10.48047/jocaaa.2024.34.01.22

AI-Powered Gesture Recognition for Human Interaction: Enhancing Natural Communication between Humans and Machines

FNU Harsh1, Smarth Behl 2, Soumya Banerjee 3
1. Software Engineer
2. Software Engineer, Mountain View,
3. Engineering Manager

Abstract

The growing demand for natural and intuitive human-machine interaction has driven significant advancements in gesture recognition technologies. This study presents an AI-powered gesture recognition framework that leverages a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) architecture to accurately interpret human hand gestures in real-time. A multimodal dataset was collected using RGB/depth cameras and EMG/inertial sensors across diverse user demographics and environmental conditions. The proposed system achieved a peak accuracy of 95.2% and demonstrated superior performance compared to baseline models such as standalone CNNs, LSTMs, and transfer learning approaches like ResNet50 and MobileNetV2. Usability evaluations confirmed high user satisfaction, with System Usability Scale (SUS) scores averaging above 80. The system also exhibited strong real-time performance across edge devices, including Jetson Nano and Raspberry Pi 4. While performance declined slightly under variable lighting and among differently-abled users, the overall results highlight the model's robustness and scalability. This study underscores the potential of AI-powered gesture recognition in transforming human-machine communication, especially in applications such as assistive technology, smart environments, and interactive robotics. Future work will focus on enhancing personalization, environmental adaptability, and inclusivity through adaptive learning and sensor fusion.

Keywords: Gesture Recognition, Human-Machine Interaction, AI, CNN-LSTM, Real-time System, Multimodal Input, Usability, Edge Computing.

Introduction

Background and motivation

The rapid evolution of artificial intelligence (AI) has transformed the way humans interact with machines, moving beyond traditional interfaces such as keyboards, touchscreens, and voice commands (Guo et al., 2023). One of the most promising frontiers in this evolution is gesture recognition, which enables machines to interpret human body language and hand movements

10.48047/jocaaa.2024.34.01.22

as inputs. Gesture-based interaction, being non-verbal and intuitive, closely resembles natural human communication and can serve as a bridge toward seamless human-machine interaction (HMI). This modality holds significant potential across various domains, including smart homes, gaming, healthcare, virtual reality, robotics, and assistive technologies for individuals with disabilities (Patil, 2024).

Despite its growing relevance, existing gesture recognition systems face challenges in accurately capturing and interpreting a wide range of dynamic human gestures across diverse users and environments (Dubey et al., 2024). Traditional rule-based or vision-based approaches often suffer from limited adaptability, sensitivity to environmental changes, and lack of contextual understanding. The advent of AI, particularly deep learning, offers a paradigm shift by allowing systems to learn complex patterns in human movement from large datasets, thereby enabling more robust and scalable gesture recognition solutions (Harshini et al., 2024).

AI in gesture recognition: a paradigm shift

AI-powered gesture recognition leverages advanced algorithms such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based architectures to process and analyze visual, motion, or sensor data (Jixuan, 2024). These models have demonstrated remarkable accuracy in detecting and classifying static and dynamic gestures in real-time. When integrated with computer vision techniques and sensor fusion methods, AI-driven systems can recognize gestures with high precision, even under varying lighting conditions, occlusions, or across different cultural gesture norms (Sadiq & Saraswathi, 2024).

Moreover, AI facilitates continuous learning and personalization, enabling gesture recognition systems to adapt to individual user behaviors over time (Mukherjee et al., 2022). This adaptability is particularly critical in applications such as healthcare monitoring or assistive robotics, where user-specific interactions are key to effectiveness and comfort. With the growth of edge computing and embedded AI, these intelligent systems are now being deployed on portable or wearable devices, making gesture-based interfaces more accessible and responsive than ever before (Nama, 2023).

Enhancing human-machine communication

The integration of AI-powered gesture recognition in human-machine communication represents a significant advancement toward natural user interfaces (NUIs) (Awad et al., 2024).

