such details should be processed locally at the edge and should not be sent online.
2. Software-wise: it is crucial to ensure its reliability, adaptivity and ubiquity with the advancement of software technology in the network design. To fit the requirements of the ultra-fast network and computing, a designated middle-ware for the safety feature of the network is needed to be further investigated.

## 6. Summary

In this paper, we surveyed existing representative commercial domestic robots and related state-of-the-art computational methods, with a focus on the gap between them. Categorizing robotic products into a taxonomy, we went through state-of-the-art works on computational intelligence related to the core abilities by surveying reputable international conferences and journals in each domain. We conclude that the gaps lie between existing robot products and most advanced techniques. This is followed by a discussion on trends and challenges of developing robot products in the future. Ethical issues, security and privacy protection are key concerns in developing domestic robots in the future. Developing reliable domestic robots is a highly interdisciplinary task that involves knowledge from different AI domains and communities.

**Author Contributions:** Writing original draft preparation, J.Z.; literature search, J.Z., C.L., A.C., A.L. and X.L.; writing—review and editing, A.C. and A.L.; supervision, A.C. and A.L.; project administration, X.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** Z.J. and A.C. are sponsored by the Guangdong International Project 2020A1414010126. A.C. is supported by the Air Force Office of Scientific Research, USAF under Award No. FA9550-19-1-7002, by the UKRI TAS Node on Trust (EP/V026682/1) and the H2020 projects PERSEO, TRAINCREASE and eLADDA. X.L. is supported in part National key research and development program 2018AAA0100800, by the Key Research and Development Program of Jiangsu under grants BK20192004, BE2018004-04, Guangdong Forestry Science and Technology Innovation Project under grant 2020KJCX005, International Cooperation and Exchanges of Changzhou under grant CZ20200035.

**Data Availability Statement:** No data is available to support this study.

**Acknowledgments:** J.Z. would like to thank his aunt for allowing to stay in her apartment to do the writing during self-quarantine.

**Conflicts of Interest:** The authors have no conflicts of interest to declare.

## References

1. Cavallo, F.; Esposito, R.; Limosani, R.; Manzi, A.; Bevilacqua, R.; Felici, E.; Di Nuovo, A.; Cangelosi, A.; Lattanzio, F.; Dario, P. Robotic services acceptance in smart environments with older adults: User satisfaction and acceptability study. *J. Med. Internet Res.* **2018**, *20*, e264. [CrossRef]
2. Zhong, L.; Verma, R. "Robot Rooms": How Guests Use and Perceive Hotel Robots. *Cornell Hosp. Rep.* **2019**, *19*, 1–8.
3. Pyae, A.; Joelsson, T.N. Investigating the usability and user experiences of voice user interface: A case of Google home smart speaker. In Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, Barcelona, Spain, 3–6 September 2018; pp. 127–131.
4. Shibata, T.; Inoue, K.; Irie, R. Emotional robot for intelligent system-artificial emotional creature project. In Proceedings of the 5th IEEE International Workshop on Robot and Human Communication (RO-MAN'96 TSUKUBA), Tsukuba, Japan, 11–14 November 1996; pp. 466–471.
5. Yamamoto, T.; Terada, K.; Ochiai, A.; Saito, F.; Asahara, Y.; Murase, K. Development of human support robot as the research platform of a domestic mobile manipulator. *ROBOMECH J.* **2019**, *6*, 1–15. [CrossRef]

