Baština Akademije nauka i umjetnosti Bosne i Hercegovine

The Industry of the Future: From Industry 4.0 to Industry 5.0 – Integration of Humans and Technology: New Technologies

Karabegović, Isak

2025

Akademija nauka i umjetnosti Bosne i Hercegovine

https://bastina.anubih.ba/handle/123456789/837

Preuzeto s Baštine Akademije nauka i umjetnosti Bosne i Hercegovine

https://bastina.anubih.ba/

DOI: 10.5644/PI2025.220.04

Enhancing Human-Robot Collaboration Through Multimodal Data and Robot Learning in Human-Centered Industry 5.0 Systems

Lejla Banjanović-Mehmedović*[1]

Abstract: *The transition from Industry 4.0 to Industry 5.0 emphasizes the development of human-cantered robotic systems designed for seamless and adaptive collaboration with human operators in complex industrial environments. Advances in multimodality within human-robot collaboration (HRC) are enabling richer and more natural interactions by leveraging diverse communication channels, including auditory inputs (speech recognition), visual perception (RGB-D and depth cameras), gestural understanding (pose estimation), haptic feedback, and even brain-computer interfaces. Moreover, the capability to understand human behaviour, particularly in terms of intent prediction and motion analysis, plays a critical role in fostering mutual human-robot assistance.*
This review provides a comprehensive analysis of state-of-the-art approaches that integrate multimodal data with advanced robot learning paradigms, such as imitation learning and reinforcement learning. Within this context, the study presents an overview of the current landscape of multimodal HRC and its applications while outlining key challenges and future research directions.

Keywords: *Human-Robot Collaboration, Industry 5.0, Multimodal Data, Robot Learning*

1. Introduction

The industrial landscape is undergoing a profound transformation, shifting from the automation-centric paradigm of Industry 4.0 to the human-cantered vision of Industry 5.0. While Industry 4.0 focused on digitalization, cyber-physical systems, and interconnected smart factories, Industry 5.0 emphasizes the synergistic collaboration between humans and intelligent machines, promoting values such as sustainability, resilience, and personalization. This evolution demands not only smarter technologies but also adaptive, intuitive, and cooperative robotic systems that can operate seamlessly alongside human workers.
At the core of this transition is Human-Robot Collaboration (HRC), a paradigm in which humans and robots share physical workspaces, coordinate actions, and jointly complete tasks in dynamic and often unstructured environments.

*[1]University of Tuzla, Faculty of Electrical Engineering, Tuzla, Bosnia and Herzegovina
E-mail: lejla.banjanovic-mehmedovic@fet.ba

Technological advancements in artificial intelligence, robotics, soft materials, and bioelectronics are key factors in enabling human-robot collaboration in industrial environments[1].
A key enabler of such capabilities lies in the integration of multimodal data - including vision, speech, gesture, and contextual information—with robot learning methods such as self-supervised learning,imitation learning, and reinforcement learning. Multimodal perception allows robots to interpret complex human communication cues, while learning algorithms enable them to improve and personalize their behaviour over time.
This review aims to provide a comprehensive overview of recent advances in the intersection of multimodal sensing and robot learning for HRC in the context of Industry 5.0. We discuss existing frameworks, representative applications, and open challenges, with a focus on creating robotic systems that are not only functional, but also trustworthy, adaptive, and human-aware.
The remainder of this study is organised as follows. Section 2 provides key characteristics of human-robot collaboration and discusses real-world applications.Challenges and open issues are presented in Section 3. Section 4presents an overview of input modalities, followed by a discussion on sensor fusion and real-time data processing. Section 5 presents the key components of effective human–robot collaboration, emphasizing how cognitive perception, safe and adaptive actions, and continuous learning enable robots to understand human intent, plan cooperative behaviours, and improve performance in dynamic environments. The discussion of future directions is presented in Section 6, outlining emerging trends, unresolved challenges, and potential advancements that could enhance the effectiveness and adaptability of human–robot collaboration systems.Finally, Section 7 concludes this study.

2. Human-Robot Collaboration in Industry 5.0

Human-robot collaboration has emerged as a transformative approach in modern manufacturing and logistics, enabling the seamless integration of human expertise with robotic precision and efficiency. By working side by side with humans, collaborative robots (cobots) address the limitations of fully automated systems and empower flexible, adaptive workflows, Figure 1.This synergy is particularly valuable in industries characterized by high variability, customization, and the need for quick reconfiguration of production lines[2].
Collaborative robots are often used in industrial processes, such as assembly, packaging, inspection, or the transportation of items from conveyor belts. The key advantages of human-robot collaboration include:

- ***High level of automation.*** Cobots complement human workers' capabilities and enable rapid automation of production steps.

- ***Reduced employee workload.*** Physically demanding, dangerous, and monotonous tasks can be taken over by cobots, thereby relieving workers.
- ***High quality.*** Repetitive processes that require high concentration are executed by cobots with maximum precision, improving production quality.
- ***Maximum flexibility.*** Collaborative robot tasks can be flexibly adapted to changing requirements.

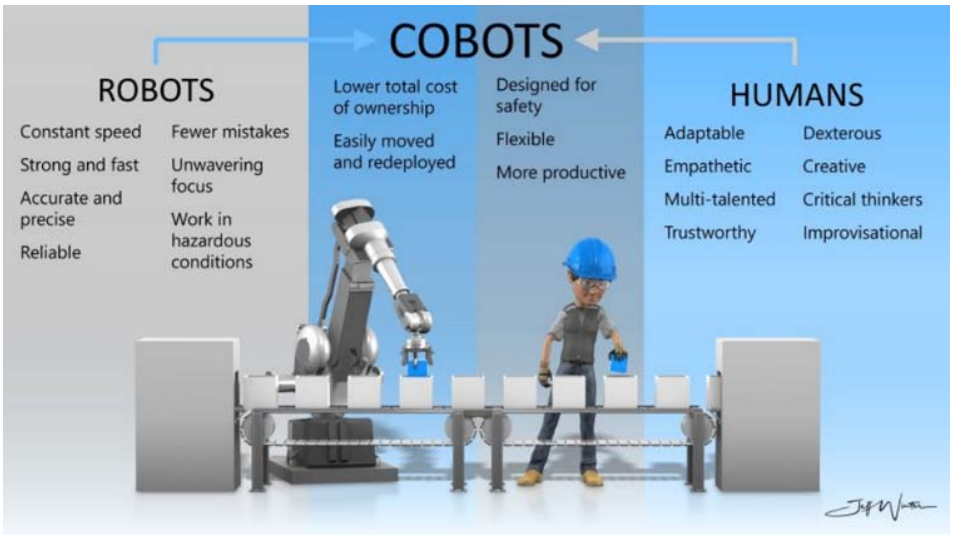


Figure 1. Traditional Robots vs. Collaborative Robots in Human–Robot Collaboration[3].

