International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Revolutionizing Human-Robot Interaction (HRI): Multimodal Intelligent Robotic System for Responsive Collaboration

Babu. G[1], E. Sathiyanarayanan[2], S. Parasuraman[3], Balamurugan. D[4], Anitha Jaganathan[5]

Submitted: 12/12/2023 **Revised:** 24/01/2024 **Accepted:** 03/02/2024

Abstract: With the advancement of robotics throughout time, human-robot interaction (HRI) is now crucial for providing optimal user experience, reducing tedious activities, and increasing public acceptance of robots. A central aspect of the investigation involves developing context-aware robotic systems that can dynamically adapt to varying environmental conditions and user contexts. By incorporating real-time adaptability into the robotic framework, the research aims to create a more responsive and intuitive human-robot collaboration experience. In order to facilitate the advancement of robots, it is imperative to adopt innovative Human-Robot Interaction (HRI) strategies, with a particular emphasis on fostering a more natural and adaptable mode of interaction. Multimodal HRI, as a recently emerging methodology, provides a means for individuals to engage with robots through diverse modalities, encompassing voice, images, text, eye movement, touch, and even bio-signals such as EEG and ECG. This approach marks a significant shift in HRI paradigms, offering a versatile framework for enhanced communication between humans and robots. In this paper, a Multi-Modal Intelligent Robotic System (MIRS) is proposed, comprising several distinct modules. Leveraging various sensors such as image, sound, and depth, these modules can operate independently or collaboratively to facilitate efficient interaction between humans and robots. Three key components are identified and implemented in this research, which includes the location and posture of the object, information extraction, gesture analysis and eye tracking. Experimental evaluations were conducted to gauge the performance of these interaction interfaces, and the findings underscored the effectiveness of the proposed approach.

Keywords: *multimodal, human-robot interaction, robotics, communication, sensors, multi-modal inputs*

1. Introduction

In the ever-evolving landscape of technology, the strides made in robotics have not only transformed industries but have also brought forth a fundamental shift in the way humans and machines interact [1]. In recent years, there has been a significant surge in the progress of robotics. However, the challenge persists in constructing a robot capable of engaging in natural communication with individuals and seamlessly generating comprehensible multimodal motions across various interaction scenarios. Achieving this requires the robot to possess a high level of multimodal recognition [2], enabling it to grasp the inner moods, goals, and character of the person it interacts with, thereby facilitating appropriate feedback. The ubiquity of devices for Human-Robot Interaction (HRI) in everyday life owes itself to the expansion of the Internet of Things. The conventional reliance on a single sense modality, such as sight, touch, sound, scent, or flavor, for HRI input and output is no longer the sole option [3].

The realm of Human-Robot Interaction (HRI) has emerged as a linchpin for delivering optimal user experiences, streamlining arduous tasks, and cultivating widespread acceptance of robotic entities [4]. As the integration of robots into our daily lives becomes increasingly prevalent, the need for a nuanced and adaptive collaboration between humans and robots has become more crucial than ever.

Multimodal Human-Robot Interaction (HRI) aims to establish communication with robots through diverse signals, encompassing voice, images, text, eye movement, and touch [5]. This interdisciplinary field spans cognitive science, ergonomics, communication technologies, and virtual reality. It involves the reception of multimodal input signals from humans to robots and the generation of multimodal output signals from robots to humans. This holistic approach ensures a nuanced and comprehensive interaction experience, where robots can interpret and respond to various cues, fostering a more natural and effective exchange between humans and machines.

At the heart of this paradigm shift lies the exploration of context-aware robotic systems [6], a focal point of investigation in this research. The endeavor is to craft systems that transcend the traditional boundaries of rigid programming, enabling robots to dynamically adapt to the multifaceted intricacies of

[1] *Professor, Department of Biomedical Engineering,*
SRM Easwari Engineering College, Chennai
Email: babu.g@eec.srmrmp.edu.in, babutry@gmail.com
ORCID: 0000-0002-7363-0607
[2] *Assistant Professor, ECE,*
Madanapalle Institute of Technology & Science,
Madanapalle, Andhra Pradesh
Email: sathiyaame@gmail.com
ORCID: 0000-0001-7182-3141
[3] *Professor, Electronics and Communication Engineering,*
Karpaga Vinayaga College of Engineering and Technology
Email: parasuengineers@gmail.com
ORCID: 0000-0002-5982-8983
[4] *Associate Professor,*
Department of Computer Science and Engineering,
Sona College of Technology, Salem, Tamilnadu, India.
Email: balamurugand_81@yahoo.com
ORCID: 0000-0001-5248-9651
[5] *Assistant Professor,*
Department of Artificial Intelligence and Data Science,
Panimalar Engineering College, Chennai
Email: anitha@panimalar.ac.in
ORCID: 0009-0002-7773-0469

environmental conditions and user contexts. The linchpin of this adaptation is real-time adjustability, woven into the very fabric of the robotic framework. The objective is clear: to foster a collaborative experience that is not just efficient but also responsive and intuitively attuned to the diverse needs and nuances of human interaction.

