Bridging the Gap between Robotic Applications and Computational Intelligence in Domestic Robotics

Junpei Zhong*‡, Ting Han*†, Ahmad Lotfi‡, Angelo Cangelosi†§¶, Xiaofeng Liu¶
*Equal contributions
‡ School of Science and Technology,
Nottingham Trent University, Nottingham, NG11 8NS, UK
† Artificial Intelligence Research Center, AIST, Aomi 2-3-26, 135-0064, Tokyo, Japan
§ School of Computer Science, University of Manchester, Manchester, M13 9PL, UK
¶ College of IoT Engineering, Hohai University, Changzhou, 213022, China

***Abstract*—With the speedy development of hardware (e.g., high performance computing, smaller and cheaper sensors) and software (e.g., deep learning techniques and interaction technologies), robotic products and IoT devices have gradually become accessible to household users. Typical application scenarios of domestic robots include: 1) Providing physical assistance such as floor vacuuming; 2) Providing social assistance to answer questions; and 3) Providing education and cognitive assistance such as offering partnerships. In this paper, we provide a brief overview of available domestic robots, particularly focusing on the services they provide and corresponding computational techniques incorporated in these services. We first provide an overview of available commercial domestic robots and state-of-the-art computational intelligence techniques, then discuss the gap between current robotic systems and advanced computational techniques. Finally, we analyze what are the next developmental stages for these techniques with the emergence and development of the domestic robotic platforms.**

***Keywords*—Domestic robotics; Intelligent robotics; Machine learning; Smart Home; Interactive robotics; Service robotics.**

I. Introduction

Commercial domestic robots have obtained growing attention over the past few years as they become more affordable while providing affectionate appearance and amiable interfaces. Designed to assist humans in domestic domains, domestic robots can assist users to finish one or several tasks, such as cleaning, delivery, communicating via speech or texts to provide users with new information (e.g. translating, online teaching, online shopping), acting as an entertainment media via social interactions or as a partner. While previous overviews on domestic robotics often focus on one application scenario or topic [2], [3], [4], we focus on discussing the gaps between existing robotic applications and state-of-the-art computing algorithms.

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 acting** (see Fig. 1). 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. As most users are not equipped with professional knowledge on how robotic systems work, it is essential that domestic robots can communicate with users in friendly and efficient manner to provide high quality assistance. In other words, domestic robots should be able to understand human users through natural conversations, recognizing intentions and emotions (i.e. the internal status) of users via automatic learning, so that users do not have to master complex programming and operating skills to obtain assistance.

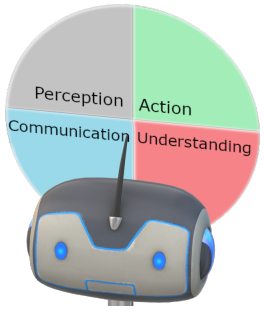

Fig. 1. Four Basic Functions of Domestic Robots

According to the assistance service they provide, domestic robots are designed with different physical appearance. Humanoid robots such as Nao[1], Pepper and HSR[2] are designed to offer cognitive assistance. In comparison, some robots designed to assist with household tasks such as vacuum cleaner robots[3] and Samsung family hub fridge[4]), also physically exist. However, they do not have human-like appearance. Robots don't have to physically exist. For instance, virtual robots, which mainly provide service via conversation, do not physically exist, but are embedded in devices such as tablets in the form of software (e.g. software hub for smart homes, such as Amazon Echo[5] and Google Home[6]) or other devices.

In following discussions, we roughly categorize domestic

[1]https://www.softbankrobotics.com/emea/en/nao
[2]https://www.toyota-global.com/innovation/partner_robot/robot/#link02
[3]https://www.irobot.com/
[4]https://www.samsung.com/uk/refrigerators/family-hub/
[5]https://en.wikipedia.org/wiki/Amazon_Echo
[6]https://store.google.com/gb/product/google_home

TABLE I
THE SPECTRUM OF SOFTWARE ROBOTS AND HARDWARE ROBOTS.