The types of human-robot cooperation are illustrated in Figure 2. They encompass varying levels of interaction, each defining different degrees of shared tasks and communication between humans and robots[4]:

- **Coexistence:** Humans and robots do not share the same workspace and operate independently on different tasks.
- **Cooperation:** In human-robot cooperation, humans and robots work in the same workspace, alternately performing different tasks within the process. There is no direct interaction.
- **Collaboration:** Humans and robots interact within a shared workspace; both work simultaneously on the same product.

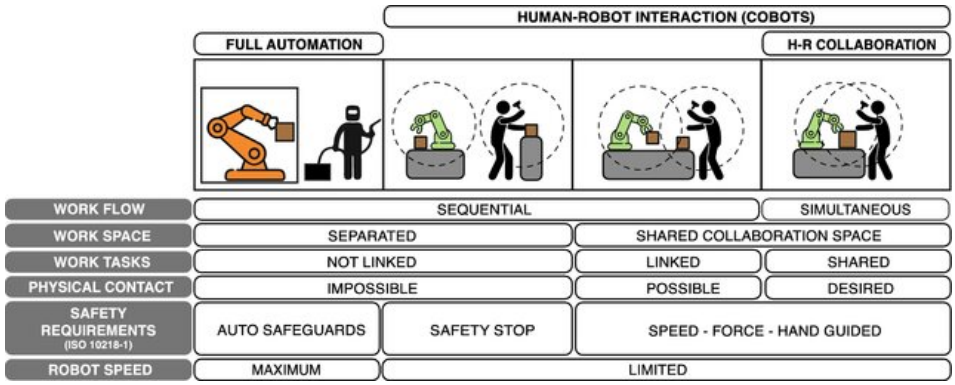


Figure 2. Typesof human-robot cooperation: coexistence, cooperation and collaboration[5,6].

In industrial assembly, HRC enhances productivity by allowing cobots to handle repetitive or ergonomically challenging tasks while humans focus on high-skill operations such as quality adjustments, problem-solving, and fine-tuning components[7]. Cobots equipped with force-torque sensors and advanced vision systems can assist with tasks like screwing, welding, and component placement, ensuring precision and consistency. For example, in automotive manufacturing, robots can preassemble complex parts or hold components in position while a human operator performs intricate manual tasks, resulting in improved efficiency and reduced cycle times.

Collaborative robots are increasingly deployed in inspection processes due to their ability to combine advanced computer vision algorithms with consistent, fatigue-free operation. In electronics manufacturing or aerospace industries, cobots can perform real-time visual inspections, detect micro-defects, or use non-destructive testing (NDT) techniques such as ultrasonic or laser scanning. When combined with human oversight, this ensures both speed and accuracy in identifying production anomalies. Through multimodal sensing, cobots can detect quality deviations, while humans make final judgments that require domain expertise or creativity[8].

In logistics, warehouse automation, and intralogistics, cobots streamline workflows by transporting goods, sorting items, and assisting with order fulfillment. Unlike traditional automated guided vehicles (AGVs), mobile cobots can operate safely alongside human workers in dynamic environments, adapting to changing layouts and workflows. Applications include palletizing, depalletizing, and collaborative picking, where robots lift heavy loads while humans handle delicate tasks such as product verification and packaging[9]. This is particularly beneficial in e-commerce and retail distribution centers, where demand for fast and flexible order processing is high.

An example of human-robot collaboration in an industrial environment is shown in Figure 3**.**

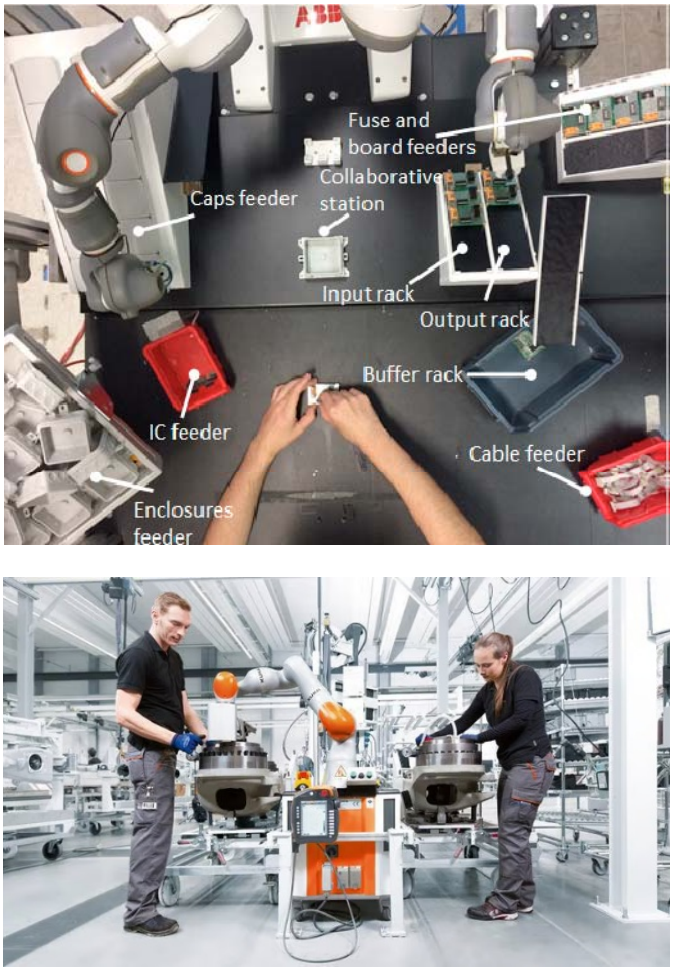


Figure 3. Examples of human-robot collaboration in industrial settings[10,11].

3. Challenges and Open Issues

Developing effective collaborative systems is a complex challenge that involves technical, cognitive, and design-oriented considerations. Key challenges include perception, communication, learning, optimization, and explainability, each of which directly impacts the safety, efficiency, and trustworthiness of HRC systems[12,13].

- ***Perception.*** Accurate perception of the environment is a fundamental requirement for successful collaboration. Robots must be capable of recognizing and classifying objects, understanding spatial relationships, and tracking dynamic elements in real time. Beyond object recognition, advanced perception includes interpreting **human activities, intentions, and emotional states** using visual cues such as facial expressions, body posture, and gestures. This requires robust multimodal sensing, integrating data from cameras, LiDAR, tactile sensors, and other devices, combined with deep learning algorithms for reliable human-robot interaction in unstructured and dynamic environments.
- ***Communication.*** Effective communication between humans and robots is essential for smooth collaboration. This includes both **verbal and non-verbal communication.** Verbal communication relies on **natural language processing (NLP)** for interpreting spoken commands, contextual understanding, and generating human-like responses. Non-verbal communication, such as interpreting hand gestures, pointing, or body posture, is equally critical in industrial and service contexts, where verbal instructions may not always be practical. A failure to understand subtle non-verbal cues can lead to inefficiencies or safety risks.
- ***Reactive Control-Based Approaches.*** In collaborative settings, robots must dynamically adjust their behaviour in response to changing conditions. **Real-time feedback mechanisms** allow robots to adapt to unexpected human actions, task changes, or environmental disruptions. This demands low-latency sensor data processing and control algorithms capable of reactive planning, ensuring both task success and human safety.
- ***Learning.*** Robots in HRC scenarios must be capable of **continuous learning** to adapt to human preferences, new tools, or novel environments. Methods such as **Learning from Demonstration (LfD)**, reinforcement learning, and human-in-the-loop training enable robots to refine their behaviours and tasks based on feedback. Continuous learning not only improves adaptability but also enhances the robot's ability to share workloads effectively with human partners.
- ***Optimization.*** Task planning and motion trajectories must be optimized to ensure efficiency and safety during collaboration. This includes minimizing energy consumption, avoiding unnecessary robot movements, and maintaining safe distances from human co-workers. Optimized planning also accounts for task sequencing and resource allocation, which are vital for high-mix, low-volume manufacturing or logistics environments.
- ***Robot Design.*** The physical design of collaborative robots plays a crucial role in building trust and ensuring usability. Cobots must feature