6. Abubshait, A.; Wiese, E. You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human—Robot interaction. *Front. Psychol.* **2017**, *8*, 1393. [CrossRef]
7. Holloway, J. Owners Really Like Their Robot Vacuums, Survey Says. 4 September 2018. Available online: https://newatlas.com/robot-vacuum-market/56200/ (accessed on 4 September 2018).
8. Chestnutt, J.; Lau, M.; Cheung, G.; Kuffner, J.; Hodgins, J.; Kanade, T. Footstep planning for the honda asimo humanoid. In Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 629–634.
9. Deng, L.; Abdel-Hamid, O.; Yu, D. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6669–6673.
10. Povey, D.; Burget, L.; Agarwal, M.; Akyazi, P.; Feng, K.; Ghoshal, A.; Glembek, O.; Goel, N.K.; Karafiát, M.; Rastrow, A.; et al. Subspace Gaussian mixture models for speech recognition. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 4330–4333.
11. Juang, B.H.; Rabiner, L.R. Hidden Markov models for speech recognition. *Technometrics* **1991**, *33*, 251–272. [CrossRef]
12. Sercu, T.; Puhrsch, C.; Kingsbury, B.; LeCun, Y. Very deep multilingual convolutional neural networks for LVCSR. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4955–4959.
13. Yu, D.; Xiong, W.; Droppo, J.; Stolcke, A.; Ye, G.; Li, J.; Zweig, G. Deep Convolutional Neural Networks with Layer-Wise Context Expansion and Attention. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 17–21.
14. Sak, H.; Senior, A.W.; Beaufays, F. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. In Proceedings of the INTERSPEECH 2014 15th Annual Conference of the International Speech Communication Association, Singapore, Singapore, 14–18 September 2014.
15. Pundak, G.; Sainath, T. Highway-LSTM and Recurrent Highway Networks for Speech Recognition. In Proceedings of the INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017.
16. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. *arXiv* **2015**, arXiv:1506.07503. .
17. Chiu, C.C.; Sainath, T.N.; Wu, Y.; Prabhavalkar, R.; Nguyen, P.; Chen, Z.; Kannan, A.; Weiss, R.J.; Rao, K.; Gonina, E.; et al. State-of-the-art speech recognition with sequence-to-sequence models. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4774–4778.
18. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
19. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; Volume 1.
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
21. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
22. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
25. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
26. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
27. Wu, J.W.; Cai, W.; Yu, S.M.; Xu, Z.L.; He, X.Y. Optimized visual recognition algorithm in service robots. *Int. J. Adv. Robot. Syst.* **2020**, *17*. [CrossRef]
28. Quan, L.; Pei, D.; Wang, B.; Ruan, W. Research on Human Target Recognition Algorithm of Home Service Robot Based on Fast-RCNN. In Proceedings of the 2017 10th International Conference on Intelligent Computation Technology and Automation (ICICTA), Changsha, China, 9–10 October 2017; pp. 369–373.
29. Li, P.; Chen, X.; Shen, S. Stereo r-cnn based 3d object detection for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7644–7652.
30. Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; Yan, S. Scale-aware fast R-CNN for pedestrian detection. *IEEE Trans. Multimed.* **2017**, *20*, 985–996. [CrossRef]
31. Schwarz, M.; Milan, A.; Periyasamy, A.S.; Behnke, S. RGB-D object detection and semantic segmentation for autonomous manipulation in clutter. *Int. J. Robot. Res.* **2018**, *37*, 437–451. [CrossRef]