In recent years, significant progress in robotics has hinged on the pivotal role of human-computer interaction (HCI) technology. HCI has proven instrumental in enhancing user experiences, streamlining tasks, and fostering widespread acceptance of robots. To propel the trajectory of robotics further, innovative HCI strategies are imperative, with an emphasis on cultivating a more natural and adaptable interaction style, as underscored by [7]**.** The evolution of HCI remains pivotal in shaping a future where human-robot collaboration is characterized by heightened intuitiveness and seamless adaptability.

To propel the advancement of robotics into this new era of adaptability, the adoption of innovative Human-Robot Interaction (HRI) strategies is imperative. Central to this shift is the emphasis on cultivating a more natural and adaptable mode of interaction between humans and robots. This paper addresses this imperative by introducing a Multimodal HRI methodology, a cutting-edge approach that serves as a conduit for individuals to engage with robots through a spectrum of communication modalities. This inclusive framework spans voice, images, text, eye movement, touch, and even extends to the integration of bio-signals such as EEG and ECG [8]. The adoption of Multimodal HRI represents a transformative leap, promising a versatile platform that transcends conventional communication barriers, elevating the quality of interaction between humans and robots to unprecedented heights.

In the pages that follow, propose a revolutionary multi-modal framework designed to reshape the landscape of human-robot collaboration. This framework is not a mere conglomeration of sensors and modules but a strategic orchestration that leverages various sensory inputs such as image, sound, and depth. These modules are designed to function independently or collaboratively, offering a dynamic and adaptive interface that facilitates seamless interaction between humans and robots.

At the core of this proposed framework are three key components, meticulously identified and implemented to fortify the pillars of human-robot collaboration. The first component addresses the fundamental need for the precise determination of the location and posture of objects, empowering robots to navigate and interact effectively within their environment. The second component revolves around information extraction, a critical capability that enables the system to decipher intricate details from verbal instructions, leading to more nuanced and context-aware responses. Finally, the third component delves into the domain of gesture analysis and eye tracking, recognizing the significance of non-verbal cues in human communication and aiming to decode these cues for a more intuitive interaction experience. Multimodal Human-Robot Interaction for Various Signals is depicted in Fig.1.

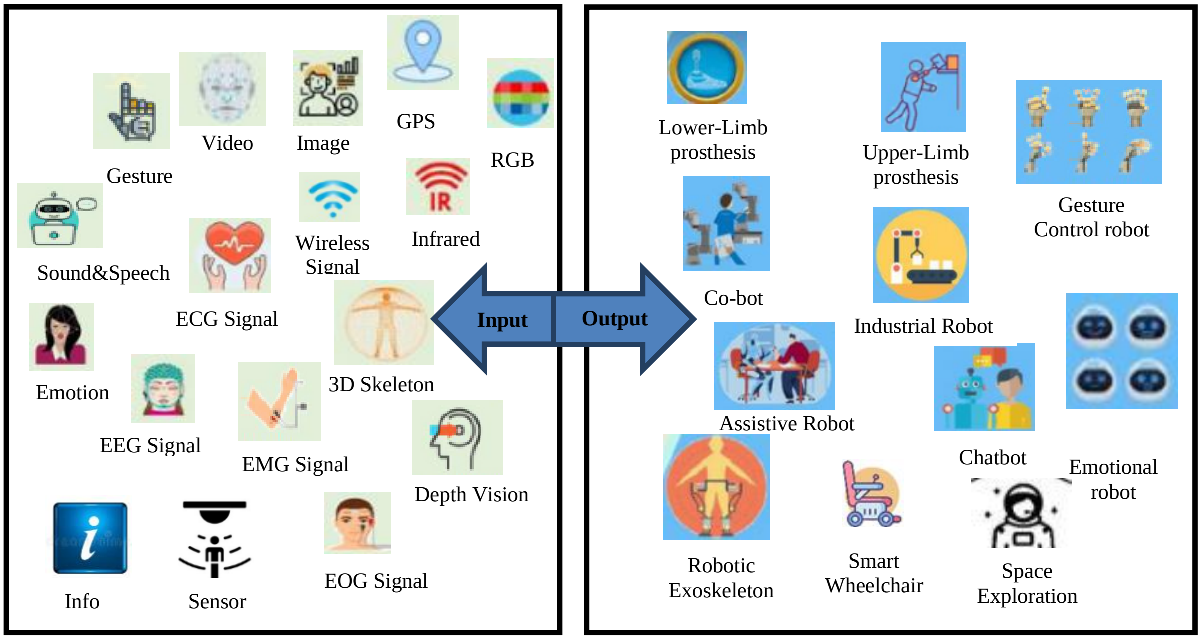

Fig. 1. Multimodal Human-Robot Interaction for Various Signals

To validate the efficacy of the proposed framework and its key components, a series of comprehensive experimental evaluations were conducted. These evaluations aimed to scrutinize the performance of the interaction interfaces in diverse scenarios and usage contexts. The findings derived from these experiments serve as a testament to the effectiveness of the research approach, affirming its potential to elevate the standard of human-robot collaboration.