ergonomic, safe, and intuitive designs, often incorporating lightweight materials, rounded edges, and limited power or force to reduce the risk of injury. User-centric design, combined with intuitive interfaces, ensures that operators can work with robots without extensive training or fear.

- ***Explainable Robotics.*** Transparency is vital for trust in HRC systems. Robots must be capable of explaining their **decisions, actions, and reasoning** in ways that are understandable to human collaborators. Explainable robotics not only enhances user confidence but also improves debugging, safety verification, and regulatory compliance in industrial settings.

4. Muti Modal Modalities in HRC

Most methods for learning robotic policies emphasize only one modality of task description, failing to exploit the wealth of information offered by cross-modal data. Modern robotic systems are designed to process and integrate data from multiple sensory modalities, such as vision, speech, text,and voice commands as well as tactile and psysiological sensing, Figure 4 [1,14].

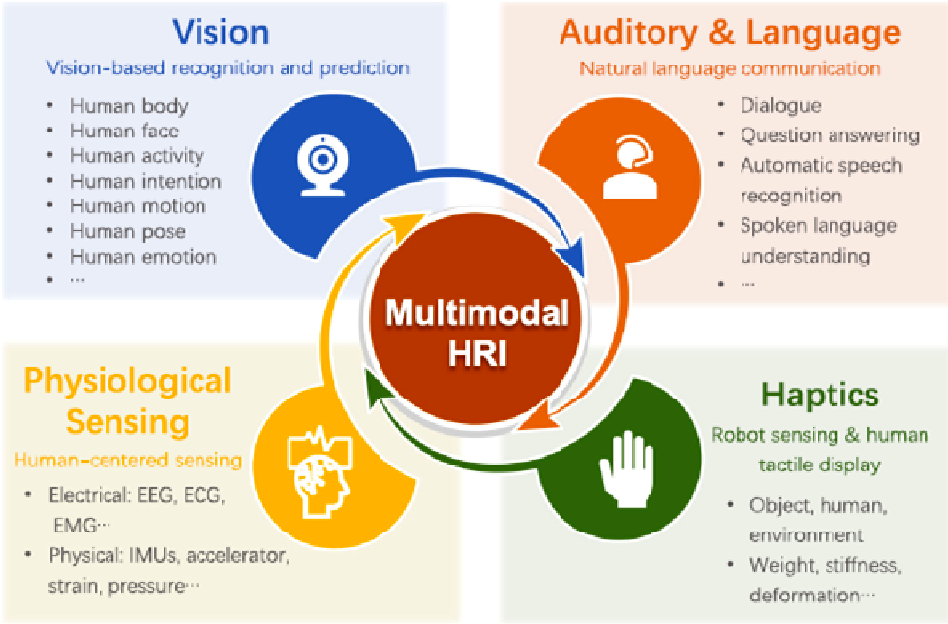


Figure 4. Four typical modalities of Human Robot Interface [1].

For example, combining visual perception with tactile sensing allows for precise manipulation of deformable objects, while coupling language and vision facilitates semantic understanding of tasks described by humans. Furthermore, the integration of physiological or biosignal data (e.g., electromyography or EEG) enables robots to anticipate human intent, optimize ergonomic factors, and respond proactively during collaboration.

Vision-based technologies in Human-Robot Collaboration (HRC) rely on advanced computer vision and deep learning techniques to detect and analyze human position, activity, pose, and emotion [1]. Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs), and Vision Transformers (ViTs) are widely used for tasks such as pose estimation, activity recognition, and facial expression analysis, enabling robots to understand and respond to human behaviour in dynamic environments.An example of multimodal fusion of gesture recognition and object classification using Vision Transformers (ViTs) in human–robot collaboration is presented in Figure 5 [15].

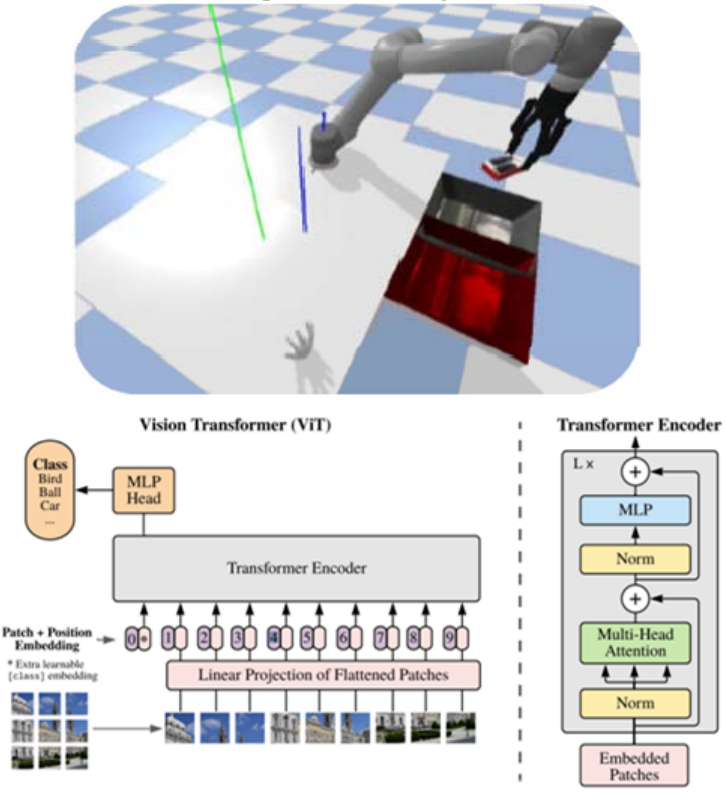


Figure 5. Multimodal Fusion of Gesture and Object Classification in Human-Robot CollaborationusingViTs[15].

Language-based interaction leverages Natural Language Processing (NLP) and machine learning models to enable natural communication between humans and robots. Automatic Speech Recognition (ASR) is powered by deep neural networks (DNNs) and Transformer-based models, while Spoken Language Understanding (SLU) often employs Recurrent Neural Networks (RNNs), BERT-like models, or sequence-to-sequence architectures. Question-answering (QA) and dialogue systems use large language models (LLMs) and reinforcement learning techniques to support context-aware and conversational HRI.