32. Martinez-Martin, E.; Del Pobil, A.P. Object detection and recognition for assistive robots: Experimentation and implementation. *IEEE Robot. Autom. Mag.* **2017**, *24*, 123–138. [CrossRef]
33. Trigueros, D.S.; Meng, L.; Hartnett, M. Face Recognition: From Traditional to Deep Learning Methods. *arXiv* **2018**, arXiv:1811.00116.
34. Wang, M.; Deng, W. Deep face recognition: A survey. *arXiv* **2018**, arXiv:1804.06655.
35. Learned-Miller, E.; Huang, G.B.; RoyChowdhury, A.; Li, H.; Hua, G. Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*; Springer: Cham, Switzerland, 2016; pp. 189–248.
36. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
37. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Closing the gap to human-level performance in face verification. deepface. In Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; Volume 5, p. 6.
38. Sun, Y.; Liang, D.; Wang, X.; Tang, X. Deepid3: Face recognition with very deep neural networks. *arXiv* **2015**, arXiv:1502.00873.
39. Zheng, Y.; Pal, D.K.; Savvides, M. Ring loss: Convex feature normalization for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5089–5097.
40. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
41. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
43. Goswami, G.; Bharadwaj, S.; Vatsa, M.; Singh, R. On RGB-D face recognition using Kinect. In Proceedings of the 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Arlington, VA, USA, 29 September–2 October 2013; pp. 1–6.
44. Min, R.; Kose, N.; Dugelay, J.L. Kinectfacedb: A kinect database for face recognition. *IEEE Trans. Syst. Man Cybern. Syst.* **2014**, *44*, 1534–1548. [CrossRef]
45. Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* **2006**, *13*, 99–110. [CrossRef]
46. Bailey, T.; Durrant-Whyte, H. Simultaneous localization and mapping (SLAM): Part II. *IEEE Robot. Autom. Mag.* **2006**, *13*, 108–117. [CrossRef]
47. Montemerlo, M.; Thrun, S.; Koller, D.; Wegbreit, B. FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem. In Proceedings of the Eighteenth National Conference on Artificial Intelligence, Menlo Park, CA, USA, 28 July–1 August, 2002; pp. 1–6.
48. Montemerlo, M.; Thrun, S.; Koller, D.; Wegbreit, B. FastSLAM 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. *IJCAI* **2003**, *3*, 1151–1156.
49. Zhong, J.; Fung, Y.F. Case study and proofs of ant colony optimisation improved particle filter algorithm. *IET Control Theory Appl.* **2012**, *6*, 689–697. [CrossRef]
50. Liu, Y.; Thrun, S. Results for outdoor-SLAM using sparse extended information filters. In Proceedings of the 2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422), Taipei, Taiwan, 14–19 September 2003; Volume 1, pp. 1227–1233.
51. Bohren, J.; Rusu, R.B.; Jones, E.G.; Marder-Eppstein, E.; Pantofaru, C.; Wise, M.; Mösenlechner, L.; Meeussen, W.; Holzer, S. Towards autonomous robotic butlers: Lessons learned with the PR2. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 5568–5575.
52. Hornung, A.; Wurm, K.M.; Bennewitz, M. Humanoid robot localization in complex indoor environments. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 1690–1695.
53. Jamiruddin, R.; Sari, A.O.; Shabbir, J.; Anwer, T. RGB-depth SLAM review. *arXiv* **2018**, arXiv:1805.07696.
54. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]
55. Ackerman, E.; Guizzo, E. iRobot Brings Visual Mapping and Navigation to the Roomba 980. 16 September 2015. Available online: https://spectrum.ieee.org/automaton/robotics/home-robots/irobot-brings-visual-mapping-and-navigation-to-the-roomba-980 (accessed on 16 September 2015).
56. Karlsson, N.; Di Bernardo, E.; Ostrowski, J.; Goncalves, L.; Pirjanian, P.; Munich, M.E. The vSLAM algorithm for robust localization and mapping. In Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 24–29.
57. Newcombe, R.A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A.J.; Kohli, P.; Shotton, J.; Hodges, S.; Fitzgibbon, A.W. Kinectfusion: Real-time dense surface mapping and tracking. In Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; Volume 11, pp. 127–136.
58. Whelan, T.; Kaess, M.; Fallon, M.; Johannsson, H.; Leonard, J.J.; McDonald, J. Kintinuous: Spatially Extended KinectFusion. In Proceedings of the RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, Sydney, Australia, 9–10 July 2012; pp. 1–8.