In the subsequent sections, a detailed exploration of the proposed multi-modal framework, its architecture, and the integration of key components will be presented. Additionally, the experimental findings and their implications for the future of HRI will be discussed in depth. This paper will unfold the layers of innovation, unveiling a vision for a future where the synergy between humans and robots transcends conventional boundaries, leading to a more responsive, adaptive, and harmonious coexistence.

2. Literature Review

[9] Proposed the pivotal role of human-robot collaboration (HRC) in Industry 4.0, highlighting the need for intuitive communication modalities. It introduces Multi-Modal Offline and Online Programming (M2O2P), a software component facilitating communication via configurable hand gestures. Evaluated within a smart factory context (SHOP4CF EU project), personalized gestures were found to reduce user-perceived physical and mental workload, with high system usability (SUS) scores (79.25) affirming overall effectiveness. Notably, M2O2P exhibited a gesture recognition accuracy of 99.05%, aligning with state-of-the-art applications, emphasizing its reliability in advancing seamless HRC in Industry 4.0 environments.

[10] Addressed the current challenges in the interaction between robots and humans during collaborative activities of daily living, despite significant progress in social robotics and autonomy. Recognizing the frequent use of multiple communication modalities in such engagements, the paper introduces a Multimodal Interaction Manager framework. At its core lies a Hierarchical Bipartite Action-Transition Network (HBATN), enabling the robot to deduce task and dialogue states from spoken utterances and pointing gestures. The proposed framework, implemented on a robot, demonstrates its potential by successfully engaging in task-oriented multimodal interactions. This research contributes to the advancement of assistive robots, bridging the gap in effective human-robot collaboration in real-world scenarios.

[11] Depicted human emotion recognition through facial expressions, vital for medical diagnostics and human-robot interaction. Introducing "ConvNet-3," a novel Convolution Neural Network (CNN) model, the focus is on optimizing training accuracy in a limited number of epochs. Trained on the FER2013 dataset, ConvNet-3 achieves 88% training and 61% validation accuracy, surpassing existing models. However, observed overfitting on the CK+48 dataset suggests potential limitations. This research contributes to emotion recognition technology, showcasing ConvNet-3's effectiveness and signaling avenues for further refinement in real-world applications.

[12] Introduced the concept of Proactive HRC, advocating for a shift from reactive operations to a symbiotic relation with five stages of intelligence: Connection, Coordination, Cyber, Cognition, and Coevolution. Proactive HRC, characterized by mutual-cognitive, predictable, and self-organizing capabilities, envisions collaborative manufacturing tasks where human and robotic agents consider each other's needs and capabilities. The paper addresses current challenges and outlines future research directions, offering valuable insights for academic and industrial practitioners navigating human-robot flexible production.

[13] Presented a neural network-based user simulator for training Reinforcement Learning agents in collaborative tasks with diverse communication modes. Trained on the ELDERLY-AT-HOME corpus, the simulator creates a multimodal interactive environment, incorporating language, pointing gestures, and haptic-ostensive actions. To address limited datasets, a novel multimodal data augmentation approach is proposed, mitigating the challenges of resource-intensive human demonstration collection. The study underscores the potential of Reinforcement Learning and multimodal user simulators in advancing domestic assistive robot development.

3. Methodology

Creating a system architecture diagram involves visually representing the key components and their interactions within the proposed Multimodal Intelligent Robotic System for Responsive Collaboration. Fig. 2 shows the overall system architecture of the proposed system. Below is a textual description of the system architecture, outlining the major modules and their relationships:

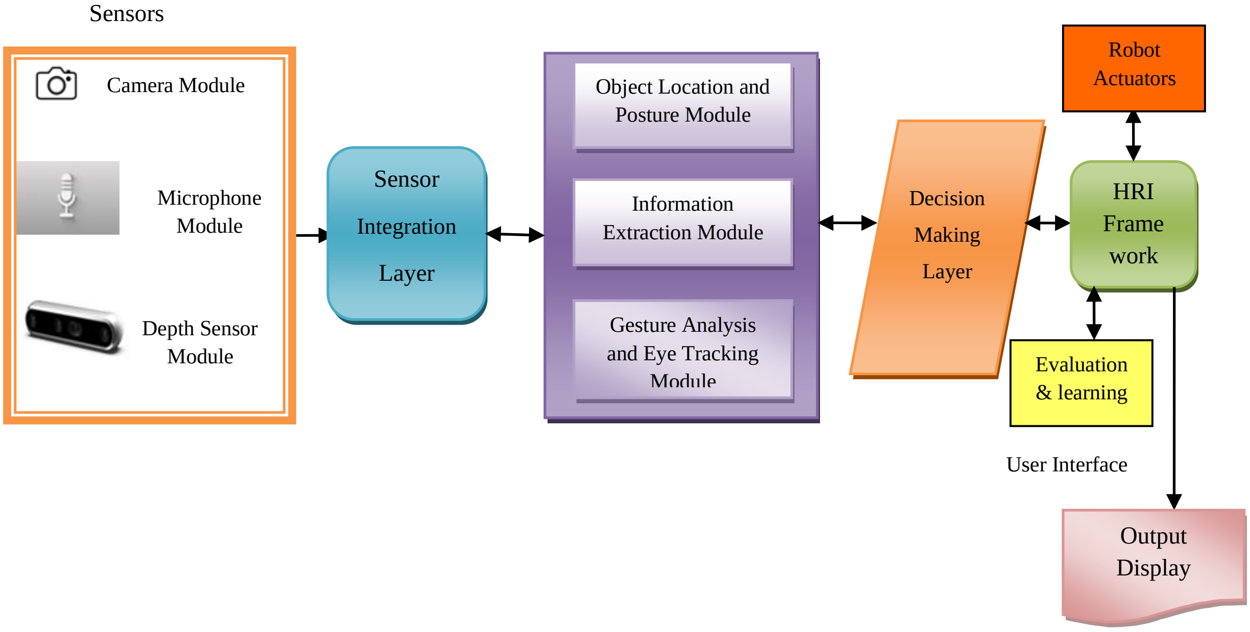

Fig. 2. Proposed System Architecture

Sensors:

The sensor array comprises vital components. The Camera Module captures visual data to facilitate object location and posture analysis. The Microphone Module records sound, enabling speech recognition and audio input. Additionally, the Depth Sensor Module contributes crucial 3D spatial data, enhancing the system's ability to perceive and understand the surrounding environment for responsive collaboration in the Multimodal Intelligent Robotic System.

Sensor Integration Layer:

The Sensor Integration Layer is pivotal in harmonizing diverse inputs. Sensor Fusion blends data from cameras, microphones, and depth sensors, forging a cohesive perception of the environment. Simultaneously, the Calibration Module guarantees precise synchronization and calibration of sensor data, ensuring the accuracy and reliability of the Multimodal Intelligent Robotic System for Responsive Collaboration.

Object Location and Posture Module:

The Object Location and Posture Module employs advanced Computer Vision Algorithms to meticulously process image data, identifying objects and gauging their spatial relationships. Additionally, Depth Estimation utilizes data from the depth sensor, precisely determining the distance between the robot and identified objects. These integrated processes form a robust foundation, enabling the Multimodal Intelligent Robotic System to dynamically adapt and collaborate responsively in various environments.

Information Extraction Module:

Within the Information Extraction Module, Natural Language Processing (NLP) scrutinizes textual input, extracting pertinent information from user interactions. Speech Recognition processes audio data captured by the microphone, comprehending spoken commands with accuracy. Bio-signal Processing delves into EEG and ECG signals, capturing nuances of emotional and physiological states. These integrated components empower the Multimodal Intelligent Robotic System to interpret and respond to users' diverse inputs with heightened contextual understanding.

Gesture Analysis and Eye Tracking Module:

The Gesture Analysis and Eye Tracking Module employs cutting-edge technologies to enhance user interaction. Leveraging Computer Vision for Gesture Analysis, the system interprets gestures through intricate image data processing. Simultaneously, Eye-tracking Technology meticulously monitors and interprets user eye movements, providing valuable insights into focus and intentions. These combined capabilities elevate the Multimodal Intelligent Robotic System, fostering a more intuitive and responsive collaboration with users.

Decision-Making Layer:

The Decision-Making Layer is pivotal in the system's responsiveness. Context Awareness integrates information from all modules, fostering a dynamic comprehension of the user's context. Decision Algorithms then leverage this contextual understanding to make informed decisions, enabling the Multimodal Intelligent Robotic System to collaboratively and adaptively interact with users in real time.

Human-Robot Interaction (HRI) Framework:

The Human-Robot Interaction (HRI) Framework orchestrates seamless collaboration. The Task Planner receives input from the Decision-Making Layer, strategizing robot tasks based on user intent. Simultaneously, the Communication Module establishes bidirectional communication, fostering a dynamic exchange of information between the Multimodal Intelligent Robotic System and the user, enhancing the overall interaction experience.

Robot Actuators:

The Robot Actuators are the execution backbone of the system. Motor Control brings planned tasks to life, orchestrating precise physical movements. Meanwhile, the Audio Output component generates spoken responses and other audio feedback, providing a dynamic and expressive channel for communication between the Multimodal Intelligent Robotic System and the user.

User Interface:

The User Interface is the bridge for effective communication. The Display offers visual feedback, conveying information about the robot's actions and responses. Simultaneously, the Audio Output channel enriches the interaction by conveying spoken responses and various audio feedback, ensuring a comprehensive and engaging experience in the Multimodal Intelligent Robotic System.

Evaluation and Learning Module:

The Evaluation and Learning Module is instrumental in system refinement. Performance Metrics collect data during user interactions, enabling quantitative evaluation. Learning Algorithms then utilize user feedback and experience to adapt and continuously enhance the Multimodal Intelligent Robotic System for optimal performance.