Main categories	Sub categories	Examples	Applications		
			Physical assistance	Social assistance	Cognitive assistance
Software Robots	Virtual Robots	Google Home	✗	Assisting, Scheduling, Calling, etc	Entertainment
		Amazon Eco	✗	Q & A Online ordering, etc	Entertainment
	IoT Robots	Nest Thermostat	Adjust heating	✗	✗
		Samsung Hub Freezer	Watch food storage	Question answering; Online ordering	Entertainment: TV & Music
	Interactive Robots	Pepper	Very limited	Assistant, Receptionist Healthcare	Limited entertainment
		Paro[1]	✗	✗	Entertaining and comforting elderly
Hardware Robots	Service Robots	HSR	Capable of different actions	Very limited	✗
		iRobot Roomba	Cleaning	✗	✗

robots into *software robots* and *hardware robots*: Software robots are software systems that run on host devices, rather than physically existing as standalone machines. They can basically communicate with the users via any device in the domestic environment. Comparing to software robots, hardware robots physically exist, while they can also embed software robots. Robots in both categories perceive and process information from the environment through the sensors they are equipped with, then respond with physical behaviors such as conversing, cleaning the floor or providing entertainment services.

Practically, there are no clear boundaries for categorizing domestic robots. Basically, human users expect a domestic robot to react in human-like manners such as efficient communicating. That is, regardless their appearance and computational approaches, robots should be able to perceive the situated environment, mutually communicate with humans, and respond with conversations or physical behaviors. In this article, we summarize and discuss four sub-categories of the domestic robots: virtual robots, internet of things (IoT) robots, service robots and interactive robots. These four sub-categories spread across a spectrum of software-hardware robots, which focusing on different services. With easier access of WiFi and 5G technologies, most of the above-mentioned robots in the market are able to provide various services, hence, position in the middle of the spectrum, making domestic robots even more suitable for the demanding market and become popular for households. For instance, the Pepper robot, targeting for social interaction, is also able to download most of its function modules through the App market on the cloud. In the meanwhile, it can also obtain the sensory information from its embedded camera. Moreover, it is able to deliver its bodily language, the conversation, as well as the screen display to better communicate with the users. Furthermore, it is a trend that these four sub-categories of robots can inter-connect with each other and the boundary may become even more blurried in the future. For instance, the Amazon Echo 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. In Tab. I, samples of each sub-category are listed and analyzed according to the services they provide. Following the aforementioned categories of services that domestic robots provide, we list three categories of assistance in the table: physical assistance, social assistance, cognitive assistance in domestic environment.

The rest of the paper is organized as follows: We first provide an overview of some representative and commercially available domestic robots, discussing the tasks and services they provide. Next, we briefly introduce the computational techniques, or computing technologies in a broader sense, that are relevant to the domestic robotic platforms. We also compare on which levels of integration of these techniques are for the domestic robot systems, and future directions to build better domestic robot products.

II. RELATED COMPUTING TECHNOLOGIES

A. Action: SLAM, Obstacle Avoidance and Path Planning

One of the basic requirements for domestic robots is the ability to move around in the domestic environment. Therefore, navigation, simultaneous localisation and mapping (SLAM) and path planning are three intersected modules and continuous steps towards building intelligent domestic robots.

Robots with SLAM algorithms build and update the map of an unknown environment while tracking their own location in the mean time. SLAM is a theoretically mature method based on the statistical estimation, however, most of its problems are within the engineering domain. For instance, the trade-off of usage of expensive sensors and accuracy of the SLAM, SLAM

in the dynamical environment and so on. On the contrary of the expensive laser sensors that previously used in unmanned vehicles, the SLAM with low-cost depth camera-based sensors has also grown rapidly over the past few years in domestic robotic applications, especially for products of commercial or academic uses where the requirement of accurate mapping is less important than the equipment costs. For example, the vacuum cleaning robot iRobot Roomba 980 [5] is based on vSLAM [6]. Also, the SLAM algorithms employing only vision sensors, such as monocular cameras, 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.

