Human Factors in Robots, Drones and Unmanned Systems, Vol. 179, 2025, 43–52
https://doi.org/10.54941/ahfe1006373

Integrating Robotics, AI, and Immersive Technologies: A Modular Framework for Human-Metahuman-Robot Collaboration

Ramisha Fariha Baki[1], Apostolos Kalatzis[2], and Laura Stanley[1]

[1]Gianforte School of Computing, Montana State University, MT 59717, USA
[2]Computer Science Department, Cleveland State University, OH 44115, USA

ABSTRACT

Collaborative robots have been increasing rapidly across industries, particularly in manufacturing settings. This advancement allows humans and robots to work side by side to complete tasks more efficiently. Moreover, with the development of synthetic actors like metahumans, humans can now enter immersive environments where these metahumans act as guides, facilitating task execution. However, there is a limited implementation of combining both technologies in the industrial field. This paper aims to demonstrate how metahumans can enhance efficiency and guidance to humans and how collaborative robots can help that person in an industrial context. In this paper, we present an illustration showcasing how a metahuman can guide a human and how a human can command a robotic arm according to the instructions of the metahuman during a simple pick-and-place task on an assembly line. The goal of this research to improve the pedagogical curve of an assembly worker by introducing metahumans and collaborative robots.

Keywords: Collaborative robot, Human computer interaction, Metahuman, Virtual reality

INTRODUCTION

Robot technology is one of the key technologies in the 4th industrial revolution along with AI (Artificial intelligence) and IoT (Internet of Things) technology. Especially robots that can collaborate with humans are drawing attention at industrial sites (Lee et al., 2019). Following this trend, the Fifth Industrial Revolution (IR5.0) will continue concentrating more on an innovative human-machine interface and not replace human workers. It brings together the best aspects of both worlds, humans and machines, working together for increased productivity (George and George, 2020). Hence, there is a strong trend in the research community and the industry toward developing collaborative robots, the so-called cobots (Anandan, 2013). A collaborative robot is an umbrella term that conveys the general idea of proximity between machines and humans for some useful tasks in a shared space, with a range of options for timing (continuously, synchronously, alternately, etc.) (Galin and Meshcheryakov, 2019). Instead of being caged,

collaborative robots work together with people in a cooperative environment to assist with complex tasks that cannot be fully automated and to fulfil tasks that could be risky for people, which results in fewer accidents on the work floor (Brussel, n.d.). Humans will no longer need to perform tedious and dangerous jobs, leading to a reduction in workplace injuries, such as musculoskeletal disorders, which affect millions of workers globally each year and cost businesses billions in revenue (Owen-Hill, 2016).

Previous research has shown that researchers have worked toward finding a way to allow people to communicate with machines in the same manner as people communicate with other people (Guzman, 2016, 2018; Oberquelle et al., 1983; Wilpon and Roe, 1994). Speech recognition, evolving since the 1950s, integrates various disciplines to process human speech effectively (Mohamad et al., 2016; Rabiner and Juang, 1999). One study demonstrated robots in the medical field as receptionist robots, nurse assistants, and a server with communication being established through speech recognition (Ahn et al., 2015). Similarly, object detection is an essential task and has been widely studied in computer vision (Sahin and Ozer, 2021). However, the integration of these components into a modular framework remains a challenge, requiring further exploration. Moreover, ensuring seamless communication and coordination between human workers and robotic systems in dynamic industrial environments is a significant challenge in Human-Robot Collaboration (HRC) (Karuppiah et al., 2023). Research has also highlighted the importance of human-centered design in HRC with studies demonstrating that well-designed interfaces, such as augmented reality-based systems, can provide insights on task performance, cognitive load, and situational awareness during collaborative tasks (Kalatzis et al., 2023a, 2023b).

Emerging technologies such as Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality (XR) are regarded as good solution candidates for increasing the communication between humans and machines during the design, commission, and operation phases (Kyrlitsias and Michael-Grigoriou, 2022). Tools like Unreal Engine, combined with AI-driven avatars such as Metahumans, have been employed to provide real-time guidance, creating a more engaging and accessible experience for users (Alcántara et al., 2024). The creation of avatars with very high levels of human resemblance has recently become more accessible, such as the UNREAL metahuman (Higgins et al., 2021). Currently, UNREAL metahumans represent one of the most advanced commercially available avatars to date, offering high human resemblance and graphical realism such that the features appear photorealistic (Fraser et al., 2024).