The Data Storage and Logging components play a crucial role in system optimization. The Database efficiently stores collected data, facilitating in-depth analysis and continuous system improvement. Concurrently, the Logging Module records system events and interactions, serving as a valuable resource for debugging and comprehensive analysis. This meticulous approach ensures that the Multimodal Intelligent Robotic System for Responsive Collaboration evolves and adapts based on insights derived from stored data.

$$\left\{\begin{array}{l} P_x = \{w/2, h/2\} \\ P_y = \{w, h/2\} \\ P_z = \{w/2, 0\} \end{array}\right. \quad (1)$$

Where w, h is the width and height of the reference plane.

$$\tilde{\imath} = \frac{\vec{y} X \vec{z}}{|\vec{y} X \vec{z}|} = [i_x i_y i_z] \quad (2)$$

$$\tilde{\jmath} = \frac{\vec{y}}{|\vec{y}|} = [j_x \ j_y \ j_z] \quad (3)$$

$$\tilde{\imath} = \frac{\vec{x} X \vec{y}}{|\vec{x} \ X \ \vec{y}|} = [k_x k_y k_z] \quad (4)$$

Based on the above equations, α, β, γ are expressed as,

$$\alpha = \tan^{-1}\left(\frac{j_z}{k_z}\right) \quad (5)$$

$$\beta = \tan^{-1}\left(\frac{-i_z}{\sqrt{1 - i_z^2}}\right) = \sin^{-1}(-i_z) \quad (6)$$

$$\gamma = \tan^{-1}\left(\frac{i_y}{i_x}\right) \quad (7)$$

Algorithm for Posture Estimation

```
Input: Images of planar objects, $I$

Detector ← describe feature detector
Descriptor ← describe feature descriptor
for i in $I$ do
    K←Detect Keypoints (i, Detector)
    D[i]← GetDescriptors (K, Descriptor)
End for
```

```
While camera is on do
    F ← RGM Image Frame
    PC← Point Cloud Data
    $K_f$← Detect Keypoints (f, Detector)
    $D_f$← GetDescriptors ($K_f$, Descriptor)
    for i in $I$ do
        matches ← FindMatches (D[i], $D_f$)
        if the total number of matches ≥ 8 then
            $K_{Pi}, K_{Pf}$← ExtractKeypoints (matches)
            H ← EstimateHomography ($K_{Pi}, K_{Pf}$)
            $P_x, P_y, P_z$ ← points on the planar object obtained using Equation (1)
            $P'_x, P'_y, P'_z$ ← corresponding projected points of $P_x, P_y, P_z$
            $\vec{x}, \vec{y}, \vec{z}$ ← corresponding 3D locations of $P'_x, P'_y, P'_z$ from PC
            $\vec{y} \leftarrow \vec{y} - \vec{x}$
            $\vec{z} \leftarrow \vec{z} - \vec{x}$
            $\tilde{\imath}, \tilde{\jmath}, \tilde{k}$ ← from equation (2,3 &4)
            $\alpha_i, \beta_i, \gamma_i$ ← from Equation (5,6 &7)
            Publish ($\vec{x}, \alpha_i, \beta_i, \gamma_i$)
    End for
End while
```

This assumption proves applicable in many practical instances within human-robot interaction, establishing a robust foundation for the technique to yield dependable and accurate results. The validity of this assumption enhances the reliability and precision of the outcomes, making the approach well-suited for a variety of real-world scenarios where human-robot interaction dynamics are at play.

4. Result and Discussion

Three key components are identified and implemented in the result section. The experimental evaluations conducted to assess the proposed multi-modal framework for Human-Robot Interaction (HRI) yielded promising results, affirming the effectiveness of the implemented components—object location and posture identification, information extraction, gesture analysis, and eye tracking.

4.1 Object Detection and Posture Identification

Object detection and posture estimation play crucial roles in Human-Robot Interaction (HRI) through a Multimodal Intelligent Robotic System.

Object Detection: Object detection involves the identification and localization of objects in the robot's environment. In HRI, this capability is essential for a robot to understand and interact with its surroundings. Advanced sensors, such as cameras and depth sensors, enable the robot to perceive and recognize various objects. The system processes the visual information, employing computer vision techniques and machine learning algorithms, to identify objects accurately. This capability is foundational for tasks like fetching or manipulating objects based on user commands, contributing to a more intuitive and responsive interaction.

Posture Estimation: Posture estimation refers to the ability to determine the spatial configuration or pose of a person or objects within the robot's field of view. In HRI, understanding the posture of humans is critical for the robot to interpret non-verbal cues and gestures accurately. Vision systems, often combined with depth sensors or even wearable devices, help in estimating the posture of users. This includes recognizing gestures, body language, and overall body posture. Accurate posture estimation enhances the robot's ability to respond appropriately to user actions, fostering more natural and effective communication.

Multimodal Intelligent Robotic System: The term “multimodal” in this context implies the integration of various sensory modalities, such as vision (object detection and posture estimation), audio, and potentially touch or other sensory inputs. A Multimodal Intelligent Robotic System leverages these modalities to create a comprehensive understanding of the environment and user interactions. For instance, while interacting with a person, the robot may combine information from visual cues (object detection and posture estimation), voice commands, and perhaps even gestures to offer a more holistic and context-aware response.

Integration for Human-Robot Interaction: In the context of HRI, the integration of object detection and posture estimation within a multimodal intelligent system enhances the overall interaction experience. For instance, if a user instructs the robot to pick up a specific object, the object detection component identifies the target, while posture estimation helps in understanding the user's gestural cues indicating the desired interaction. This integrated approach enables the robot to respond more intelligently, adapting its actions based on both the identified objects and the user's posture.