Haptics-based technologies utilize AI algorithms for tactile signal interpretation and adaptive control. Machine learning methods, such as Support Vector Machines (SVMs), Random Forests, or CNNs, are applied to process tactile sensor data for recognizing object properties (e.g., texture, stiffness) and human touch patterns. Reinforcement Learning (RL) is often integrated for dynamic adjustment of haptic feedback and robotic manipulation, ensuring safe and effective physical collaboration in real and virtual environments.

Physiological sensing incorporates AI-driven signal processing and classification techniques to monitor human states using EEG, ECG, and EMG signals [1]. Methods like convolutional and recurrent neural networks (CNNs, LSTMs) or hybrid deep learning models are employed for emotion recognition, stress detection, and cognitive state assessment.

Efficient fusion of multimodal data remains one of the primary challenges in achieving robust performance, particularly in complex tasks such as human–robot collaboration. Various strategies have been proposed to address this issue, with the most used approaches - early fusion, late fusion, and hybrid (middle) fusion[1].

- *Early Fusion:*In early fusion, all modalities (e.g., images, audio, text, or sensor data) are integrated at the input stage of the model, Figure 6a).Typically, feature vectors extracted from different sensors are concatenated into a single high-dimensional vector, which is subsequently processed by a unified neural network (e.g., a combination of CNN and LSTM layers or a transformer-based architecture). The main advantage of early fusion is that it enables the model to directly learn cross-modal correlations. However, this approach can be challenging due to differences in the nature and dimensionality of the input data, potentially leading to training inefficiencies or model overfitting.
- *Late Fusion:*In late fusion, each modality is processed independently through dedicated subsystems (e.g., separate neural networks designed for visual and audio inputs), Figure 6b). The outputs of these subsystems—typically feature embeddings or intermediate decisions—

are then combined at a higher level, either through concatenation of latent vectors or decision-level integration. The primary strengths of late fusion lie in its modularity and flexibility, making it straightforward to add or remove modalities. Nonetheless, it may fail to capture deeper inter-modal correlations that emerge in earlier stages of processing.

- *Hybrid (Middle) Fusion:*Hybrid fusion represents a compromise between early and late fusion. In this approach, individual modalities are first partially processed through their respective encoders, after which the resulting intermediate representations are fused at a middle layer of the network, Figure 6c). This strategy aims to balance the advantages of both early and late fusion, offering improved flexibility while maintaining the ability to learn efficient multimodal representations.
-

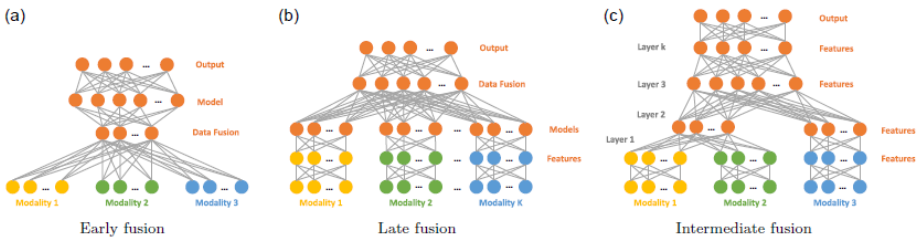


Figure 6. Forms of Multimodal Fusion: a). Early fusion, b). Late fusion; c). Intermediate fusion [1].

Multimodal fusion often relies on specialized architectures that integrate heterogeneous data streams into a unified representation.

a. *Multimodal encoders* process each modality through dedicated networks (e.g., CNNs for visual inputs, RNNs or transformers for audio and text), after which the resulting embeddings are combined via concatenation, attention mechanisms, or element-wise operations.
b. *Attention mechanisms* allow the model to dynamically prioritize modalities based on contextual relevance, improving robustness when certain inputs are noisy or missing.
c. *Gating mechanisms* adaptively regulate the contribution of each modality using learned weights, often implemented as sigmoid-based layers.
d. *Multimodal transformers* are increasingly adopted due to their ability to capture complex inter-modal dependencies through self-attention and hierarchical representation learning.

The choice of method depends on the application, with early fusion suited for strongly correlated data and attention-based approaches excelling in context-aware tasks.

Figure 7 illustrates multimodal learning that integrates voice recognition, hand movement, and human body posture. This approach leverages various deep learning techniques, including convolutional neural networks (CNNs), recurrent neural networks with specialized architectures such as long short-term memory (LSTM), and related transfer learning methods[16].

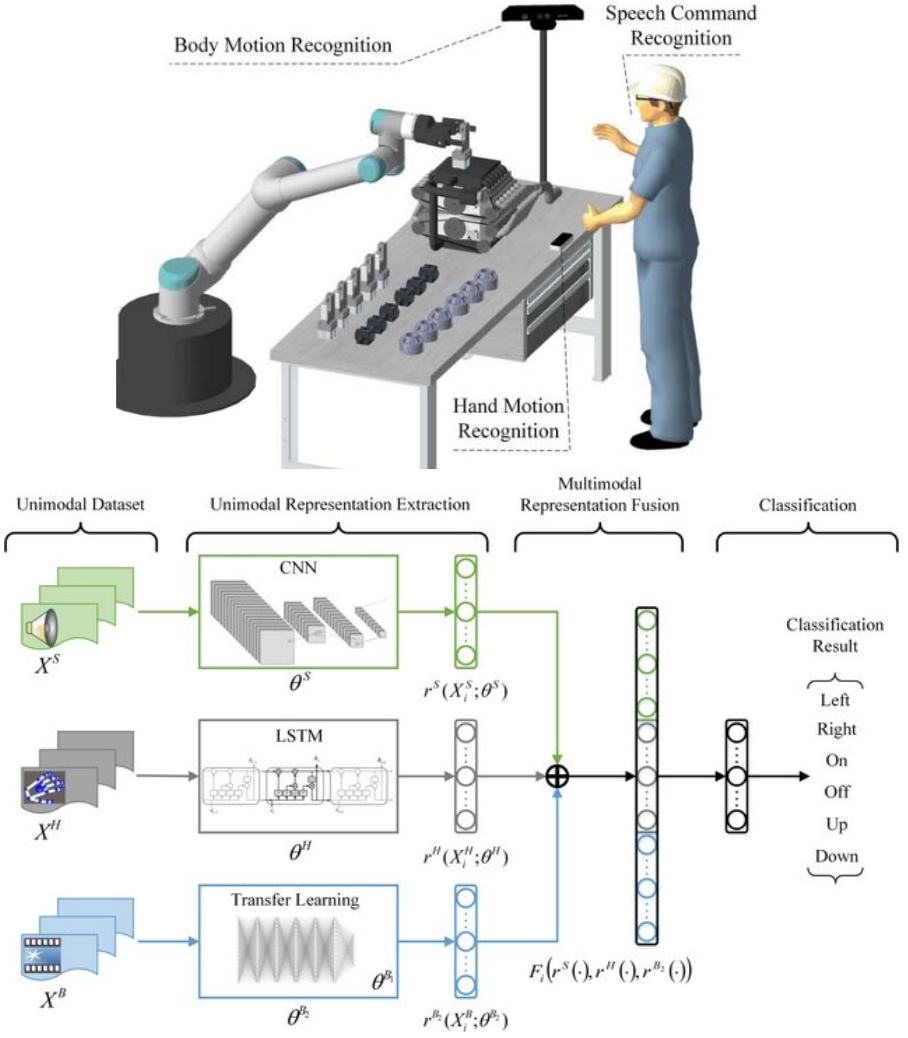


Figure 7. Example of multimodal fusion combining three types of deep learning approaches: (1) voice command recognition, (2) hand position detection, and (3) body pose estimation, enabling comprehensive interpretation of human intent and actions [16].