10.48047/jocaaa.2024.34.01.22

Such systems enable users to communicate with machines in a way that mimics human-to-human interaction, thereby reducing the cognitive load and learning curve associated with traditional input methods. In smart environments, users can control devices with simple hand motions; in virtual or augmented reality, immersive gesture-based control enhances user experience and realism (Kotti et al., 2024). For the differently-abled community, gesture recognition offers a powerful tool to bridge communication gaps, such as translating sign language into speech or text.

In human-robot interaction (HRI), gesture recognition plays a crucial role in facilitating intuitive commands and cooperative tasks (Srikanth et al., 2023). For instance, robots that understand gestures can assist humans in industrial settings, healthcare scenarios, or domestic environments without the need for complex programming or spoken commands. The synergy between AI and gesture recognition thus not only increases the efficiency of interactions but also fosters trust, engagement, and emotional connection between humans and machines (Al-Shayeb et al., 2024).

Challenges and research opportunities

While AI-powered gesture recognition systems have shown promising results, several challenges remain. These include high computational costs, privacy concerns related to continuous video monitoring, and the need for large annotated datasets for training. Furthermore, ensuring fairness and minimizing bias in gesture interpretation across different demographic groups remains a critical research focus. Real-world deployment also requires models to be energy-efficient and capable of operating under low-resource constraints.

Consequently, this research aims to develop a scalable, accurate, and context-aware gesture recognition framework powered by AI that enhances natural communication between humans and machines. By exploring the integration of multimodal data inputs, adaptive learning techniques, and efficient model architectures, the study seeks to address current limitations and pave the way for more inclusive and intuitive HMI systems.

Methodology

This study adopts a comprehensive and iterative methodology to design, implement, and evaluate an AI-powered gesture recognition system aimed at enhancing natural communication between humans and machines. The methodology involves several key stages, including data

10.48047/jocaaa.2024.34.01.22

collection, preprocessing, model development, system integration, and performance evaluation. A combination of visual and sensor-based data was used to build a robust gesture recognition framework capable of real-time interaction in dynamic environments.

Research design

The research follows a design-based research (DBR) approach, allowing for iterative refinement of the gesture recognition system through continuous testing and development. It is experimental in nature and geared toward practical implementation, integrating AI techniques with human-computer interaction (HCI) principles to ensure real-world applicability. The system is evaluated across both lab-based setups and semi-natural environments to validate its robustness and generalizability.

Data collection

To build a diverse and reliable dataset, gesture data were collected from 50 participants across different demographic groups. Two primary input modalities were used: video-based capture using Azure Kinect DK and a standard HD webcam, as well as inertial and electromyography (EMG) signals obtained using the Myo armband. Participants were asked to perform 25 predefined static and dynamic hand and body gestures, resulting in a total of 3,000 recorded gesture samples. These gestures were selected based on their frequency and relevance in daily interactions and human-robot communication contexts.

Preprocessing and data augmentation

All collected data underwent rigorous preprocessing to prepare it for model training. Visual data were normalized and resized to a standard resolution of 224×224 pixels. Gesture sequences were segmented using a temporal sliding window approach, isolating individual gestures from continuous video streams. For sensor data, normalization techniques were applied to ensure consistency across sessions. Data augmentation was crucial for improving model generalization and included transformations such as rotation, mirroring, Gaussian noise injection, and frame dropout, simulating variations in gesture execution and environmental conditions.

Model development

10.48047/jocaaa.2024.34.01.22

A deep learning-based architecture was developed for gesture recognition, utilizing a hybrid CNN-LSTM model. Convolutional Neural Networks (CNNs) were employed to extract spatial features from gesture images and video frames, while Long Short-Term Memory (LSTM) networks captured temporal dynamics in sequential gesture data. To improve robustness, especially in complex or occluded environments, a multimodal fusion strategy was employed. Visual data were combined with EMG and inertial sensor inputs using an attention-based fusion layer. Transfer learning techniques, particularly pre-trained models such as ResNet50 and MobileNetV2, were utilized for efficient feature extraction, enabling faster convergence and improved accuracy. The model was trained using the Adam optimizer with a learning rate of 0.001 and batch size of 32, incorporating early stopping to prevent overfitting.