On top of the results from SLAM, navigation for most of the physical domestic robots should be equipped with two particular abilities: obstacle avoidance and path planning. Among these two abilities, the obstacle avoidance usually requires robots to react in a short-time frame to avoid static objects. In robotic applications nowadays, it requires reasonable sensors deployments using ultrasound, tactile and infrared sensors. In academic research, the object avoidance function has also been relatively well-studied. And it is integrated with object recognition, object following, or with fast-moving objects. Some of them have been developed and used in domestic robots where in some scenarios, they face a lot of moving objects surrounding, for instance, to finish a delivery in a dynamic environment [7].

On the contrary of the object avoidance ability that reacts in a short time, the robots' path planning ability usually requires the embedded system to plan ahead from one location to another, with the consideration of the factors such as time and energy. Using the map and the location information obtained from SLAM, the path planning algorithm determines appropriate motion actions 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 solution is considered as complete if the planner in finite time either produces a solution or correctly reports that there is no solution for the task. Most of the complete algorithms are geometry-based. These algorithms are usually planned directly on the occupant grid map in the field of mobile robots (i.e., a matrix composed of pixels) with the search methods, the Dijkstra algorithm [8], the A* algorithm [9] and so on. Particularly, the Dynamical A* (called D*) algorithm [10], [11], focusing on the path planning in a dynamical environment, has been successfully applied in mobile robots and autonomous navigation applications, such as the Mars rovers and the CMU team who participated the DARPA urban challenge. The representative methods of sampling-based algorithms are RRT and PRM algorithms. Since in some of the scenarios, high-complexity and high-dimensional space in the complete path-planning algorithm exists, the random sampling algorithms [12] can effectively solve this problem by drawing random samples to form a graph (PRM) [13] or a tree (RRT) [14] 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 is needed. For instance, the D* algorithm has been embedded as part of the navigation module (e.g. SLAMWare[7]) and widely used in many domestic robot products which require mobility.

B. Perception: Object Detection, Face Detection and Face Recognition

The tasks of object recognition and face recognition are relevant, as both of them aim to identify objects/faces in videos or images. They have been hot topics in the field of CV since three decades ago. Before convolutional neural networks (CNN) based methods achieved prominent success, most of the classical methods of object detection extract features with feature extraction algorithms, such as SIFT (Scale-invariant feature transform) [15], and Viola-Jones object detection framework [16]. Designing feature extraction algorithms is key to obtain a good performance of these algorithms. In comparison, deep learning based methods (e.g., CNN-based methods) adopt an end-to-end learning strategy, thus feature extraction is integrated into the learning procedure. These methods often perform faster, consume less resource, and outperform classical methods on tasks of real-time object detection.

Current deep learning based object detection methods fall into two categories: 2-stage methods and 1-stage methods, which are distinguished by whether the selective search for the objects is required before the object recognition is processed. For instance, the R-CNN [17], Fast R-CNN [18] use search methods [19] and the Faster R-CNN [20] uses the RoI (Region of Interest) pooling layer to find the bounding boxes with various aspect ratios and sizes for the possible object appearances. Afterwards, the object classification methods based on deep learning (e.g. CNN) are applied to recognize the objects within bounding boxes. Compared with the 2-stage methods, the 1-stage methods (e.g. YOLO [21] and SSD [22]) usually perform even faster due to the fact that the search and classification tasks are done with one single network. However, they often achieve lower accuracy than the 2-stage methods. As a result, they are usually employed in devices with strictly requirements in processing speed and low-power. Since the state-of-the-art object detection methods have achieve satisfactory performances in terms of both their accuracy and processing speed, they have been widely integrated in the perception part of the domestic robotic systems, such as autonomous navigation [23], pedestrian detection [24], manipulation [25] and other robotic applications [26].