However, to our knowledge, there is no existing research where Metahumans have been utilized in a manufacturing environment. This paper seeks to fill this gap by integrating Metahuman with a collaborative robot to guide a human in performing a simple pick-and-place task. Thus, this paper builds on existing literature by integrating these technologies into a single modular system. By combining speech recognition, object detection, motion planning, and AI-enabled Metahuman guidance within an immersive

environment, the proposed system addresses gaps in previous studies and introduces a novel framework for intuitive human-robot interaction.

SYSTEM ARCHITECTURE

This section explores integrating Unreal Engine's Metahuman Creator with a collaborative robotic arm in creating an immersive assembly training environment. Here, the Metahuman guides human commands to fetch and position components via the robotic arm.

Virtual Environment

The virtual environment serves as the platform for the Metahuman to interact with users through the Meta Quest Headset 2. The Unreal Engine can model a fully immersive virtual workspace to simulate scenarios such as a factory assembly line. There can be realistic elements like tools, workstations, and components required for training, ensuring a comprehensive and interactive experience for users. An example of such a workplace is shown in Figure 1. This scene was created by the Convai website (a company specializing in tools and APIs for creating conversational AI systems, particularly for virtual characters and game development) that can be used in the Unity Platform (Convai, n.d.).

Figure 1: Virtual environment created by Convai website (Convai, n.d.).

Metahuman Guide

The Metahuman is developed in Unreal Engine 5.4.4. version Platform. The Metahuman is AI enabled with ChatGPT integrated using Convai Plugin. The Metahuman developed in the Unreal Engine is shown in Figure 2.

Collaborative Robot Arm

For easier demonstration purposes, the collaborative robot arm will perform a pick-and-place action. The collaborative robot arm utilized in this system is a UR3e robotic arm (Universal Robots, n.d.) equipped with a 2F Adaptive Gripper (Robotiq, n.d.) for handling tasks (Figure 3). The arm is controlled by the human through voice recognition, which enables the user to command the arm to detect an object and fetch it to a desired position. An Intel RealSense Depth Camera (Intel, n.d.) is employed for object detection and localization.

Figure 2: Metahuman with ChatGPT integration, created in Unreal Engine 5.4.4 for interactive dialogue with users.

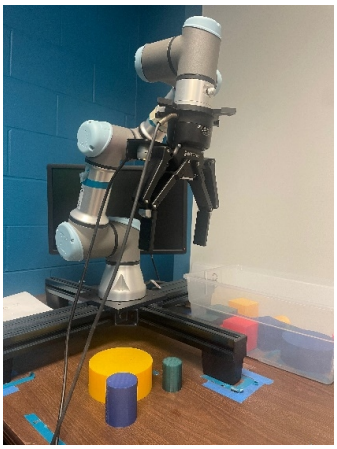

Figure 3: Robotic arm with a 2F adaptive gripper.

The modular architecture is designed to facilitate communication between the human, robotic arm, and the virtual environment. The human serves as the intermediary, using voice recognition to command the robotic arm to perform pick-and-place tasks. The Meta Quest 2 headset enhances interaction by bridging the human experience with the virtual environment through immersive visuals.

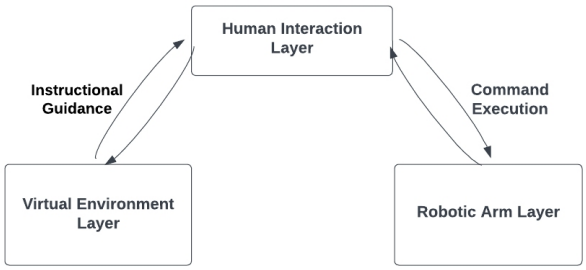

Figure 4: System architecture.

USE CASE

This section demonstrates how the system/tool is used in practical training scenarios, where a human trainee wearing the Meta Quest Headset 2 is guided by the Metahuman to command the robotic arm to pick and drop some simple objects. The process unfolds as follows:

Initialization: The human trainee enters the immersive environment and is greeted by the metahuman, who provides an overview of the task.