Figure 8 depicts a human-centered human-robot collaboration (HRC) scenario that leverages multiple modalities to achieve collaborative objectives, structured into four main phases. Initially, within a shared workspace, the human and robotic arm jointly execute a complex assembly task, utilizing both visual and tactile feedback for precise coordination. In the subsequent phase, an autonomous guided vehicle (AGV) moves toward the storage zone to locate necessary materials while engaging with the human operator through a visual-language navigation (VLN) framework that combines visual and auditory cues. The third phase involves the AGV-mounted robotic arm retrieving the identified material and handing it to the human, again relying on visual and tactile modalities. In the final phase, the human operator is equipped with EMG electrodes to enable ergonomics evaluation through integrated physiological sensing.

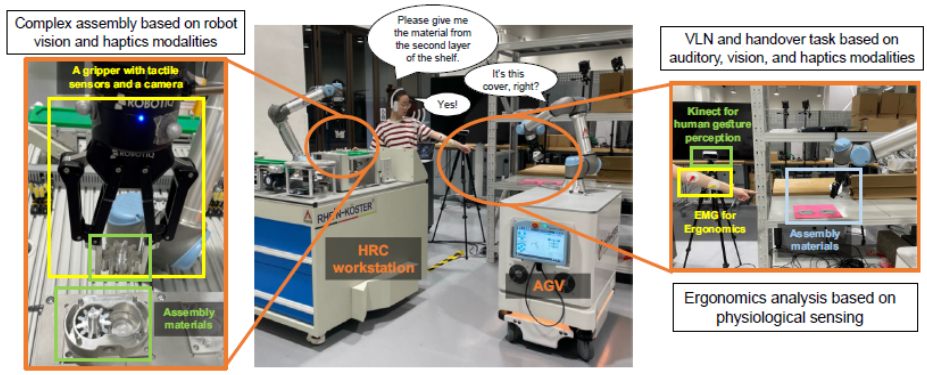


Figure 8. Typical human-centric smart manufacturing application scenario based on multimodal HRI[1].

The core challenge of multimodal approaches with AI in robotic learning lies in the **effective fusion of heterogeneous data modalities**—such as visual input, speech, haptic feedback, and natural language—into a coherent representation that enables **reliable, robust, and adaptive robot behaviour in real-world environments**. Achieving this fusion requires addressing a set of complex, interrelated challenges[17]:

- ***Semantic misalignment between modalities.*** Different modalities (e.g., video and speech) operate on distinct temporal scales, levels of semantic granularity, and structural representations. Achieving temporal and conceptual alignment remains difficult, particularly when interpreting commands such as "grab that," where the robot must correctly link the verbal cue to the visual context (e.g., the object currently detected by the camera).

- ***Fusion of heterogeneous representations.*** Each modality is typically processed by specialized neural architectures—e.g., Vision Transformers (ViT) for vision, LSTMs for speech, and BERT for language. Integrating these diverse feature spaces into a unified, task-relevant representation is inherently nonlinear and computationally demanding.
- ***Dynamic attention and modality selection.*** A robot must be able to dynamically determine which modality to prioritize depending on the context. For instance, speech may dominate in low-noise settings, whereas visual information should take precedence when acoustic signals are unreliable. This necessitates **modality-adaptive attention mechanisms** capable of weighting information in real time.
- ***Lack of large, synchronized multimodal datasets.*** Collecting realistic, high-quality multimodal datasets for robotics is challenging, resource-intensive, and often task- or domain-specific, which limits generalization and transfer learning capabilities.
- ***Real-time learning and adaptability.*** Multimodal policies must achieve efficient, low-latency learning to operate under real-world constraints, especially in reinforcement learning (RL) scenarios where interaction with the environment is costly.
- ***Explainability and transparency.*** It remains difficult to disentangle the contribution of each modality to the final decision, which complicates debugging, performance evaluation, and the establishment of trust in safety-critical domains such as industrial automation or healthcare.
- ***Robustness to modality degradation or failure.*** Ensuring continuous functionality when one or more modalities fail (e.g., camera malfunction, speech input loss) requires the design of **graceful degradation mechanisms** and fallback strategies.

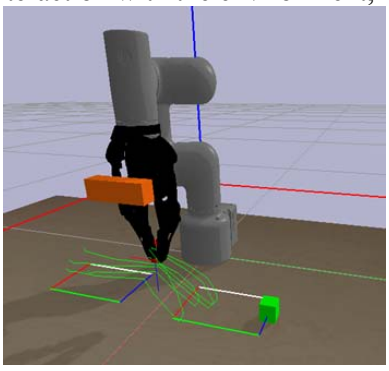
5. Cognition, Action, and Learning in HRC

Following multimodality, cognition, action, and learning are essential components of effective human–robot collaboration (HRC). *Cognition* involves a robot's ability to perceive, reason, and interpret human behaviour and environmental context. This includes understanding task goals, predicting human intentions, and adapting to dynamic conditions. Cognitive models often integrate multimodal perception with semantic reasoning, probabilistic inference, or graph-based representations to create a human-aware decision-making process [18]. For example, recognizing gestures, gaze direction, or speech commands can provide contextual cues that guide collaborative tasks.
Action refers to the generation of safe and contextually appropriate behaviours based on cognitive insights. In collaborative environments, this includes motion

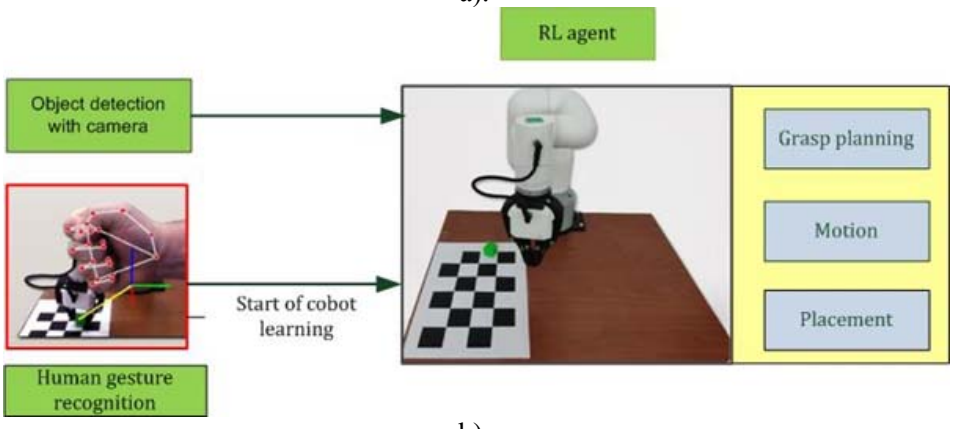
planning, trajectory adaptation, and compliant control to ensure fluid cooperation with humans. Modern approaches leverage real-time feedback loops and optimization algorithms, enabling robots to perform tasks such as handovers or co-manipulation in a natural and intuitive manner [19]. Shared autonomy strategies and dynamic re-planning ensure that robots remain responsive to human inputs and environmental changes.