Compared to the object detection and facial detection methods, face recognition is more challenging. Subtle differences

[7]https://www.slamtec.com/en/Slamware

on faces might indicate different identities, and consequentially affect the recognition results. An important requirement for the facial recognition is to perform "in-the-wild": when deployed in a random environment, human faces in the real world can be highly variable, making face recognition one of the most challenging tasks in CV (for more comprehensive reviews, see [27], [28]). Nevertheless, the facial recognition nowadays using deep learning methods also achieve impressive performance which are on par with human performance. The deep learning methods for face recognition (e.g. DeepFace [29], DeepID [30] and RingLoss [31]) usually use mainstream CNN networks (e.g. AlexNet [32], VGGNet [33], ResNet [34]) 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 face recognition methods are usually achieved by improving the CNN architectures (see also [28]). In the field of service robotics, the face recognition tasks face additional challenges: first, in the dynamic environment, most of the face recognition methods embedded in robotic systems require users to stand in the view field of the camera (assuming they are also at a reasonable distance from the camera), which might be inconvenient/unfriendly, despite that it still struggles to achieve real-time recognition. Second, the usages of other sensors, such as RGB-D cameras (e.g. [35], [36]), voice recognition can also be used for person identification to avoid the in-the-wild problem.

C. Understanding: Action Recognition, Emotion Recognition

At present, the most popular sensing techniques in robotic systems for action recognition are still camera-based. In the CV community, the action recognition task aims to identify various actions which are performed throughout or during part of the entire procedure based on video sensors. Similar to other tasks in CV, the state-of-the-art methods for action recognition mostly employ deep learning methods. Most of the successful deep learning methods for action recognition are based on or extend the work of either Two-Stream Convolutional Networks [37] (TSN) or 3D-convNet (C3D) [38]. The TSN method is inspired by the neuroscience findings in the visual cortex [39] that the separation of ventral and dorsal pathways to deal with the visual information for perception and action. In practice, the two-stream idea was proposed according to two independent recognition streams (space and time). 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 are implemented by CNN. Besides, separating the time stream and the spatial stream improves the generalization ability of the network. In the C3D and other related work, a 3-dimensional convolution kernel appears to be an effective video descriptor for the actions in videos. Particularly, the size of the convolution kernel should be $3 \times 3 \times 3$. The C3D method is supposed to be universal, and easy to comprehensive because it owns the same principle of general CNN. Besides action recognition based on visual information, some other sensors could also be possible to be deployed where the visual information could not be fully utilised due to privacy-related issues. In these scenarios, some ubiquitous sensing techniques such as the accelerometer [40], inertial sensors [41], microphones [42] or the combination of more than two modalities [43], [44] can also be used for action/activity recognition, especially when the privacy of the users is taken into consideration.

Emotion recognition is an important factor while the robot is engaged with the social interaction with humans. The emotion status can be expressed in facial and bodily features. Facial expression is thus an essential way for the robots to identify the emotion information. Similar to 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 achieve more robust performance, while the main recognition techniques are still CNN architectures (e.g. [32], [33], [45]) plus some fine-tuning tricks . For instance, [46] consider the temporal relation between 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 also a GoogleNet [45] which achieves good performances on many data-sets.

Besides recognizing emotions through facial expressions, it is worth mentioning that other forms of presentation (e.g. the bodily behavior [47], [48], [49], voice [50] and conversation [51]) can also be adopted as cues for emotion recognition tasks. From the perspective of computer engineering, using facial expression as a cue to recognize emotion is also the most straightforward and easy-to-labeled way. Nevertheless, we cannot ignore other possibilities of other substantial cues for emotion recognition when it is used as one of the functions in human-robot interaction, where multiple cues of sensor signals can complement each other. Hence, the challenges are: when applied in a robotic system, how to utilize multi-modal signals, social contexts, and common knowledge to identify the emotional status of users?

D. Communication: Speech Recognition and dialogue system

Natural language is perhaps the most natural way to communicate in our daily life. Automatic speech recognition (ASR) is the first step to enable natural communication between robots and humans. A speech recognition system usually includes five components: acoustic analysis for feature extraction, acoustic model, language model, pronunciation dictionary and the recognizer for recognition.