Step-by-Step Instruction: The Metahuman gives verbal instructions like "Ask the Robotic Arm to fetch the Yellow Circle Block". The Metahuman also confirms with the user whether the correct block was fetched by the robotic arm. If not, it repeats the instruction until the correct block is fetched.

Robot Interaction: The robot waits for the human command and fetches the object specified by the human. For example, the human can say, "Fetch the Yellow Circle Block". After hearing the command, the robotic arm moves to a designated place and, using object detection, finds the Yellow Circle Block, picks it up, and places it in front of the user.

The use case diagram of this immersive task system is shown in Figure 5.

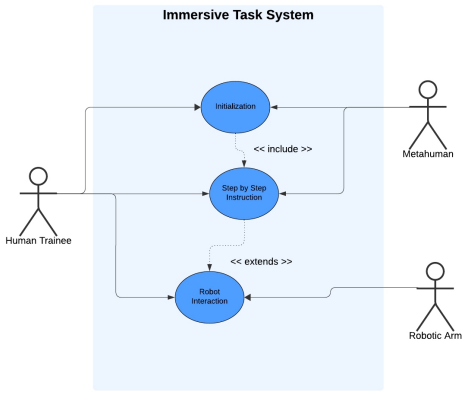

Figure 5: Use case diagram of the immersive task system.

TECHNICAL DETAILS

This section provides insights into the underlying technologies used in the entire system.

Metahuman Integration

Metahuman technology plays a critical role in this project as a platform for developing synthetic actors. The purpose of creating Mr. Hamilton (the metahuman) is to assist assembly workers by demonstrating tasks for them. Assembly workers can wear a VR headset (Meta Quest Headset) and learn to perform tasks by following his step-by-step guidance, adapting quickly to cobot-integrated processes and technologies.

Figure 6: Metahuman interacting within a designed environment, engaging in real-time conversations with users.

The narrative design is provided (Figure 7) to the Metahuman to function as a trigger and a conditional (if-else statement), enabling it to operate in all types of scenarios presented by the user.

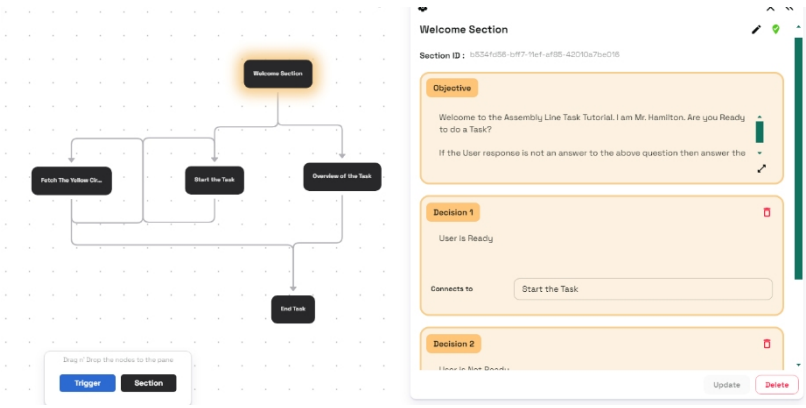

Figure 7: Flow diagram of the narrative design for the assembly line task tutorial in a Metahuman interface.

Immersive Environment

An immersive environment was created using the Unreal Engine to place the Metahuman inside the scene. Afterward the MetaQuest Headset was connected with the Unreal Engine using OpenXR API and Steam VR to enable real-time rendering and hand-tracking interactions. In the immersive environment, the user can talk to the Metahuman by pressing the A button on the right hand controller. The immersive environment is shown in Figure 8.

Figure 8: The interface displayed on the Meta Quest headset in an immersive environment.

Robot Interaction

Voice Recognition Module: The Voice Recognition Module uses ROS Noetic to enable a robot to fetch a "yellow circle block" based on human voice commands. Google Speech-to-Text captures and processes the command, which is then parsed to identify the task and object, with the information published to the robot via a topic.

Object Detection Module: The Object Detection Module uses OpenCV and an Intel RealSense Depth Camera to locate predefined objects, such as a red rectangle block or yellow circle block, and determine their coordinates. Without training, the model relies on hard-coded recognition. These coordinates are shared with the motion planner node, which utilizes ROS MoveIt! to calculate and execute the UR3e robotic arm's trajectory for pick-and-place tasks, as shown in Figure 9.