Learning allows robots to improve their performance over time, adapting to novel situations and user preferences. *Learning from demonstration (LfD)*or *imitation learning*provides a direct way to acquire skills by observing human behaviour [20,21]. This paradigm facilitates fast adaptation in flexible manufacturing and has proven effective for trajectory generalization and task sequencing.

In contrast, *reinforcement learning (RL)* allows robots to learn optimal policies through trial-and-error interaction with the environment, Figure 9.



a).



b).

Figure 9. Training process of a robotic agent using DRL within the PyBullet environment to develop robust pick-and-place strategies, emphasizing the integration of policy refinement and simulation-to-real transfer techniques [22].

While RL has demonstrated success in collaborative assembly and manipulation tasks, its application in HRC requires safety-aware mechanisms to mitigate risks during exploration [22]. In human-robot collaboration systems for assembly tasks, reinforcement learning (RL) and deep reinforcement learning (DRL) methods are increasingly being utilized. A collaborative reinforcement learning approach was applied to evaluate the use of a fixed-arm robot to determine the optimal strategy for emptying the contents of a plastic bag [23]. Furthermore, *online and incremental learning* enables robots to update their models during operation, allowing them to adapt to evolving human behaviour, tools, or workflows—critical for unpredictable industrial settings [24].

To enhance human alignment, *human-in-the-loop learning* integrates feedback and corrections from human operators, allowing real-time adaptation and personalized behaviour shaping [25]. Additionally, *transfer learning and simulation-to-reality (Sim2Real) techniques* are increasingly used to pre-train models in virtual environments, minimizing costly or dangerous real-world trials and improving deployment efficiency[26].

Hybrid frameworks that combine DRL with LfD and transfer learning have shown great promise in reducing sample complexity and enhancing generalization [27]. These methods, when integrated with cognitive reasoning and adaptive action control, create robust HRC systems capable of operating effectively in unstructured, human-centric environments.

5.1. Multimodality-Based Robot Learning

The incorporation of multimodal data significantly strengthens robot learning methodologies. By combining information from various sensory channels, these systems develop a richer understanding of their environment, enabling better generalization to diverse scenarios and the successful execution of complex tasks such as object detection, gesture recognition, and natural language processing. Multimodal fusion enhances both the resilience and flexibility of learning algorithms while promoting more natural and efficient interactions between humans and robots. Consequently, multimodal approaches drive higher levels of autonomy and performance across multiple domains, including industrial automation, assistive technologies, and service robotics.

One example of robot policy learning from multimodal task specifications is presented in Figure 10. It trains a transformer-based architecture to facilitate cross-modal reasoning, combining masked modeling and cross-modal matching objectives in a two-stage training procedure[28]. After training, MUTEX can follow a task specification in any of the six learned modalities (video demonstrations, goal images, text goal descriptions, text instructions, speech goal descriptions, and speech instructions) or a combination of them. This approach systematically evaluated the benefits of MUTEX in a newly designed

dataset with 100 tasks in simulation and 50 tasks in the real world, annotated with multiple instances of task specifications in different modalities, and observed improved performance over methods trained specifically for any single modality.

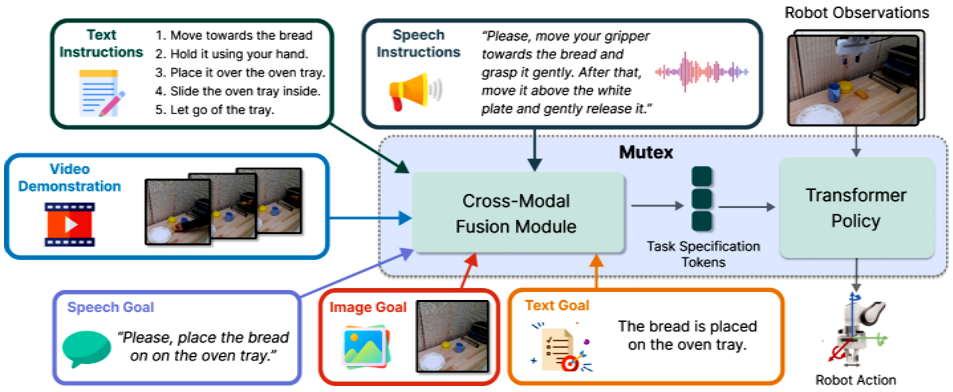


Figure 10. Learning Unified Policies from Multimodal Task Specifications[28].

Everyday tasks involving extensive physical interactions—such as peeling, cleaning, and writing—require robust multimodal perception to ensure accurate and effective execution, Figure 11. For robots, however, such contact-rich tasks pose significant challenges due to their limited capability to integrate and interpret diverse sensory modalities. Existing learning-based approaches for contact-rich manipulation have attempted to address this issue but typically rely on large datasets and task-specific reward functions, which constrain their scalability and generalization. To overcome these limitations, the paper [29]introduced a generalizable, model-free learning-from-demonstration framework **t**hat enables robots to acquire contact-rich skills without the need for explicit reward engineering. A novel multimodal sensor data representation was proposed, enhancing the efficiency and accuracy of the learning process. The framework is validated through experiments on a Sawyer robot across three representative contact-rich tasks: cleaning, writing, and peeling. The results demonstrate a 100% success rate for both peeling and writing, and an 80% success rate for cleaning. These findings indicate that the proposed approach offers a scalable foundation for extending skill acquisition to a wide range of physical manipulation tasks. Humans skillfully manipulate deformable objects by relying on multimodal perception, enabling them to accomplish everyday tasks such as opening bags, unwrapping candy, or retrieving keys from pockets. Transferring these abilities to robots is highly challenging due to the complex and unpredictable properties of deformable materials.