An acoustic model processes extracted audio features to estimate probability of how the speech is formed from the acoustic symbol(phoneme). Even before deep learning methods were applied to speech recognition task, there haven been mature ASR models. For example, the classic Gaussian Mixture Model (GMM) (e.g. [52]) and Hidden Markov Model

(HMM) [53]. With the rise of deep neural networks, the CNN (e.g. [54]), RNN (e.g. [55]), attention mechanism (e.g.[56]), encoder-decoder architectures [57] and other acoustic models have reduced the error rate dramatically compared with the traditional acoustic models. The basic principle is to analyze and predict the acoustic features, for example the features in the frequency domain through Fourier Transform, obtained from acoustic analysis.

Specifically, the current dialogue system can be roughly divided into two types: (1) task-oriented dialogue systems, and (2) open-domain dialogue systems (also known as chat-bots).

Different scenarios require different levels of communication skills. For simple physical assistance, pressing buttons or using graphical user interface (GUI) is sufficient. For more complex tasks such as cognitive assistance (e.g., reduce the symptoms of Alzheimer via chatting), more natural and human-like communication skills such as interactiveness in dialogues[58] and incrementatility in situated dialogue systems [59], [60] are essential. To achieve more natural communication, especially in complex tasks such as understanding or giving navigation descriptions[61], [62], [63], incorporating hand gestures would be beneficial[64].

III. Robotic Applications and the Gaps

In this section, we discuss the gap from perspectives of robotic applications. First, we provide an overview of the techniques implemented in existing robotic applications. Then, we discuss the gap between such techniques and current state-of-the-art algorithms in academic research.

A. Virtual Robots

The current virtual robots (e.g., Siri[8], Amazon Echo[9], Xiaoice[10]), especially those from the commercial companies, integrates a lot of services from their own, for instance, question answering, online shopping and playing online music. They are highly complex intelligent engineering systems which include a couple of existing services, among which the conversation robot is one of the interfaces. As such, the conversation robot uses primarily the communication function, but has include functions that from the cloud. From the perspective of implementing a conversation robot, there are various challenges, such as 1) how to reasonably use the context information, 2) how to generate respons according to the common knowledge, and so on. Some commercial products have been exploring methods to use common sense knowledge to build better dialogue systems. These systems rely on large scale datasets, thus they are able to estimate the preferences of the normal users and give recommendations of different products according to users' past behaviors. However, these systems are also suffering from noises in the datasets such as gender biases. To build virtual agents that are specialized for individual users, a virtual robot should be able to track communication history and learn to adapt while communicating.

A further development of the Q&A function may include the conversation system, which automatically narrow the search result with the development of the conversation or dialog systems, which finish one or several tasks.

B. IoT Robots

Similar to virtual robots, current IoT robots also connect and retrieve information from the Internet. The most significant difference between them is that the IoT robots usually also own one or more sensors, and probably have the ability to execute some basic actions, especially actions of home appliance as well. Perception, communication, and possibly action modules are key parts of IoT robots.

As mentioned before, the perception solely depends on what kind of sensors they are integrated with. For instance, there are cameras in the family hub freezer. And it is possible that the object recognition techniques can be used for this camera to identify what kinds of food remain and what should be ordered. This can be found as a separate product in the market[11], but the performance in recognition, the convenience of use and the integration with internet seems far from perfect. Furthermore, another existing problem is that with the excellent performance of object recognition tasks with specific datasets, how such networks can be implemented in the IoT robots. Specifically, two problems are worth being investigated:

1) How to deal with the objects in the wild with different conditions in household? For instance, object recognition in the smart fridge with limited lighting conditions and limited space?
2) How to implement the state-of-the-art CNN object recognition models with limited computational power?

C. Service Robots

Perception and action modules are key components of service robots in domestic environments, as they are designed to perceive the environment and act to accomplish certain tasks to assist the users. With the progress of research in robotic perception and action, service robotic applications nowadays can at least finish particular pre-defined domestic tasks, such as vacuum cleaning, delivery etc. We believe that with the fast development, a successful engineering integration of them and all the gap we mentioned in the Section 2, will result in more successful service robotic products. Besides the aforementioned works, to build better service robots, promising future research and engineering directions are:

1) Service robots for broader application scenarios: most of the service robots are designed to carry out one particular task at present[12]. In the future, with the development of motor control, mechanical design and robot perception, a

[8]https://www.apple.com/siri/

[9]https://en.wikipedia.org/wiki/Amazon_Echo

[10]https://en.wikipedia.org/wiki/Xiaoice

[11]https://store.smarter.am/products/fridgecam

[12]Here we define the term 'task' as a collection of similar behaviours. For instance, 'cooking' is one task for a service robot, but it includes a few different behaviours of robot arms.