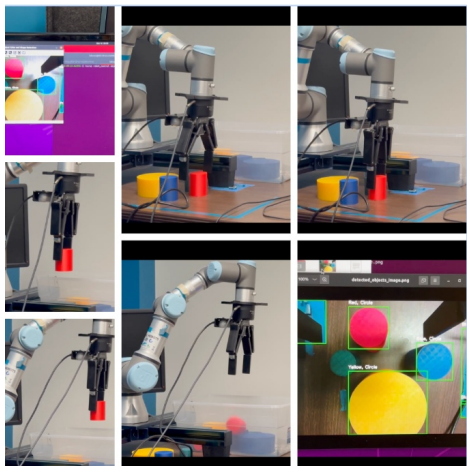

Figure 9: UR3e robot performing object detection and pick-and-place tasks.

After completing the task, the robotic arm returns to a listening state, waiting for the user's next command. This architecture effectively integrates human voice commands with robotic motion for intuitive and efficient task execution.

CONCLUSION

This paper presents a modular and interactive human-robot interaction integrating human voice commands and object detection. Moreover, there is also a human-computer interaction between the human and the AI-enabled Metahuman in an immersive environment. The system demonstrates a practical solution for intuitive human-robot collaboration by utilizing ROS (Noetic version) along with advanced tools for speech recognition, object detection, and motion planning for seamless communication between nodes. The architecture of the system allows each component—voice recognition, object detection, and robotic motion—to function independently while collaborating through ROS communication protocols. Such architecture ensures flexibility, scalability, and ease of maintenance, making the system adaptable to a variety of environments and use cases.

However, the system has some limitations that warrant future exploration. At the current stage of the system, the bridge between the Metahuman and the robot is the human. As a result, the user needs to verify with the Metahuman after every step whether the robotic arm has completed its task successfully or not. If direct communication can be established between the Metahuman and the robotic arm, the Metahuman could modify its commands in real time based on the robot's performance. Another limitation of the system is the dependency on pre-defined object detection models, which may struggle in cluttered or dynamic environments. Similarly, the speech recognition module could benefit from enhanced capabilities to understand complex or domain-specific commands.

In conclusion, this work demonstrates a practical approach to combining robotics, artificial intelligence, and immersive technologies to create an intuitive and efficient human-robot collaboration system. By bridging the gap between humans and robots, this system paves the way for innovative applications in industrial automation, education, and beyond. Moreover, the integration of the Metahuman enables non-technical users to effectively interact with advanced robotics, making the system accessible and user-friendly.

FUTURE WORK

Future work can focus on addressing the current limitations and expanding the system's capabilities for more complex and autonomous operations. A key area for exploration is establishing direct communication between the Metahuman and the robotic arm to eliminate the need for the human as an intermediary. By enabling the Metahuman to receive real-time feedback from the robotic arm, it can dynamically adjust commands and guide the system autonomously. This will improve efficiency and reduce user dependency. Enhancements in speech recognition could also be a priority,

incorporating advanced Natural Language Processing (NLP) models to handle more complex, multi-step, or domain-specific commands. These improvements would make the system more robust and capable of handling a wider range of tasks. Another potential direction involves expanding the immersive environment to include multimodal feedback (e.g., haptic feedback) could further improve the user experience by making human-metahuman interaction more natural and intuitive. These advancements would not only improve the current system but also open doors for broader applications in industrial automation, logistics, education, and healthcare, showcasing the transformative potential of human, Metahuman, and robot collaboration.

ACKNOWLEDGMENT

This research was supported by the National Science Foundation (NSF Award #2013651) and the National Institutes of Health. We extend our gratitude to the Gallatin College manufacturing instruction faculty at Montana State University for their role in developing the holographic synthetic actor. We also acknowledge the use of Meta Quest 3, UR3e robotic arm and other tools that supported our research. Finally, we express our appreciation to the National Science Foundation for its continuous support in advancing interdisciplinary approaches to human-robot collaboration and immersive technology.

REFERENCES

Ahn, H. S., Lee, M. H., MacDonald, B. A.: Healthcare robot systems for a hospital environment: CareBot and ReceptionBot. In: 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Kobe, Japan, pp. 571–576 (2015).