Overall, the combination of object detection and posture estimation in a Multimodal Intelligent Robotic System facilitates a nuanced and responsive interaction between humans and robots. This not only streamlines tasks and commands but also contributes to the creation of a more natural and intuitive collaboration, which is essential for the advancement and acceptance of robots in various human-centric environments.

4.2 Information Extraction

Information extraction involves the process of retrieving meaningful details from various sources, such as verbal instructions, to enhance the robot's understanding and response capabilities.

Verbal Instruction Processing: Information extraction often begins with the interpretation of verbal instructions provided by the user. Speech recognition algorithms analyze spoken language, converting it into text. Natural Language Processing (NLP) techniques then extract semantic meaning from the text, discerning the user's intent and commands.

Context Awareness: Beyond literal content, the system aims to grasp the contextual nuances embedded in verbal instructions. Understanding the context surrounding a command enables the robot to generate more accurate and contextually relevant responses. This involves considering factors such as the user's tone, emphasis, and the broader conversation context.

Intent Recognition: Information extraction includes identifying the user's intent behind a given instruction. This involves categorizing the extracted information into specific commands or actions that the robot should undertake. Machine learning models are often employed to train the system to recognize various intents, improving accuracy over time.

User Profiling: To personalize interactions, the system may extract information about the user, such as preferences, habits, or past interactions. By building a profile of the user, the robot can tailor responses and actions to align with the individual's specific needs and preferences.

4.3 Gesture Analysis and Eye Tracking

Gesture analysis involves interpreting and understanding the gestures made by humans, such as hand movements, body language, and facial expressions, to facilitate more natural and intuitive interactions in a Multimodal Intelligent Robotic System. This component enhances the robot's ability to comprehend non-verbal cues, adding a layer of richness to the communication process.

Algorithmic Steps for Gesture Analysis:

Data Acquisition:

Capture visual data using cameras or depth sensors to track and record gestures made by the user.

Pre-processing:

Clean and enhance the captured visual data, ensuring optimal input quality for subsequent analysis.

Feature Extraction:

Identify key features of gestures, such as hand positions, movements, and facial expressions, using computer vision techniques.

Gesture Recognition:

Train machine learning models (e.g., using deep learning frameworks like TensorFlow or OpenCV) to recognize specific gestures based on the extracted features.

Integration:

Combine gesture analysis results with other modalities, such as speech or eye tracking, to create a more comprehensive understanding of user intent.

Eye Tracking in Human-Robot Interaction (HRI):

Eye tracking involves monitoring and recording the movements of a person's eyes to understand where they are looking. In HRI, eye tracking contributes to the system's ability to discern focus, attention, and user preferences, enhancing the overall interaction experience.

Algorithmic Steps for Eye Tracking:

Eye Data Acquisition:

Utilize eye-tracking devices or cameras to capture data related to eye movements, including gaze points and fixations.

Calibration:

Calibrate the eye-tracking system to account for individual differences and ensure accurate tracking of gaze positions.

Data Processing:

Clean and preprocess the recorded eye-tracking data, filtering out noise and irrelevant information.

Feature Extraction:

Identify relevant features, such as gaze direction, duration of fixations, and eye movement patterns, through signal processing and analysis.

Gaze Estimation:

Utilize algorithms, including machine learning models or mathematical computations, to estimate the user's gaze point and track changes over time.

Integration:

Integrate eye tracking data with other modalities, such as gesture analysis or speech recognition, to create a holistic understanding of the user's intentions and preferences.

In Figure 3, the cumulative loss across epochs is computed for the proposed Multimodal Intelligent Robotic System (MIRS) and contrasted with benchmark methods such as LSTM, RNN, and Random Forest. The graph illustrates the superior performance of the MIRS method, showcasing lower loss values compared to the alternative algorithms. This reduced loss trajectory signifies enhanced system efficiency, affirming that the proposed technique consistently outperforms others. The notable advantage of minimizing loss across epochs positions the MIRS method as a promising approach for optimizing performance and reliability in various applications, particularly in the context of human-robot interaction.

