Evolving Roles of Intelligent Robots in Social and Industrial Domains

Giuseppe De Simone[1,*,†], Pasquale Foggia[1,†], Francesco Rosa[1,*,†], Alessia Saggese[1,†] and Mario Vento[1,†]

[1]*Department of Information Engineering, Electrical Engineering, and Applied Mathematics (DIEM), University of Salerno*

Abstract

The development of intelligent robotic systems that can interact naturally with humans and adapt to dynamic environments is a central goal of modern robotics. This paper presents ongoing research at MIVIA Lab aimed at addressing these challenges in both social and industrial contexts. In the social domain, the focus is on enabling humanoid robots to engage in meaningful interactions through speech understanding, emotional recognition, and autonomous behavior. In the industrial domain, the emphasis is on developing flexible, task-adaptive robots capable of collaborating with human workers and performing effectively in unstructured environments. Some videos showing the activities conducted are available on the project website sites.google.com/evolving-roles-of-intelligent-robots.

Keywords

Human-Robot Interaction, Learning from Demonstration, Social Robotics, Autonomous Behaviors

1. Introduction

The current trend in robotics focuses on developing systems that exhibit intelligent behaviors. This requires robots to interact seamlessly with humans [1], engaging in conversations, interpreting speech, and recognizing emotions. Robots must also perceive and interact with their environment, identifying and manipulating relevant objects to complete tasks.

These capabilities are valuable in both social and industrial settings. In social contexts, the primary role of a social robot is to assist humans through intelligent and cognitively meaningful interactions and communication. To fulfill this role, a social robot must be capable of analyzing its surrounding environment, with particular attention to the individuals present, and dynamically adapting its behavior based on contextual information. A fundamental capability in this regard is the implementation of cognitively and empathically driven dialogues, enabled by the analysis of soft-biometrics [2] data such as gender, emotional state, and age of the interlocutors. Owing to these intelligent and socially aware features, social robots are increasingly deployed across various domains. In retail environments, they assist customers in shopping malls; in healthcare, they provide support to elderly individuals and children with autism; and at public events, they serve as interactive staff members.

These advanced systems significantly surpass traditional, non-cognitive robots, often referred to as "dumb" robots, which lack the ability to interpret or adapt to human behavior and real-world scenarios. In industrial settings, robots are as well nowadays expected to go beyond repetitive, pre-programmed tasks. A key challenge is adapting to Industry 4.0, where robots must handle diverse tasks and collaborate directly with human workers in dynamic, often unstructured environments. Unlike traditional factories with fixed processes, modern industry demands flexibility, adaptability, and minimal operational constraints [3, 4]. This shift highlights the need for intuitive, user-friendly programming methods, allowing workers to teach robots new tasks quickly and efficiently.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*Corresponding author.

†These authors contributed equally.

gidesimone@unisa.it (G. D. Simone); pfoggia@unisa.it (P. Foggia); frosa@unisa.it (F. Rosa); asaggese@unisa.it (A. Saggese); mvento@unisa.it (M. Vento)

One promising approach is inspired by human learning, where individuals acquire skills by observing others. In robotics, task demonstration offers a powerful paradigm [5]: a human performs a task (e.g., 'pick up the green box and place it in the first bin'), and the robot must autonomously infer and replicate the action. This involves key steps: recognizing and interpreting the demonstration, localizing relevant objects in its own (possibly different) environment, and planning and executing the task accurately. Learning from demonstration has strong potential to enhance robot autonomy and simplify programming, making intelligent robotic systems more practical and accessible across social and industrial domains.

In this paper we are going to describe how at MIVIA Lab these challenges are going to be address. Specifically in Section 2, we are going to describe all the research activities related to social applications, while in Section 3, we are going to describe those research activities related to industrial applications. For additional information and videos, refer to the project website sites.google.com/evolving-roles-of-intelligent-robots.