TABLE II
FOUR CATEGORIES OF ROBOTS AND THE RELATED TECHNIQUES

✗ indicates that this category of robots does not need such kind of technique; → 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	IoT	Interactive	Service
SLAM and Navigation	✗	✗	→	→
Object Recognition	→	→	↗	↗
Face Recognition	→	↗	↗	↗
Action Recognition	✗	↗	↗	↗
Emotion Recognition	✗	↗	↗↗	✗
Speech Recognition	→	→	→	→
Dialog System	→	→	↗	↗

single robot assistant with an affordable price should be able to finish a few different tasks. Furthermore, a service assistant should also be adaptive enough to switch among tasks.

2) Building closer relationships with users. With frequent and varied collaborations between service robots and users, more friendly and safe service robots will improve user experience in terms of effective communication.

D. Interactive Robots

A few interactive robotic platforms such as Nao, Pepper and Jibo have also been commercially available. Researchers and engineers have developed interactive robots to provide reception, health care or other services based on these platforms. While these robots share similar system architecture, they often adopt most mature and reliable existing algorithms (e.g. ASR and emotion recognition embedded in Nao robot), rather than utilizing the algorithms with the best performance (e.g., Google ASR based on deep learning).

On one hand, the embedded SDKs of the platforms might not be flexible enough to customize algorithms. On the other hand, when implementing interactive robots, the priority is to provide reliable service while minimizing risks of misunderstanding. Hence, most dialogue systems implemented in the interactive robots are task-oriented dialogue systems rather than open-domain dialogue systems. Interacting with such robots is far less satisfying than interacting with a real person, as non-task-oriented aspects of interactions, such as humor and chit-chatting, are still missing in most of existing interactive robots.

Furthermore, human communications are multi-modal in nature. Non-verbal communication approaches such as hand gestures are tightly integrated in our daily communication. Enabling non-verbal communication in interactive robots would lead to more natural and intuitive interactions. Currently, the Pepper robot provides a limited set of built-in gestural behaviors that performing pre-defined functions. To enable natural multi-modal communications, free and context-dependent gestural behaviors are essential. While there are research works on characteristics of the human-human interaction to build human-robot interaction (e.g. [65], [66]), it is still a long way to go from research prototypes to applications in realistic scenarios.

Similar case also exist in the understanding module of the interactive robots, which encounters even higher uncertainties when interacting with users from different cultures, backgrounds and individual differences of the communication counterpart. In the CV community, lots of effort have been put in the facial emotion recognition and action recognition, their successful and robust integration in a robotic system can also be found in some interactive robots (e.g. Pepper), although the performance is still far from satisfaction. Besides the engineering challenges, the culture and context differences also affect their results [67]. To solve this problem, perhaps employing novel learning techniques (e.g., active learning) and collaborating with researchers from linguistic and cognitive science background will lead to more individualized systems.

Tab.II summarizes how the above-mentioned categories of domestic robots use the state-of-the-art computing techniques, and promising challenges should be tackled in the future. We show that the action module has been well developed. Although some engineering problems still exist in robotic navigation, most of its algorithms are well developed and integrated. Its counterpart, the perception module which includes object and face recognition functions, are quite useful in all of the four sub-categories. The state-of-the-art deep learning algorithms have been utilized in virtual and IoT robots which do not have high mobility. On the other hand, it's still unclear how to solve the problem with limited energy and computational power within limited time with the deep learning based methods.

Comparing with conventional modules of perception and action, the communication and understanding components are still far from perfect. We conjecture the main reason that the users are not fully convinced to buy an interactive or service robots 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 domestic robots as machines or toys.