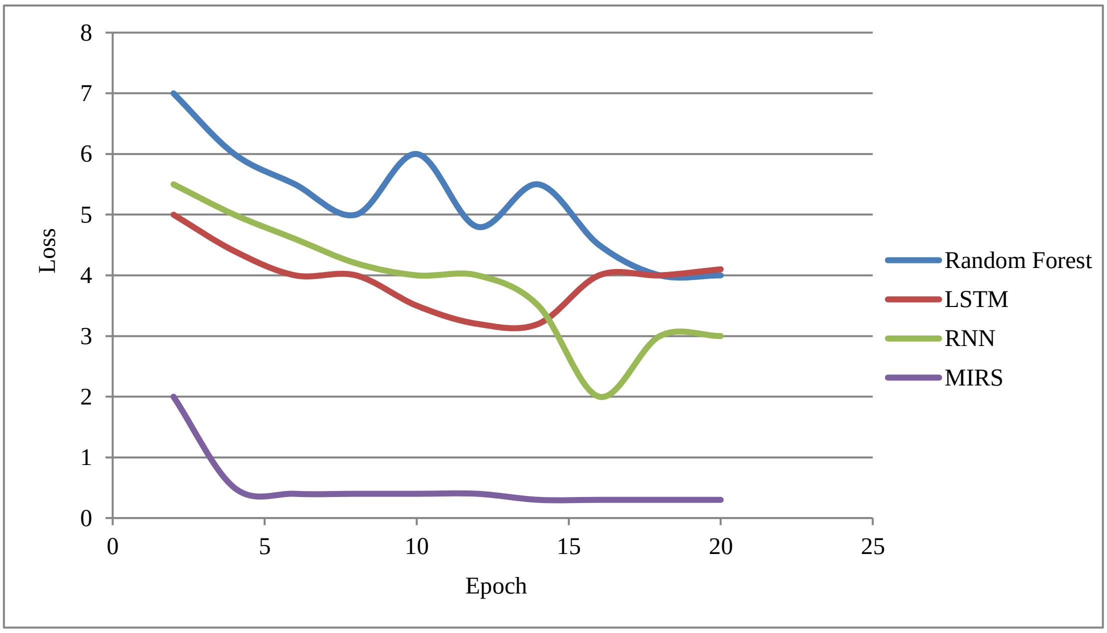

Fig. 3. Total Loss across Epochs

Table 1 organizes a comprehensive overview of the proposed Multimodal Intelligent Robotic System (MIRS) and its comparison with other methods, including detailed insights into their respective accuracy and execution times. Notably, the MIRS stand out as the top performer, achieving an impressive accuracy of 98.90%. Following closely are LSTM and Random Forest, securing the second and third positions, respectively. These results underscore the superior performance of the proposed MIRS method, emphasizing its efficacy in achieving high accuracy levels, a pivotal factor in the success of applications, particularly within the domain of human-robot interaction.

Table 1. Performance Comparison of Different Methods

Methods	**Accuracy (%)**	**Processing Time (ms)**
Random Forest	85.21	0.025
LSTM	95.70	0.036
RNN	80.65	0.011
MIRS	98.90	0.055

In Fig. 4, the cumulative accuracy of diverse models is depicted, with a particular focus on comparing the proposed Multimodal Intelligent Robotic System (MIRS) against Random Forest, LSTM, and RNN. Notably, the MIRS method emerges as the leader in accuracy among all models. While acknowledging that the processing time for MIRS is marginally higher than alternative methods, this discrepancy is justified by its superior accuracy achievement. This nuanced analysis reinforces the effectiveness of the MIRS approach, affirming its capability to outperform others in terms of accuracy, a pivotal metric in applications such as human-robot interaction.

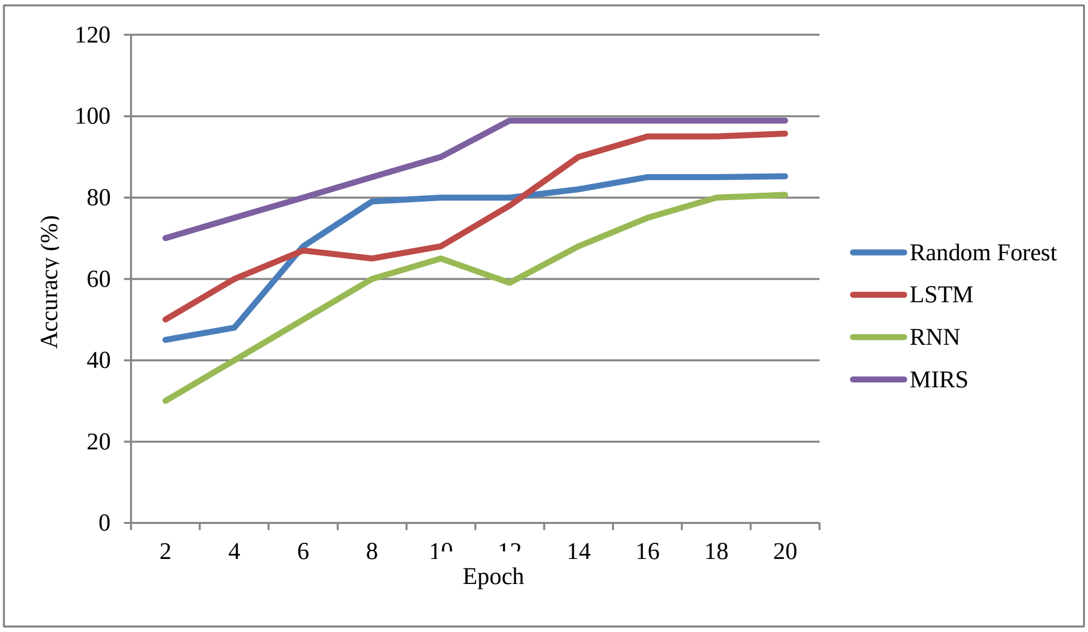

Fig. 4. Accuracy across Epochs

Multimodal Integration:

The collaborative operation of various modules within the proposed multi-modal framework showcased the power of integrating sensors such as image, sound, and depth. The system seamlessly combined information from different modalities, allowing for a more nuanced interpretation of user interactions and preferences.