2. Artificial Intelligence for Social Robotics

A social robot can be described as a "human-made technology that can take physical or digital form, resemble people in form or behavior to some degree, and is designed to interact with people" [6]. According to this definition, and as outlined in Section 1, it is essential for a social robot to emulate human behavior by extracting contextual information from both the surrounding environment and the user. This information must then be leveraged to enable communication with interlocutors that is effective and cognitively appropriate, primarily through the implementation of personalized and context-aware dialogues. In light of these considerations, an important and not negligible feature is equipping the robot with soft-biometric analysis capabilities [2], including gender and emotion recognition, and age estimation, so as to enable empathic and customized dialogues by adapting the generated robot's responses based on the specific characteristic of the interlocutor [7]. Notably, the proposed approach received positive feedback in terms of human–robot interaction, contributing to increased user satisfaction and engagement. While engaging in dialogue with a single user can be effectively handled by state-of-the-art conversational systems, managing conversations in multi-user scenarios remains a complex challenge. Humans are naturally able to estimate suitable interlocutors for interaction and to manage conversational turns [8]. Therefore, it is crucial for social robots to acquire similar capabilities. In this regard, in [9] we proposed a multimodal engagement system that combines both visual and verbal cues to estimate potential interlocutors, such as users who are speaking and making eye contact with the robot, and to manage conversational turns among them. The system integrates the high precision of convolutional neural networks for candidate detection with the flexibility of rule-based classifiers, such as behavior trees [10]. The proposed framework was evaluated using standard benchmark datasets, including the widely adopted UE-HRI database [11], and demonstrated promising results in terms of quantitative performance metrics. Specifically, the system achieved a recall of **77%**, a precision of **88%**, and an F1-score of **82%**. Furthermore, real-world testing was conducted during a public event in Italy, the *Maker Faire*, where users interacted with the social robot *Pepper*. At the end of the interaction, participants completed a survey using a 5-point Likert scale to assess the perceived cognitive abilities of the robot and the quality of the interaction. The results confirmed the system's effectiveness, with high scores reported for the robot's cognitive performance, thereby supporting the quantitative analysis and validating the impact of the proposed approach in real-world human–robot interaction scenarios.

Futhermore, in recent years, social robots have increasingly been employed to support elderly individuals, often functioning as robotic companions that yield a range of positive outcomes, including reductions in loneliness, anxiety, stress, and depressive symptoms. In this context, particular attention has been given to the role of so-called *casual games* [12], which, when integrated with social robots, have successfully engaged older users, generating the aforementioned benefits. Inspired by this line of work, we extended the capabilities of our social robot, *MIVIA-BOT* [13], which already supports

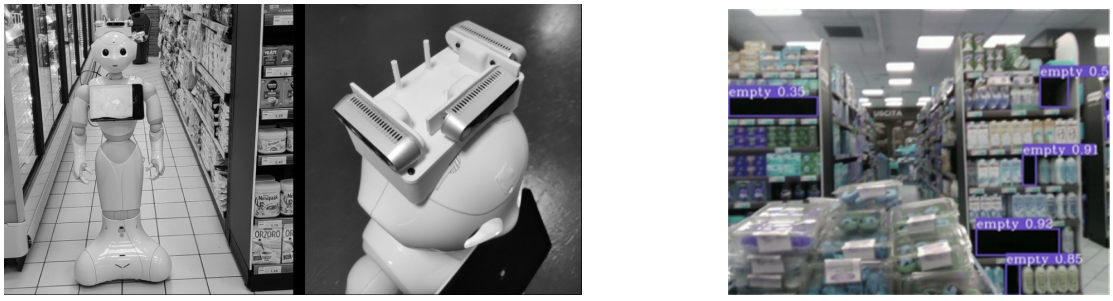
cognitively and empathically rich interactions, by enabling it to play the game "rock, paper, scissors" with users. This enhanced system was evaluated in two public exhibitions in Italy: *Fiera Sicurezza* and *Futuro Remoto*, where it was met with significant interest and engagement from attendees. As in prior evaluations, users interacted with the social robot *Pepper* through an initial conversational phase, during which they could ask personal or factual questions, such as "What is my gender?" or "What is the capital of Italy?" Users could also initiate a game session by requesting to play "rock, paper, scissors" with the robot. At the end of the interaction, participants were asked to complete a questionnaire based on a 5-point Likert scale to assess the quality of the experience (Table 1). In this case as well, the results indicated positive outcomes in terms of human-robot interaction. Specifically, the introduction of the proposed system contributed to an improvement in the overall quality of the interaction, enhancing user satisfaction.

Table 1
Summary of the survey responses for each question, including statistics such as mean, standard deviation (std), minimum (min), maximum (max).