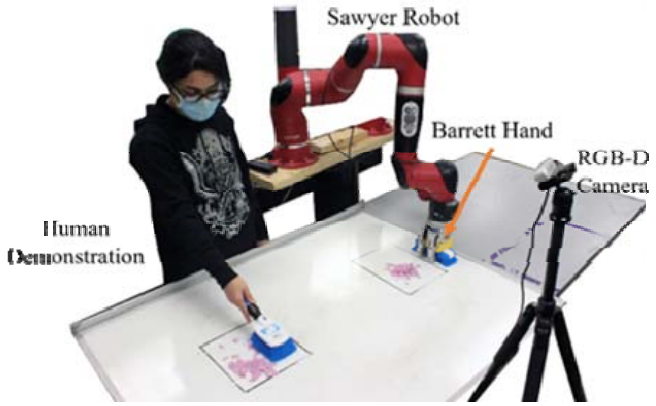


Figure 11. The experimental setup for learning-from-demonstration framework [29].

To tackle this problem, a human-inspired exploration and purposeful manipulation framework has been developed for robots, focusing on multimodal learning and adaptation[30]. As illustrated in Figure 12, the framework enables robots to autonomously explore and learn the characteristics of a class of deformable objects. Using the knowledge acquired during this exploration phase, the robot can execute purposeful manipulations to complete specific tasks.

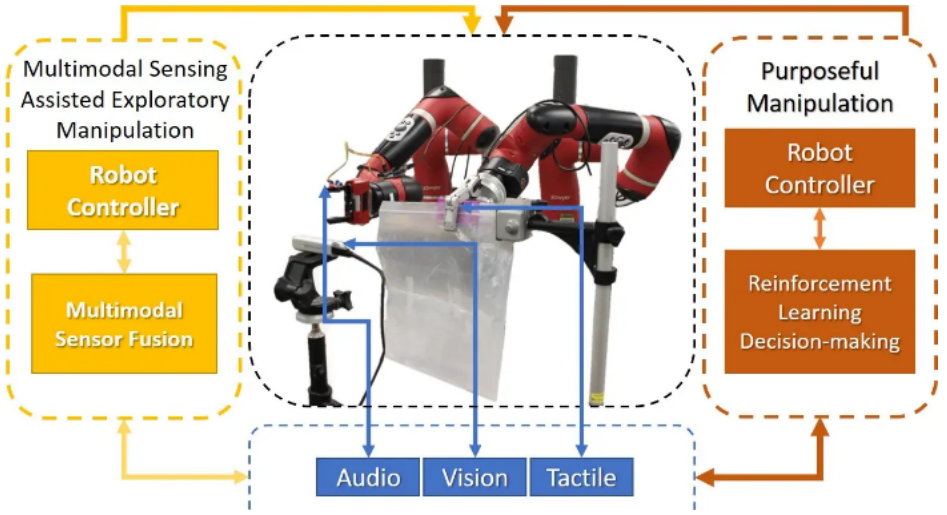


Figure 12. Multimodal Reinforcement Learning and Decision-making [30].

6. Future Directions

Future research in human-robot collaboration (HRC) aims to address the limitations of current systems by enhancing adaptability, intelligence, and trustworthiness in dynamic real-world environments. Emerging trends focus on integrating multimodal perception, large-scale foundation models, and lifelong learning mechanisms to create robots capable of personalized, context-aware, and safe collaboration with humans [14].

- *Personalized and Context-Aware HRC.*Future robotic systems must adapt to individual human users by learning their preferences, expertise levels, and working styles. This requires integration of online learning, user modeling, and context-aware adaptation, enabling robots to act as personalized collaborators rather than generic tools.
- *Continual and Lifelong Learning.*Robots in real-world settings must operate in non-stationary environments where tasks, users, and workflows change over time. Lifelong learning mechanisms will allow robots to incrementally acquire knowledge, avoid catastrophic forgetting, and build richer experience-based policies without exhaustive retraining.
- *Large Language Models and Natural Communication.*The rise of Large Language Models (LLMs) opens new possibilities for natural, high-level interaction between humans and robots. By embedding LLMs into HRC systems, robots can better interpret verbal instructions, generate contextual responses, and support collaborative dialogue, paving the way toward semantic-level understanding in industrial environments.
- *Robotic Foundation Models for Science and Generalization.*Inspired by foundation models in natural language processing, the future of robot learning will be shaped by Robotic Foundation Models (RFMs)—large, pre-trained multimodal models that can generalize across tasks, environments, and embodiments. These models will integrate vision, language, force, and state data into a unified representation, enabling zero-shot or few-shot adaptation to novel contact-rich tasks. RFMs will also accelerate scientific discovery, allowing robots to autonomously conduct experiments, simulate hypotheses, and collaborate with human researchers in domains such as material science, chemistry, and advanced manufacturing.
- *Multimodal and Cross-Modal Reasoning.*Next-generation collaborative robots will increasingly rely on joint reasoning over multiple modalities (e.g., speech, vision, force feedback) to robustly interpret ambiguous situations or resolve conflicts. Techniques such as multimodal transformers and cross-modal attention mechanisms will be critical for fusing diverse signals.
- *Explainable and Trustworthy AI.*Building trust between human workers and robotic systems requires transparency and interpretability. Future HRC

systems must provide explainable actions and justifications, particularly in high-risk or safety-critical environments. Human-understandable feedback and behaviour prediction are key components of trustworthy collaboration.

- *Edge AI and Real-Time Decision Making.*To meet the latency and reliability requirements of industrial HRC, future solutions will increasingly leverage Edge AI architectures that enable on-device processing of multimodal inputs and learned policies. This supports scalable, real-time decision making in decentralized and bandwidth-limited factory floors.
- *Open Benchmarks and Standardized Evaluation.*There is a growing need for shared datasets, simulation platforms, and evaluation protocols that reflect realistic HRC scenarios. Open-source frameworks and benchmarking environments (e.g., Isaac Sim, ROS2) will support reproducibility, comparison, and rapid innovation.

By aligning technological development with human needs and values, future HRC systems will not only improve productivity, but also foster a new generation of collaborative, inclusive, and ethically grounded industrial workspaces. The integration of multimodal intelligence and adaptive learning stands as a cornerstone of this human-cantered industrial revolution.

7. Conclusion

Human-centered Industry 5.0 systems demand seamless collaboration between humans and robots, where flexibility, safety, and adaptability are key priorities. By leveraging multimodal data - including vision, haptics, audio, and language - robots can develop a richer contextual understanding of their environment, enabling more intuitive and responsive interactions with human operators. Robot learning approaches, such as deep reinforcement learning, imitation learning, and foundation models, play a pivotal role in equipping robots with the ability to generalize across complex, unstructured tasks and adapt to individual user needs. This study emphasizes that the fusion of multimodal sensing and advanced learning methods not only improves task efficiency and precision but also contributes to building trustworthy, explainable, and ergonomic human-robot collaboration frameworks. The integration of these technologies is central to achieving the vision of Industry 5.0, where robots act as collaborative partners rather than passive tools, supporting human creativity and decision-making.

Future research will focus on lifelong learning, multimodal reasoning, and large-scale robotic foundation models to further enhance adaptability and cross-domain generalization. The synergy between multimodal AI, edge computing, and human-centered design will be the cornerstone for developing intelligent robotic systems capable of safe and meaningful collaboration in next-generation industrial environments.