59. Newcombe, R.A.; Fox, D.; Seitz, S.M. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 343–352.
60. Alajlan, A.M.; Almasri, M.M.; Elleithy, K.M. Multi-sensor based collision avoidance algorithm for mobile robot. In Proceedings of the 2015 Long Island Systems, Applications and Technology, Farmingdale, NY, USA, 1 May 2015; pp. 1–6.
61. Amditis, A.; Polychronopoulos, A.; Karaseitanidis, I.; Katsoulis, G.; Bekiaris, E. Multiple sensor collision avoidance system for automotive applications using an IMM approach for obstacle tracking. In Proceedings of the Fifth International Conference on Information Fusion, Annapolis, MD, USA, 8–11 July 2002; Volume 2, pp. 812–817.
62. Borenstein, J.; Koren, Y. Real-time obstacle avoidance for fast mobile robots. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 1179–1187. [CrossRef]
63. Borenstein, J.; Koren, Y. The vector field histogram-fast obstacle avoidance for mobile robots. *IEEE Trans. Robot. Autom.* **1991**, *7*, 278–288. [CrossRef]
64. Heinla, A.; Reinpõld, R.; Korjus, K. Mobile Robot Having Collision Avoidance System for Crossing a Road from a Pedestrian Pathway. U.S. Patent 10/282,995, 14 March 2019.
65. Dijkstra, E.W. A note on two problems in connexion with graphs. *Numer. Math.* **1959**, *1*, 269–271. [CrossRef]
66. Hart, P.E.; Nilsson, N.J.; Raphael, B. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 100–107. [CrossRef]
67. Stentz, A. Optimal and efficient path planning for partially known environments. In *Intelligent Unmanned Ground Vehicles*; Springer: Cham, Switzerland, 1997; pp. 203–220.
68. Stentz, A. The focussed D^* algorithm for real-time replanning. *IJCAI* **1995**, *95*, 1652–1659.
69. Elbanhawi, M.; Simic, M. Sampling-based robot motion planning: A review. *IEEE Access* **2014**, *2*, 56–77. [CrossRef]
70. Kavraki, L.; Latombe, J.C. Randomized preprocessing of configuration for fast path planning. In Proceedings of the 1994 IEEE International Conference on Robotics and Automation, San Diego, CA, USA, 8–13 May 1994; pp. 2138–2145.
71. Siciliano, B.; Sciavicco, L.; Villani, L.; Oriolo, G. *Robotics: Modelling, Planning and Control*; Springer-Verlag London: London, UK, 2009.
72. Brooks, R. A robust layered control system for a mobile robot. *IEEE J. Robot. Autom.* **1986**, *2*, 14–23. [CrossRef]
73. Brooks, R.A.; Connell, J.H. Asynchronous distributed control system for a mobile robot. In Proceedings of the Cambridge Symposium Intelligent Robotics Systems, Cambridge, MA, USA, 28–31 October 1986.
74. Maes, P. How to do the right thing. *Connect. Sci.* **1989**, *1*, 291–323. [CrossRef]
75. Maes, P. Situated agents can have goals. *Robot. Auton. Syst.* **1990**, *6*, 49–70. [CrossRef]
76. Arkin, R. Motor schema based navigation for a mobile robot: An approach to programming by behavior. In Proceedings of the 1987 IEEE International Conference on Robotics and Automation, Raleigh, NC, USA, 31 March–3 April 1987; Volume 4, pp. 264–271.
77. Rusu, P.; Petriu, E.M.; Whalen, T.E.; Cornell, A.; Spoelder, H.J. Behavior-based neuro-fuzzy controller for mobile robot navigation. *IEEE Trans. Instrum. Meas.* **2003**, *52*, 1335–1340. [CrossRef]
78. Aguirre, E.; González, A. Fuzzy behaviors for mobile robot navigation: Design, coordination and fusion. *Int. J. Approx. Reason.* **2000**, *25*, 255–289. [CrossRef]
79. Nattharith, P.; Güzel, M.S. Machine vision and fuzzy logic-based navigation control of a goal-oriented mobile robot. *Adapt. Behav.* **2016**, *24*, 168–180. [CrossRef]
80. Kim, J.; Mishra, A.K.; Limosani, R.; Scafuro, M.; Cauli, N.; Santos-Victor, J.; Mazzolai, B.; Cavallo, F. Control strategies for cleaning robots in domestic applications: A comprehensive review. *Int. J. Adv. Robot. Syst.* **2019**, *16*. [CrossRef]
81. Grigorescu, S.M.; Lüth, T.; Fragkopoulos, C.; Cyriacks, M.; Gräser, A. A BCI-controlled robotic assistant for quadriplegic people in domestic and professional life. *Robotica* **2012**, *30*, 419–431. [CrossRef]
82. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
83. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79. [CrossRef]
84. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
85. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.
86. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 4489–4497.
87. Goodale, M.A.; Milner, A.D. Separate visual pathways for perception and action. *Trends Neurosci.* **1992**, *15*, 20–25. [CrossRef]
88. Bayat, A.; Pomplun, M.; Tran, D.A. A study on human activity recognition using accelerometer data from smartphones. *Procedia Comput. Sci.* **2014**, *34*, 450–457. [CrossRef]
89. Florentino-Liano, B.; O'Mahony, N.; Artés-Rodríguez, A. Human activity recognition using inertial sensors with invariance to sensor orientation. In Proceedings of the 2012 3rd International Workshop on Cognitive Information Processing (CIP), Baiona, Spain, 28–30 May 2012; pp. 1–6.