	Q1	Q2	Q3	Q4	QG1	QG2	QF
mean	4,34	4,13	3,85	4,39	4,65	4,45	4,25
std	0,76	0,82	0,89	0,82	0,65	0,76	0,71
min	2	2	1	1	3	2	2
max	5	5	5	5	5	5	5

Finally, with a social robot we also addressed the problem of out-of-stock detection in retail environments [14]. Inventory management is a critical activity within supermarkets, and while automated systems are increasingly adopted for this purpose, sensor-based solutions may become costly over the long term. Although computer vision-based systems are a viable alternative, covering the entire area of interest using fixed cameras can be a challenging task due to installation complexity and limited coverage. In this context, cameras mounted on mobile platforms provide a more flexible and scalable solution. Social robots are particularly well-suited to this task, as their human-like form and behavior enable them to operate within customer-facing environments without causing disruption to shoppers. In [15], we employed the mobile robot *TurtleBot 3*, equipped with a set of sensors, to perform initial mapping of a supermarket environment. Subsequently, the social robot *Pepper* , equipped with a head-mounted camera array Figure 1a, navigated autonomously through the mapped environment to perform out-of-stock detection using dedicated computer vision algorithms as illustrated in Figure 1b, showing the algorithm in action.

Figure 1: Figure (a) shows Pepper equipped with three head-mounted cameras, allowing it to capture images of both the shelves and the corridors while autonomously navigating the store aisles. Figure (b) illustrates the real-time out-of-stock (OOS) detection algorithm in action, where empty shelf areas are highlighted with red bounding boxes, each accompanied by the corresponding prediction confidence score.



(a) (b)

In addition, the system leveraged Pepper's front camera to detect the presence of users within store aisles and dynamically adapted its navigation behavior to avoid disturbing customers during their shopping experience. The proposed approach demonstrated promising results, outperforming human

operators in the same task.

3. Learning by Examples in Industrial Robots

As noted in Section 1, a key goal in robotics is to develop autonomous systems that can perform diverse manipulation tasks based on specific commands, often involving varying objects and dynamic environments. In traditional industrial settings, robots rely on fixed, manually programmed control-rules as the performed tasks are repetitive, and objects and positions are predetermined in static workspaces. In contrast, modern social and advanced industrial robots require greater flexibility: robots now work alongside humans, responding to commands and adapting in real time [16].

To address these challenges, research has focused on data-driven methods, especially Reinforcement Learning (RL) [17] and Imitation Learning (IL) [18]. While promising, these approaches traditionally produce task-specific policies with no mechanism for task selection, limiting adaptability [19].

To improve generalization, Multi-Task Learning [20] trains agents on a variety of tasks (e.g., pick-and-place, nut assembly), each with multiple variations (e.g., different objects and placements). A central challenge is specifying tasks and their variations. Inspired by human learning, two main modalities have emerged: (1) natural language descriptions [21] and (2) visual demonstrations [5]. Natural language provides explicit textual descriptions, e.g., "pick [object1] and place in [object2]", defining both the skills and objects. Visual demonstrations, by contrast, offer sequences of frames showing the task, often performed by different agents in different environments. In both cases, the core challenge is grounding the input, linking commands or demonstrations to specific objects and regions in the robot's environment.

Within this context, our research focuses on video-conditioned systems, which offer a natural and intuitive way to program robotic behavior. This approach allows the demonstrator to convey not only what task should be performed but also provide cues on how to execute it. In particular, we observed that existing methods [22, 23] often struggle in complex scenarios, such as environments containing multiple similar objects without clear indications of which task variation to perform. For example, in Figure 2, given the same initial state, the agent receives no explicit cue about which object to interact with. Our research specifically addressed the problem of target object misidentification [24]. We found that while existing methods could control the robot to reach and manipulate objects, they frequently selected the wrong one.

To tackle this challenge, we propose a novel architectural approach [25] that diverges from traditional end-to-end models. Instead, our method uses a modular architecture with two main components: • **Command Analysis Module**: Identifies and localizes the region of interest within the environment based on the video demonstration. • **Control Module**: Generates precise actions focused on the localized region to ensure accurate object interaction. A key innovation of our approach is the Conditioned Object Detector (COD), which localizes objects of interest within the scene using the visual demonstration as input (Figure 2). Unlike category-based detectors that identify objects along with their predefined categories (e.g., "blue box," "red box," "blue nut"), COD operates in a category-agnostic manner. It focuses solely on detecting the specific objects manipulated in the demonstration, regardless of their category. Critically, the meaning and relevance of the object are defined purely by its role in the demonstration video.