User Experience and Adaptability:

The real-time adaptability embedded in the robotic framework played a pivotal role in enhancing the overall user experience. The system showcased a remarkable ability to dynamically adjust to changing environmental conditions and user contexts, contributing to a more responsive and intuitive collaboration.

Versatility of Multimodal HRI:

The adoption of Multimodal HRI emerged as a cornerstone in this research, offering users a diverse array of communication modalities, including voice, images, text, eye movement, touch, and bio-signals. This versatility marked a significant paradigm shift in HRI, fostering enhanced communication and engagement between humans and robots.

Future Implications and Challenges:

The positive outcomes of the experimental evaluations open avenues for future research in the realm of adaptable and context-aware robotics. Challenges, such as further refining gesture recognition algorithms and expanding the range of bio-signals for integration, represent exciting prospects for future developments.

5. Conclusion

In conclusion, the research underscores the pivotal role of human-robot interaction (HRI) in advancing robotics to new heights. The integration of context-aware robotic systems and real-time adaptability has paved the way for a more responsive and intuitive collaboration between humans and robots. The emphasis on innovative HRI strategies, particularly the adoption of multimodal approaches, signifies a paradigm shift towards fostering natural and adaptable interactions. The proposed Multi-Modal Intelligent Robotic System (MIRS) demonstrates the efficacy of leveraging diverse modalities, including voice, images, text, eye movement, touch, and bio-signals, to enhance communication between humans and robots. A multimodal intelligence robotic system is developed and evaluated to find the best possible method. The result shows that the proposed framework provides the highest accuracy than the state-of-the- art techniques with proper execution time. Also the loss is very less in the proposed approach when compared to other methods. Therefore, the proposed multimodal intelligent robotic system revolutionizes human-robot interaction and enhances collaboration.

References

[1] Tonk, A., Dhabliya, D., Sheril, S., Abbas, A. H., & Dilsora, A. (2023). Intelligent Robotics: Navigation, Planning, and Human-Robot Interaction. In *E3S Web of Conferences* (Vol. 399, p. 04044). EDP Sciences.

[2] Su, H., Qi, W., Chen, J., Yang, C., Sandoval, J., & Laribi, M. A. (2023). Recent advancements in multimodal human–robot interaction. *Frontiers in Neurorobotics*, *17*, 1084000.

[3] Chandan, K. D. (2023). *Bridging the Observability Gap: Augmented Reality Policies for Human Robot Collaboration* (Doctoral dissertation, State University of New York at Binghamton).

[4] Tallat, R., Hawbani, A., Wang, X., Al-Dubai, A., Zhao, L., Liu, Z., ... & Alsamhi, S. H. (2023). Navigating Industry 5.0: A Survey of Key Enabling Technologies, Trends, Challenges, and Opportunities. *IEEE Communications Surveys & Tutorials*.

[5] Frijns, H. A., Schürer, O., & Koeszegi, S. T. (2023). Communication models in human–robot interaction: an asymmetric MODel of ALterity in human–robot interaction (AMODAL-HRI). *International Journal of Social Robotics*, *15*(3), 473-500.

[6] Shafti, A., Orlov, P., & Faisal, A. A. (2019, May). Gaze-based, context-aware robotic system for assisted reaching and grasping. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 863-869). IEEE.

[7] Fang, B., Wei, X., Sun, F., Huang, H., Yu, Y., & Liu, H. (2019). Skill learning for human-robot interaction using wearable device. *Tsinghua Science and Technology*, *24*(6), 654-662.

[8] Zheng, T. W. P., Wang, S. L. L., Wang, T., Zheng, P., Li, S., & Wang, L. Multimodal Human-Robot Interaction for Human-centric Smart Manufacturing: A Survey.

[9] Rautiainen, S., Pantano, M., Traganos, K., Ahmadi, S., Saenz, J., Mohammed, W. M., & Martinez Lastra, J. L. (2022). Multimodal interface for human–robot collaboration. *Machines*, *10*(10), 957.

[10] Abbasi, B., Monaikul, N., Rysbek, Z., Di Eugenio, B., & Žefran, M. (2019, November). A multimodal human-robot interaction manager for assistive robots. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 6756-6762). IEEE.

[11] M D, R. ., Kenchannavar, H. H. ., & Kulkarni, U. P. . (2022). Facial Emotion Recognition using Three-Layer ConvNet with Diversity in Data and Minimum Epochs. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(4), 264–268. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2225

[12] Li, S., Zheng, P., Liu, S., Wang, Z., Wang, X. V., Zheng, L., & Wang, L. (2023). Proactive human–robot collaboration: Mutual-cognitive, predictable, and self-organising perspectives. *Robotics and Computer-Integrated Manufacturing*, *81*, 102510.

[13] Shervedani, A. M., Li, S., Monaikul, N., Abbasi, B., Di Eugenio, B., & Zefran, M. (2023). An End-to-End Human Simulator for Task-Oriented Multimodal Human-Robot Collaboration. *arXiv preprint arXiv:2304.00584*.