IV. SUMMARY

We presented an overview of existing domestic robotics, focusing on the gaps between currently available commercial robots and state-of-art computing techniques. Due to the limited space, many topics remain to be discussed. For instance, 1) How we can adapt the computational techniques, especially the machine learning techniques, in a dynamic environment where the domestic robots are facing? 2) How we can optimize the machine learning methods in terms of the design of their architectures? 3) How much further we can achieve using machine learning techniques? The works we have mentioned in this paper are by no means exhaustive. We will extend it to a longer version in the future. By discussing representative works on domestic robotics, we aim to point out that the final target of filling such potential gaps is to improve user experience of domestic robotics.

ACKNOWLEDGMENT

TH is supported by a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). XF is supported by the Key Research and Development Program of Jiangsu under grants BE2017071, BE2017647 and BE2018004-04, the Projects of International Cooperation and Exchanges of Changzhou under grant CZ20170018, the Fundamental Research Funds for the Central Universities under grant 2018B47114, the Open Research Fund of State Key Laboratory of Bioelectronics, Southeast University under grant 2019005, and the State Key Laboratory of Integrated Management of Pest Insects and Rodents under grant IPM1914.

REFERENCES

[1] T. Shibata, K. Inoue, and R. Irie, "Emotional robot for intelligent system-artificial emotional creature project," in *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*. IEEE, 1996, pp. 466–471.

[2] E. G. Christoforou, A. S. Panayides, S. Avgousti, P. Masouras, and C. S. Pattichis, "An overview of assistive robotics and technologies for elderly care," in *Mediterranean Conference on Medical and Biological Engineering and Computing*. Springer, 2019, pp. 971–976.

[3] D. Paulius and Y. Sun, "A survey of knowledge representation in service robotics," *Robotics and Autonomous Systems*, vol. 118, pp. 13–30, 2019.

[4] A. P. van der Veere *et al.*, "Presenting domestic care technology and elderly care in japanese newspapers," 2018.

[5] E. Ackerman and E. Guizzo, "irobot brings visual mapping and navigation to the roomba 980," https://spectrum.ieee.org/automaton/robotics/home-robots/irobot-brings-visual-mapping-and-navigation-to-the-roomba-980, published 16 Sep 2015.

[6] N. Karlsson, E. Di Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M. E. Munich, "The vslam algorithm for robust localization and mapping," in *ICRA*, 2005, pp. 24–29.

[7] A. Heinla, R. Reinpõld, and K. Korjus, "Mobile robot having collision avoidance system for crossing a road from a pedestrian pathway," May 7 2019, uS Patent App. 10/282,995.

[8] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[9] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[10] A. Stentz, "Optimal and efficient path planning for partially known environments," in *Intelligent Unmanned Ground Vehicles*. Springer, 1997, pp. 203–220.

[11] A. Stentz *et al.*, "The focussed dˆ* algorithm for real-time replanning," in *IJCAI*, vol. 95, 1995, pp. 1652–1659.

[12] M. Elbanhawi and M. Simic, "Sampling-based robot motion planning: A review," *Ieee access*, vol. 2, pp. 56–77, 2014.

[13] L. Kavraki and J.-C. Latombe, "Randomized preprocessing of configuration for fast path planning," in *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*. IEEE, 1994, pp. 2138–2145.

[14] S. M. LaValle, "Rapidly-exploring random trees: A new tool for path planning," 1998.

[15] D. G. Lowe *et al.*, "Object recognition from local scale-invariant features." in *iccv*, vol. 99, no. 2, 1999, pp. 1150–1157.

[16] P. Viola, M. Jones *et al.*, "Rapid object detection using a boosted cascade of simple features," *CVPR (1)*, vol. 1, no. 511-518, p. 3, 2001.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[18] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[19] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[23] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7644–7652.

[24] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017.

[25] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 437–451, 2018.

[26] E. Martinez-Martin and A. P. Del Pobil, "Object detection and recognition for assistive robots: Experimentation and implementation," *IEEE Robotics & Automation Magazine*, vol. 24, no. 3, pp. 123–138, 2017.