We evaluated the proposed system in a simulated environment under both single-task multi-variation and multi-task multi-variation training scenarios, and also validated it with real-world experiments. Table 2 presents the simulation results. Overall, the modular system, namely **COD-Policy**, built on top of the MOSAIC baseline, demonstrates an average improvement of **+28.78%** in the single-task scenario and **+33.86%** in the multi-task scenario. For the real-world validation (Figure 3), on the Pick-Place task with 6 variations, we achieved an average reaching rate of 86.67% and an average success rate of **55.00%**, marking a significant improvement over the baseline, which reported a 0.00% success rate.

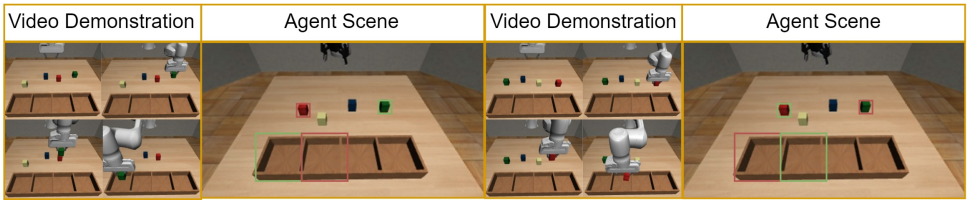


Figure 2: An example illustrating the concept behind the Conditioned Object Detector: Based on the provided video demonstrations, the system classifies objects of interest as target objects (marked with green bounding boxes) and objects to ignore as non-targets (marked with red bounding boxes). For the same object configuration in the agent scene, the meaning of the objects changes based on the video demonstration.

Table 2
Performance comparison (all measures in %) of MOSAIC and COD-Policy across different tasks and variations. The left block shows results for Single-Task Multi-Variation training, while the right block reports results for Multi-Task Multi-Variation training. Metrics include Reaching, Success, and Success with Wrong Object, averaged over multiple trials with standard error.

Task	Model	Reaching	Success	Success with Wrong Object	Reaching	Success	Success with Wrong Object
Pick-Place (16 variations)	MOSAIC	62.90±0.95	58.75±1.87	37.71±0.72	25.83±1.30	22.71±0.72	67.92±1.30
	COD-Policy	100.00±0.00	**93.33±0.72**	0.00±0.00	100.00±0.00	**89.58±3.55**	0.00±0.00
Nut-Assembly (9 variations)	MOSAIC	38.89±1.11	33.33±1.11	51.48±2.31	32.96±1.28	28.53±2.31	41.67±1.66
	COD-Policy	100.00±0.00	**81.11±3.84**	0.00±0.00	99.63±0.64	**70.74±4.49**	0.00±0.00
Stack-Block (6 variations)	MOSAIC	60.56±0.96	53.33±1.66	36.11±0.96	57.78±1.92	55.56±2.54	41.11±1.67
	COD-Policy	100.00±0.00	95.00±1.66	0.00±0.00	98.33±0.00	**85.00±5.00**	0.00±0.00
Press-Button (6 variations)	MOSAIC	100.00±0.00	**100.00±0.00**	0.00±0.00	78.33±1.66	**77.22±2.54**	22.78±2.54
	COD-Policy	100.00±0.00	91.11±1.92	0.00±0.00	83.89±2.55	71.67±2.88	0.00±0.00

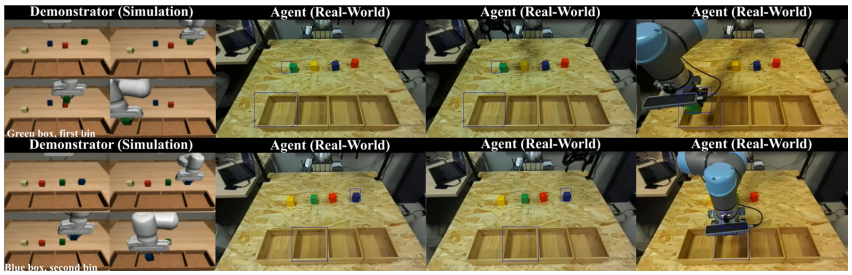


Figure 3: Example of real-world rollouts. The top row illustrates the agent placing the green box into the first bin, while the bottom row shows the blue box being placed into the second bin. It is important to note how the COD module is able to localize objects of interest despite the fact that the demonstrator is a simulated robot.