System integration

To facilitate real-time application, the gesture recognition model was embedded into a user interface developed in Python using TensorFlow, Keras, and OpenCV libraries. The system was deployed on edge devices, including Raspberry Pi 4 and NVIDIA Jetson Nano, enabling portable and responsive implementation. Gesture-to-command mapping allowed for seamless control of smart devices, robotic actuators, and virtual environments. Inference speed was optimized using TensorRT acceleration, reducing latency to below 150 milliseconds and supporting natural, fluid interactions.

Evaluation metrics

The system's performance was evaluated using multiple quantitative and qualitative metrics. Accuracy, precision, recall, and F1-score were used to measure classification effectiveness. Latency was measured to ensure real-time performance. To assess user experience, usability testing was conducted using the System Usability Scale (SUS). Robustness was further tested under variable lighting, background clutter, and occlusion conditions. Statistical analysis, including ANOVA, was used to compare system performance against baseline gesture recognition approaches and validate the significance of improvements.

Validation and benchmarking

To verify the effectiveness of the proposed model, comparative validation was performed using conventional rule-based and vision-only gesture recognition systems. The hybrid AI-powered model consistently outperformed baseline systems, demonstrating superior adaptability,

10.48047/jocaaa.2024.34.01.22

accuracy, and responsiveness. These findings establish the potential of the proposed methodology to contribute meaningfully to the development of natural user interfaces and intelligent human-machine interaction platforms.

Results

Table 1: Gesture Accuracy by Type

Gesture Type	Accuracy (%)	Precision (%)	Recall (%)
Wave	95.4	94.7	95.8
Thumbs Up	93.2	92.5	93.0
Point	91.8	91.0	91.5
Fist	92.5	91.8	92.0
Palm Open	94.1	93.5	94.0

The AI-powered gesture recognition system demonstrated strong performance across various evaluation metrics and test conditions. As shown in Table 1, the system achieved high accuracy in recognizing common hand gestures such as Wave (95.4%), Palm Open (94.1%), and Thumbs Up (93.2%), with corresponding precision and recall values also exceeding 91% for all gestures. The "Wave" gesture was the most accurately recognized, with a precision of 94.7% and recall of 95.8%, indicating the model's strong capability in classifying distinct gesture patterns.

Table 2: Model Performance Comparison

Model	Accuracy (%)	F1 Score (%)	Training Time (min)
CNN	89.6	89.2	22
LSTM	87.3	86.5	30
CNN-LSTM (Proposed)	95.2	95.0	45
ResNet50	93.8	93.2	38
MobileNetV2	92.5	91.9	34

The comparative evaluation of different models is presented in Table 2. The proposed CNN-LSTM hybrid model outperformed the standalone CNN and LSTM models, as well as transfer

10.48047/jocaaa.2024.34.01.22

learning models like ResNet50 and MobileNetV2. It achieved the highest accuracy of 95.2% and an F1 score of 95.0%, although it required a slightly longer training time (45 minutes). Among the baseline models, ResNet50 performed well with an accuracy of 93.8% and F1 score of 93.2%, followed closely by MobileNetV2.

Table 3: Lighting Condition Impact

Lighting Condition	Accuracy (%)	Latency (ms)
Bright	95.6	120
Dim	91.2	135
Variable	89.5	145

Environmental factors also influenced the system's performance, as detailed in Table 3. Under bright lighting conditions, the recognition accuracy reached 95.6% with a latency of 120 ms. However, in variable lighting, the accuracy dropped slightly to 89.5%, and latency increased to 145 ms. This indicates the system's resilience to ambient conditions but highlights the need for further optimization in dynamic environments.

Table 4: Device-wise Latency and Speed

Device	Latency (ms)	Inference Speed (FPS)
Raspberry Pi 4	148	15
Jetson Nano	132	18
PC (i7 GPU)	90	30
Smartphone (Edge AI)	160	12

System deployment on various hardware platforms was assessed and summarized in Table 4. The lowest latency (90 ms) and highest inference speed (30 FPS) were observed on a PC equipped with an i7 GPU. The Jetson Nano also delivered respectable performance (132 ms latency and 18 FPS), making it suitable for portable and edge-based deployment. Raspberry Pi 4 and smartphones exhibited slightly higher latency, suggesting they are better suited for low-frequency or less time-critical applications.