90. Stork, J.A.; Spinello, L.; Silva, J.; Arras, K.O. Audio-based human activity recognition using non-markovian ensemble voting. In Proceedings of the 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, Paris, France, 9–13 September 2012; pp. 509–514.
91. Chen, C.; Jafari, R.; Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimed. Tools Appl.* **2017**, *76*, 4405–4425. [CrossRef]
92. Garcia-Ceja, E.; Galván-Tejada, C.E.; Brena, R. Multi-view stacking for activity recognition with sound and accelerometer data. *Inf. Fusion* **2018**, *40*, 45–56. [CrossRef]
93. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
94. Zhao, X.; Liang, X.; Liu, L.; Li, T.; Han, Y.; Vasconcelos, N.; Yan, S. Peak-Piloted Deep Network for Facial Expression Recognition, In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 425–442.
95. Zhong, J.; Canamero, L. From continuous affective space to continuous expression space: Non-verbal behaviour recognition and generation. In Proceedings of the 4th International Conference on Development and Learning and on Epigenetic Robotics, Genoa, Italy, 13–16 October 2014; pp. 75–80.
96. Li, J.; Yang, C.; Zhong, J.; Dai, S. Emotion-Aroused Human Behaviors Perception Using RNNPB. In Proceedings of the 2018 10th International Conference on Modelling, Identification and Control (ICMIC), Guiyang, China, 2–4 July 2018; pp. 1–6.
97. Noroozi, F.; Kaminska, D.; Corneanu, C.; Sapinski, T.; Escalera, S.; Anbarjafari, G. Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput.* **2018**. [CrossRef]
98. Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **2018**, *61*, 90–99. [CrossRef]
99. Poria, S.; Majumder, N.; Mihalcea, R.; Hovy, E. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *arXiv* **2019**, arXiv:1905.02947.
100. Schirmer, A.; Adolphs, R. Emotion perception from face, voice, and touch: Comparisons and convergence. *Trends Cogn. Sci.* **2017**, *21*, 216–228. [CrossRef] [PubMed]
101. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.
102. Barros, P.; Churamani, N.; Lakomkin, E.; Siqueira, H.; Sutherland, A.; Wermter, S. The omg-emotion behavior dataset. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–7.
103. Dhall, A.; Goecke, R.; Joshi, J.; Wagner, M.; Gedeon, T. Emotion recognition in the wild challenge 2013. In Proceedings of the 15th ACM on International Conference on Multimodal Interaction, Sydney, Australia, 9–13 December 2013; pp. 509–516.
104. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [CrossRef]
105. Zhu, W.; Mo, K.; Zhang, Y.; Zhu, Z.; Peng, X.; Yang, Q. Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv* **2017**, arXiv:1709.04264.
106. Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
107. Lin, G.; Walker, M. Stylistic variation in television dialogue for natural language generation. In Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark, 7–11 September 2017; pp. 85–93.
108. Akama, R.; Inada, K.; Inoue, N.; Kobayashi, S.; Inui, K. Generating stylistically consistent dialog responses with transfer learning. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Taipei, Taiwan, 27 November–1 December 2017; pp. 408–412.
109. McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*; University of Chicago Press: London, UK, 1992.
110. Kita, S. *Pointing: Where Language, Culture, and Cognition Meet*; Psychology Press: East Sussex, UK, 2003.
111. Bergmann, K.; Kopp, S. Increasing the expressiveness of virtual agents: Autonomous generation of speech and gesture for spatial description tasks. In Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems—Volume 1. International Foundation for Autonomous Agents and Multiagent Systems, Budapest, Hungary, 10–15 May 2009; pp. 361–368.
112. Chiu, C.C.; Morency, L.P.; Marsella, S. Predicting co-verbal gestures: A deep and temporal modeling approach. In Proceedings of the International Conference on Intelligent Virtual Agents, Delft, The Netherlands, 26–28 August; pp. 152–166.
113. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *arXiv* **2013**, arXiv:1310.4546.
114. Gers, F.A.; Schmidhuber, E. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Trans. Neural Netw.* **2001**, *12*, 1333–1340. [CrossRef]
115. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
116. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.