4. Acknowledgements

This work has been partially supported by MIUR PRIN – PNRR, No. E63C22002150007 - MUMBLE: MUltimodal, Multitask, unBalanced Learning Environments

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT 4.1 in order to: Grammar and spelling check or rephrasing. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Mori, The uncanny valley, IEEE Robotics & Automation Magazine 19 (2012) 98–100.

[2] P. Foggia, A. Greco, A. Saggese, M. Vento, Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition, Eng. Appl. Artif. Intell. 118 (2023).
[3] S. Bini, G. Percannella, A. Saggese, M. Vento, A multi-task network for speaker and command recognition in industrial environments, Pattern Recognition Letters 176 (2023) 62–68.
[4] S. Bini, V. Carletti, A. Saggese, M. Vento, Robust speech command recognition in challenging industrial environments, Computer Communications 228 (2024) 107938.
[5] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, S. Levine, One-shot imitation from observing humans via domain-adaptive meta-learning, in: 6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings, 2018.
[6] J. Fox, A. Gambino, Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human-Robot Interaction, Cyberpsychology, behavior and social networking (2021).
[7] P. Foggia, A. Greco, A. Roberto, A. Saggese, M. Vento, A social robot architecture for personalized real-time human–robot interaction, IEEE Internet of Things Journal 10 (2023) 22427–22439.
[8] G. Skantze, Turn-taking in Conversational Systems and Human-Robot Interaction: A Review, Computer Speech & Language 67 (2021) 101178.
[9] G. De Simone, L. Greco, A. Saggese, M. Vento, Robot to human interaction with multi-modal conversational engagement, in: SIMPAR, 2025, pp. 1–8.
[10] M. Colledanchise, P. Ögren, Behavior Trees in Robotics and AI: An Introduction, CRC Press, 2018.
[11] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, A. Lim, UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions, in: ACM International Conference on Multimodal Interaction, New York, NY, USA, 2017, pp. 464–472.
[12] H. Groenewoud, J. de Lange, Y. Schikhof, A. Astell, P. Joddrell, M. Goumans, People with dementia playing casual games on a tablet, Gerontechnology 16 (2017) 37–47.
[13] G. De Simone, A. Saggese, M. Vento, Empowering human interaction: A socially assistive robot for support in trade shows, in: IEEE ROMAN, 2024, pp. 901–908.
[14] J. Chen, S.-L. Wang, H.-L. Lin, Out-of-stock detection based on deep learning, in: Intelligent Computing Theories and Application: 15th International Conference, ICIC, 2019, pp. 228–237.
[15] G. De Simone, P. Foggia, A. Saggese, M. Vento, Autonomous mobile robot for automatic out of stock detection in a supermarket, in: IEEE/CVF ICCV Workshops, 2023, pp. 1829–1838.
[16] G. De Simone, A. Greco, F. Rosa, A. Saggese, M. Vento, Context-aware data augmentation for enhanced speech command recognition in industrial environments, Scientific Reports 15 (2025).
[17] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, S. Levine, How to train your robot with deep reinforcement learning: lessons we have learned, The International Journal of Robotics Research 40 (2021) 698–721.
[18] M. Zare, P. M. Kebria, A. Khosravi, S. Nahavandi, A survey of imitation learning: Algorithms, recent developments, and challenges, IEEE Transactions on Cybernetics (2024).
[19] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, P. Abbeel, Deep imitation learning for complex manipulation tasks from virtual reality teleoperation, in: IEEE ICRA, 2018.
[20] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, C. Finn, Bc-z: Zero-shot task generalization with robotic imitation learning, in: Conference on Robot Learning, PMLR, 2022, pp. 991–1002.
[21] D. Driess, et al., Palm-e: an embodied multimodal language model, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.
[22] S. Dasari, A. Gupta, Transformers for one-shot visual imitation, in: Conference on Robot Learning, PMLR, 2021, pp. 2071–2084.
[23] Z. Mandi, F. Liu, K. Lee, P. Abbeel, Towards more generalizable one-shot visual imitation learning, in: IEEE ICRA, 2022, pp. 2434–2444.
[24] P. Foggia, F. Rosa, M. Vento, Enhancing robotic demonstration-based learning method with preliminary visual target localization, in: European Robotics Forum, Springer, 2024, pp. 212–217.
[25] P. Foggia, F. Rosa, M. Vento, Improving learning from visual demonstration methods by target localization, in: IEEE ROMAN, 2024, pp. 740–747.