[27] D. S. Trigueros, L. Meng, and M. Hartnett, "Face recognition: From traditional to deep learning methods," *arXiv preprint arXiv:1811.00116*, 2018.

[28] M. Wang and W. Deng, "Deep face recognition: A survey," *arXiv preprint arXiv:1804.06655*, 2018.

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Closing the gap to human-level performance in face verification. deepface," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, vol. 5, 2014, p. 6.

[30] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873*, 2015.

[31] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5089–5097.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, "On rgb-d face recognition using kinect," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013, pp. 1–6.

[36] R. Min, N. Kose, and J.-L. Dugelay, "Kinectfacedb: A kinect database for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 11, pp. 1534–1548, 2014.

[37] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[39] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.

[40] A. Bayat, M. Pomplun, and D. A. Tran, "A study on human activity recognition using accelerometer data from smartphones," *Procedia Computer Science*, vol. 34, pp. 450–457, 2014.

[41] B. Florentino-Liano, N. O'Mahony, and A. Artés-Rodríguez, "Human activity recognition using inertial sensors with invariance to sensor orientation," in *2012 3rd International Workshop on Cognitive Information Processing (CIP)*. IEEE, 2012, pp. 1–6.

[42] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 509–514.

[43] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4405–4425, 2017.

[44] E. Garcia-Ceja, C. E. Galván-Tejada, and R. Brena, "Multi-view stacking for activity recognition with sound and accelerometer data," *Information Fusion*, vol. 40, pp. 45–56, 2018.

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[46] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision*. Springer, 2016, pp. 425–442.

[47] J. Zhong and L. Canamero, "From continuous affective space to continuous expression space: Non-verbal behaviour recognition and generation," in *4th International Conference on Development and Learning and on Epigenetic Robotics*. IEEE, 2014, pp. 75–80.

[48] J. Li, C. Yang, J. Zhong, and S. Dai, "Emotion-aroused human behaviors perception using rnnpb," in *2018 10th International Conference on Modelling, Identification and Control (ICMIC)*. IEEE, 2018, pp. 1–6.

[49] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.

[50] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[51] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *arXiv preprint arXiv:1905.02947*, 2019.

[52] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow *et al.*, "Subspace gaussian mixture models for speech recognition," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4330–4333.

[53] B. H. Juang and L. R. Rabiner, "Hidden markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

[54] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion." in *ICASSP*, 2013, pp. 6669–6673.

[55] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.

[56] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[57] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[58] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *arXiv preprint arXiv:1905.05709*, 2019.

[59] D. Schlangen and G. Skantze, "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, 2009, pp. 710–718.

[60] S. Kopp, L. Gesellensetter, N. C. Krämer, and I. Wachsmuth, "A conversational agent as museum guide–design and evaluation of a real-world application," in *International Workshop on Intelligent Virtual Agents*. Springer, 2005, pp. 329–343.

[61] M. Marge, S. Nogar, C. Hayes, S. Lukin, J. Bloecker, E. Holder, and C. Voss, "A research platform for multi-robot dialogue with humans," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 132–137.

[62] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3674–3683.

[63] R. Hu, D. Fried, A. Rohrbach, D. Klein, K. Saenko *et al.*, "Are you looking? grounding to multiple modalities in vision-and-language navigation," *arXiv preprint arXiv:1906.00347*, 2019.

[64] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[65] T. L. Q. Dang, N. T. V. Tuyen, S. Jeong, and N. Y. Chong, "Encoding cultures in robot emotion representation," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 547–552.

[66] A. Beck, L. Cañamero, and K. A. Bard, "Towards an affect space for robots to display emotional body language," in *19th International symposium in robot and human interactive communication*. IEEE, 2010, pp. 464–469.

[67] N. Ambady, M. Weisbuch, A. Calder, G. Rhodes, M. Johnson, and J. Haxby, "On perceiving facial expressions: The role of culture and context," *Oxford handbook of face perception*, pp. 479–488, 2011.