117. Winograd, T. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. Available online: http://dspace.mit.edu/handle/1721.1/7095 (accessed on 1 January 1971).
118. Huang, M.; Zhu, X.; Gao, J. Challenges in Building Intelligent Open-domain Dialog Systems. *arXiv* **2019**, arXiv:1905.05709.
119. Schlangen, D.; Skantze, G. A general, abstract model of incremental dialogue processing. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Athens, Greece, 30 March–3 April 2009; pp. 710–718.
120. Kopp, S.; Gesellensetter, L.; Krämer, N.C.; Wachsmuth, I. A conversational agent as museum guide–design and evaluation of a real-world application. In Proceedings of the International Workshop on Intelligent Virtual Agents, Kos, Greece, 12–14 September 2005; pp. 329–343.
121. Marge, M.; Nogar, S.; Hayes, C.; Lukin, S.; Bloecker, J.; Holder, E.; Voss, C. A Research Platform for Multi-Robot Dialogue with Humans. *arXiv* **2019**, arXiv:1910.05624.
122. Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; van den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3674–3683.
123. Hu, R.; Fried, D.; Rohrbach, A.; Klein, D.; Saenko, K. Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation. *arXiv* **2019**, arXiv:1906.00347.
124. Parisi, G.I.; Kemker, R.; Part, J.L.; Kanan, C.; Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Netw.* **2019**, *113*, 54–71. [CrossRef]
125. Konyushkova, K.; Sznitman, R.; Fua, P. Learning active learning from data. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4225–4235.
126. Xin, D.; Ma, L.; Liu, J.; Macke, S.; Song, S.; Parameswaran, A. Accelerating human-in-the-loop machine learning: Challenges and opportunities. In Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning, Houston, TX, USA, 15 June 2018; pp. 1–4.
127. Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; Jurafsky, D. Deep reinforcement learning for dialogue generation. *arXiv* **2016**, arXiv:1606.01541.
128. Yao, K.; Zweig, G.; Peng, B. Attention with intention for a neural network conversation model. *arXiv* **2015**, arXiv:1510.08565.
129. Han, T.; Hough, J.; Schlangen, D. Natural Language Informs the Interpretation of Iconic Gestures. A Computational Approach. In Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Taipei, Taiwan, 27 November–1 December 2017.
130. Wagner, P.; Malisz, Z.; Kopp, S. Gesture and Speech in Interaction: An Overview. *Speech Commun.* **2014**, *57*, 209–232. [CrossRef]
131. Picard, R.W. Affective computing: Challenges. *Int. J. Hum. Comput. Stud.* **2003**, *59*, 55–64. [CrossRef]
132. Breazeal, C.L. *Designing Sociable Robots*; MIT Press: London, UK, 2004.
133. Lowe, R.; Andreasson, R.; Alenljung, B.; Lund, A.; Billing, E. Designing for a wearable affective interface for the NAO Robot: A study of emotion conveyance by touch. *Multimodal Technol. Interact.* **2018**, *2*, 2. [CrossRef]
134. Battarbee, K.; Koskinen, I. Co-experience: User experience as interaction. *CoDesign* **2005**, *1*, 5–18. [CrossRef]
135. Lakomkin, E.; Zamani, M.A.; Weber, C.; Magg, S.; Wermter, S. On the robustness of speech emotion recognition for human-robot interaction with deep neural networks. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 854–860.
136. Liu, Z.; Wu, M.; Cao, W.; Chen, L.; Xu, J.; Zhang, R.; Zhou, M.; Mao, J. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA J. Autom. Sin.* **2017**, 4, 668–676. [CrossRef]
137. Zhong, J.; Yang, C. A Compositionality Assembled Model for Learning and Recognizing Emotion from Bodily Expression. In Proceedings of the 2019 IEEE 4th International Conference on Advanced Robotics and Mechatronics (ICARM), Toyonaka, Japan, 3–5 July 2019; pp. 821–826.
138. Gallagher, S. Empathy, simulation, and narrative. *Sci. Context* **2012**, *25*, 355–381. [CrossRef]
139. Asada, M. Development of artificial empathy. *Neurosci. Res.* **2015**, *90*, 41–50. [CrossRef] [PubMed]
140. Asada, M. Artificial Pain May Induce Empathy, Morality, and Ethics in the Conscious Mind of Robots. *Philosophies* **2019**, *4*, 38. [CrossRef]
141. Aytar, Y.; Vondrick, C.; Torralba, A. See, hear, and read: Deep aligned representations. *arXiv* **2017**, arXiv:1706.00932.
142. Kaiser, L.; Gomez, A.N.; Shazeer, N.; Vaswani, A.; Parmar, N.; Jones, L.; Uszkoreit, J. One model to learn them all. *arXiv* **2017**, arXiv:1706.05137.
143. Kralik, J.D. Architectural design of mind & brain from an evolutionary perspective. *Common Model Cogn. Bull.* **2020**, *1*, 394–400.
144. Jackson, P.C. *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language*; Courier Dover Publications: Mineola, NY, USA, 2019.
145. Flavell, J.H. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *Am. Psychol.* **1979**, *34*, 906. [CrossRef]
146. Cangelosi, A. Grounding language in action and perception: From cognitive agents to humanoid robots. *Phys. Life Rev.* **2010**, *7*, 139–151. [CrossRef] [PubMed]
147. Müller, V.C. Which symbol grounding problem should we try to solve? *J. Exp. Theor. Artif. Intell.* **2015**, *27*, 73–78. [CrossRef]
148. Chalmers, D. The hard problem of consciousness. In *The Blackwell Companion to Consciousness*; Wiley-Blackwell: Hoboken, NJ, USA, 2007; pp. 225–235.