Alcántara, J. C., Tasic, I. and Cano, M. D., 2024. Enhancing digital identity: Evaluating avatar creation tools and privacy challenges for the metaverse. *Information*, *15*(10), p. 624.

A. Owen-Hill, Robots can help reduce 35, 2016, [online] Available: http://robohub.org/robots-can-help-reduce-35-of-work-days-lost-to-injury/.

Convai. (n.d.). Build interactive AI-driven virtual humans. Retrieved from https://convai.com.

E. Higgins, D., Egan, D., Fribourg, R., Cowan, B., & McDonnell, R. (2021, September). Ascending from the valley: Can state-of-the-art photorealism avoid the uncanny? In ACM Symposium on Applied Perception 2021 (pp. 1–5).

Fraser, A. D., Branson, I., Hollett, R. C., Speelman, C. P., & Rogers, S. L. (2024). Do realistic avatars make virtual reality better? Examining human-like avatars for VR social interactions. Computers in Human Behavior: Artificial Humans, 2(2), 100082.

George, A. S., & George, A. H. (2020). Industrial revolution 5.0: The transformation of the modern manufacturing process to enable man and machine to work hand in hand. Journal of Seybold Report ISSN NO, 1533, 9211.

Guzman, A. L. (2018). What is human-machine communication, anyway. Human-machine communication: Rethinking communication, technology, and ourselves, 1–28.

Guzman, A. L. (2016). The messages of mute machines: Human-machine communication with industrial technologies. communication+ 1, 5(1).
H. Messe, Robot or cobot: The five key differences, 2016, [online] Available: http://www.hannovermesse.de/en/news/robot-or-cobot-the-five-key-differences.xhtml.
Kalatzis, A., Prabhu, V. G., Stanley, L., & Wittie, M. P. (2023a). A multimodal approach to investigate the role of cognitive workload and user interfaces in human-robot collaboration. Proceedings of the International Conference on Multimodal Interaction (ICMI '23).
Kalatzis, A., Prabhu, V. G., Stanley, L., & Wittie, M. P. (2023b). Effect of augmented reality user interface on task performance, cognitive load, and situational awareness in human-robot collaboration. 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).
Karuppiah, K., Sankaranarayanan, B., Ali, S. M., & Bhalaji, R. K. A. (2023). Decision modeling of the challenges to human–robot collaboration in industrial environment: A real world example of an emerging economy. Flexible Services and Manufacturing Journal, 35(4), 1007–1037.
Kyrlitsias, C., & Michael-Grigoriou, D. (2022). Social interaction with agents and avatars in immersive virtual environments: A survey. Frontiers in Virtual Reality, 2, 786665.
Lee, E., Barthelmey, A., Reckelkamm, T., Kang, H., & Son, J. (2019, October). A study on human-robot collaboration based hybrid assembly system for flexible manufacturing. In IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society (Vol. 1, pp. 4197–4202). IEEE.
L. R. Rabiner and B. -H. Juang, Fundamentals of Speech Recognition, 1993.
Mohamad, S. N. A., Jamaludin, A. A., & Isa, K. (2016, October). Speech semantic recognition system for an assistive robotic application. In 2016 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS) (pp. 90–95). IEEE.
Oberquelle, H., Kupka, I., & Maass, S. (1983). A view of human—machine communication and co-operation. International journal of man-machine studies, 19(4), 309–333.
R. I. Association, The end of separation: Man and robot as collaborative coworkers on the factory floor, 2013, [online] Available: http://www.robotics.org/content-detail.cfm/Industrial-Robotics-Industry-Insights/The-End-of-Separation-Man-and-Robot-as-Collaborative-Coworkers-on-the-Factory-Floor/content_id/4140.
Sahin, O., & Ozer, S. (2021, July). Yolodrone: Improved yolo architecture for object detection in drone images. In 2021 44th International Conference on Telecommunications and Signal Processing (TSP) (pp. 361–365). IEEE.
Vrije Universiteit Brussel. (n.d.). Collaborative robots (cobots). Retrieved January 19, 2025, from http://mech.vub.ac.be/multibody/topics/cobots.htm.
Wilpon, J. G., & Roe, D. B. (Eds.). (1994). Voice communication between humans and machines. National Academies Press.