The future of Industry 5.0 will be defined by collaborative intelligence, where humans and robots form symbiotic ecosystems that push beyond traditional automation toward a more resilient, innovative, and human-centered industrial landscape.

8. References

[1] Wang, T., Zheng, P., Li, S., Wang, L. (2024).*Multimodal Human–Robot Interaction for Human-Centric Smart Manufacturing: A Survey*. Adv. Intell. Syst., 6, 2300359.

[2] Matheson, E., Minto, R. Zampieri, E.G.G., Faccio, M. and Rosati, G. (2019).*Human-Robot Collaboration in Manufacturing Applications: A Review*. Robotics 2019, 8(4), 100.

[3] https://www.scapetechnologies.com/blog/robots-vs-cobots-what-are-differences(Accessed July 2025)

[4] Zamboni, M. Valente, A. (2020). *Collaborative Robots: Overview and Future Trends*In Book: Industrial Robots: Design, Applications and Technology (Eds: Karabegović, I., Banjanović-Mehmedović, L.), Nova Science Publisher, USA.

[5] Burden, A.G., Caldwell, G.A., Guertler, M.R. (2022). *Towards Human–Robot Collaboration in Construction: Current Cobot Trends and Forecasts*. Constr. Robot. 6, 209–220.

[6] Liu, Y., Caldwell, G., Rittenbruch, M., Belek Fialho Teixeira, M., Burden, A., Guertler, M. (2024). *What Affects Human Decision Making in Human–Robot Collaboration?: A Scoping Review*. Robotics 2024, 13, 30. https://doi.org/10.3390/ robotics13020030

[7] Wang, L., Gao, R. X., Váncza, J., Krüger, J., Wang, X. V., Makris, S., and Chryssolouris, G. (2019). *Symbiotic human–robot collaborative assembly*. CIRP Annals, 68(2), 701–726, https://doi.org/10.1016/j.cirp.2019.05.002

[8] Puttero, S., Verna, E., Genta, G. et al. (2025). *Collaborative robots for quality control: an overview of recent studies and emerging trends*. J Intell Manuf https://doi.org/10.1007/s10845-025-02600-w

[9] Pietrantoni, L., Favilla, M., Fraboni, F., Mazzoni, E., Morandini, S., Benvenuti, M., De Angelis, M. (2024). *Integrating collaborative robots in manufacturing, logistics, and agriculture: Expert perspectives on technical, safety, and human factors*. Front Robot AI.;11:1342130. doi: 10.3389/frobt.2024.1342130.

[10] Casalino, A., Cividini, F. Zanchettin, A.M., Piroddi, L., Rocco, P., (2018). *Human-robot collaborative assembly: a use-case application*, IFAC-PapersOnLine, Volume 51, Issue 11, Pages 194-199, ISSN 2405-8963, https://doi.org/10.1016/j.ifacol.2018.08.257.

[11] https://www.kuka.com/en-us/future-production/human-robot-collaboration (AccessedJuly 2025)
[12] Dhanda, M., Rogers, B.A., Hall, S., Dekoninck, E., Dhokia, V. (2025). *Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of industry 5.0*, Elsevier Robotics and Computer–Integrated Manufacturing 93, 102937
[13] Banjanović-Mehmedović, L., Gurdić, A. (2021). *Collaborative Service Robots: Challenges, Paradigms and Applications*, in Book: Service Robots: Advances in Research and Application (Eds. Karabegović, I., Banjanović-Mehmedović, L.), Nova Science Publisher, USA.
[14] Liu, S.(2025). *Multimodal human-robot collaboration: Advancements and future directions*. Int. J. Manufacturing Research, Vol. 20, No. 1.
[15] Subašić, S., Banjanović-Mehmedović, L., Subašić, H., Karabegović, I., Husak, E.*Vision Transformer-Based Data Fusion for Gesture and Object Classification in Human-Robot collaboration*, RAAD2025 Conference, Serbia, 2025.
[16] Liu, H., Fang, T., Zhou, T., and Wang, L. (2018). *Towards Robust Human-Robot Collaborative Manufacturing : Multimodal Fusion*, IEEE Access 6,74762-74771,2018.
[17] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019*). Multimodal Machine Learning: A Survey and Taxonomy*. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
[18] Lemaignan, S., et al. (2017). *Artificial cognition for social human–robot interaction: An implementation*. Artificial Intelligence, 247, 2017.
[19] Ajoudani, A., Zanchettin, A.M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., Khatib, O. (2018). *Progress and prospects of human–robot collaboration*. Autonomous Robots, 42(5).
[20] Lee, J. (2017). *A survey of robot learning from demonstrations for human–robot collaboration*. arXiv preprint arXiv:1710.08789. https://arxiv.org/abs/1710.08789
[21] Sutton, R.S., Barto, A.G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.
[22] Husaković, A., Banjanović-Mehmedović, L., Gurdić-Ribić, A., Prljača, N. and Karabegović, I. (2025) *Reinforcement learning for robot manipulation tasks in human-robot collaboration using the CQL/SAC algorithms*, Advances in Production Engineering & Management (APEM), Volume 20, 2025, Issue 1.
[23] Kartoun, U., Stern, H. and Edan, Y. (2010). *A Human-Robot Collaborative Reinforcement Learning Algorithm*. Journal of Intelligent Robot System, 2010.

[24] Castro, A., Silva, F., Santos, V. (2021). *Trends of Human-Robot Collaboration in Industry Contexts: Handover, Learning, and Metrics*. Sensors 21, 4113. https://doi.org/10.3390/s21124113
[25] Semeraro, F., Griffiths, A., & Cangelosi, A. (2021). *Human–robot collaboration and machine learning: A systematic review of recent research*. arXiv preprint arXiv:2110.07448. https://arxiv.org/abs/2110.07448
[26] Mukherjee, D., Gupta, K., Chang, L.-H., & Najjaran, H. (2022). *A survey of robot learning strategies for human–robot collaboration in industrial settings*. Robotics and Computer-Integrated Manufacturing, 73, 102231. https://doi.org/10.1016/j.rcim.2021.102231
[27] Vecerik, M., Hester, T., Scholz, J., Wang, F., Pietquin, O., Piot, B., Heess, N., Rothörl, T., Lampe, T., Shah, R., Martín, R.M., Zhu, Z. *MUTEX: Learning Unified Policies from Multimodal Task Specifications*,Conference on Robot Learning (CoRL), November 2023. https://rpl.cs.utexas.edu/publications/2023/11/06/shah-corl23-mutex/
[28] Shah R., Martín-Martín, R. and Zhu, Y. (2023). *MUTEX: Learning Unified Policies from Multimodal Task Specifications*, arXiv, https://arxiv.org/abs/2309.14320
[29] Balakuntala, M.V., Kaur, U., Ma, X., Wachs, J. and Voyles, R.M. *Learning Multimodal Contact-Rich Skills from Demonstrations Without Reward Engineering*, 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021, pp. 4679-4685, doi: 10.1109/ICRA48506.2021.9561734.
[30] https://upinderkaur22.github.io/projects/3_project/(Accessed July 2025)