149. Cubek, R.; Ertel, W.; Palm, G. A critical review on the symbol grounding problem as an issue of autonomous agents. In Proceedings of the Joint German/Austrian conference on artificial intelligence (Künstliche Intelligenz), Dresden, Germany, 21–25 September 2015; pp. 256–263.
150. Wang, P. On defining artificial intelligence. *J. Artif. Gen. Intell.* **2019**, *10*, 1–37. [CrossRef]
151. Frith, C.; Frith, U. Theory of mind. *Curr. Biol.* **2005**, *15*, R644–R645. [CrossRef]
152. Pagallo, U. The impact of domestic robots on privacy and data protection, and the troubles with legal regulation by design. In *Data Protection on the Move*; Springer: Cham, Switzerland, 2016; pp. 387–410.
153. Voigt, P.; Von dem Bussche, A. The eu general data protection regulation (gdpr). In *A Practical Guide*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2017; Volume 10, p. 3152676.
154. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
155. Selsam, D.; Liang, P.; Dill, D.L. Developing bug-free machine learning systems with formal mathematics. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 3047–3056.
156. Sun, X.; Khedr, H.; Shoukry, Y. Formal verification of neural network controlled autonomous systems. In Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control, Montreal, QC, Canada, 16–18 April 2019; pp. 147–156.
157. Platanios, E.; Poon, H.; Mitchell, T.M.; Horvitz, E.J. Estimating accuracy from unlabeled data: A probabilistic logic approach. *arXiv* **2017**, arXiv:1705.07086.
158. Edmonds, M.; Gao, F.; Liu, H.; Xie, X.; Qi, S.; Rothrock, B.; Zhu, Y.; Wu, Y.N.; Lu, H.; Zhu, S.C. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Sci. Robot.* **2019**, *4*, eaay4663. [CrossRef] [PubMed]
159. Naser, A.; Lotfi, A.; Zhong, J. Adaptive Thermal Sensor Array Placement for Human Segmentation and Occupancy Estimation. *IEEE Sens. J.* **2020**, *21*, 1993–2002. [CrossRef]
160. Naser, A.; Lotfi, A.; Zhong, J.; He, J. Heat-map based occupancy estimation using adaptive boosting. In Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 19–24 July 2020; pp. 1–7.
161. Cheng, Y.; Wang, G.Y. Mobile robot navigation based on lidar. In Proceedings of the 2018 Chinese Control And Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; pp. 1243–1246.