Table 5: User Group-wise Performance

10.48047/jocaaa.2024.34.01.22

User Group	Recognition Accuracy (%)	SUS Score (/100)
Adults	96.0	85
Children	92.3	82
Elderly	90.5	78
Differently-abled	88.4	80

User-centric evaluation results in Table 5 revealed that the gesture recognition system was well-received across different demographic groups. Adults achieved the highest recognition accuracy at 96.0% and reported the highest usability score of 85 on the System Usability Scale (SUS). Performance was also strong among children and elderly users, though slightly reduced in the differently-abled group, where recognition accuracy was 88.4% and SUS score was 80. These findings suggest that while the system is broadly effective, future iterations should consider personalization and adaptive learning to better serve users with diverse physical and cognitive needs.

Discussion

Effectiveness of the proposed gesture recognition system

The results clearly indicate that the AI-powered gesture recognition system demonstrates a high level of accuracy and reliability in classifying a variety of hand gestures. As highlighted in Table 1, common gestures such as “Wave” and “Palm Open” were recognized with over 94% accuracy, suggesting that the system is well-suited for intuitive and natural human-machine interaction. The strong performance in precision and recall values confirms the model's ability to correctly identify gestures with minimal false positives or missed detections. These results reinforce the effectiveness of using a hybrid CNN-LSTM architecture for extracting both spatial and temporal features of gestures, providing superior pattern recognition compared to traditional models.

Model comparison and superiority of CNN-LSTM architecture

The comparative analysis in Table 2 shows that the CNN-LSTM hybrid model significantly outperforms other baseline models in both accuracy and F1 score. While CNNs are effective for spatial feature extraction, their limitations in handling temporal dependencies were mitigated by the integration of LSTM layers. This synergy between spatial and temporal

10.48047/jocaaa.2024.34.01.22

modeling is what enabled the CNN-LSTM model to outperform even powerful pre-trained models like ResNet50 and MobileNetV2. Although the proposed model had a longer training time, the resulting accuracy of 95.2% justifies the additional computational effort. This finding confirms the hypothesis that combining multiple neural architectures can yield better results in complex recognition tasks like gesture classification (Wang et al., 2024).

Impact of environmental conditions on system performance

As shown in Table 3, lighting conditions had a noticeable impact on system accuracy and latency. Bright environments yielded the highest accuracy (95.6%) and lowest latency (120 ms), while variable lighting reduced performance to 89.5% accuracy with increased latency. These findings underscore the sensitivity of vision-based AI systems to environmental noise and highlight the importance of integrating adaptive preprocessing techniques or sensor fusion methods to maintain performance across diverse settings. This also suggests that while the system is robust, real-world deployment would benefit from dynamic calibration or multi-modal input strategies to offset fluctuations in ambient lighting (Devi et al., 2024).

Hardware deployment and real-time capabilities

The practical utility of the system across different platforms is evident from Table 4. High-end devices like PCs with GPU support naturally offered the best performance in terms of latency (90 ms) and inference speed (30 FPS), making them ideal for deployment in high-performance settings like VR environments or industrial robotics. On the other hand, Jetson Nano provided a favorable balance between portability and performance, enabling deployment in mobile and edge AI applications (Xu et al., 2023). The results also show that the system remains functional, albeit slower, on devices like Raspberry Pi and smartphones, demonstrating its scalability and flexibility in resource-constrained scenarios (Zhou et al., 2024).

User group variability and usability considerations

The system's performance across diverse user groups, as reported in Table 5, is encouraging, especially with adults achieving 96% accuracy and the highest usability score of 85. The slightly lower performance among the elderly and differently-abled users (90.5% and 88.4% accuracy, respectively) suggests that individual variability in motor control or physical limitations may affect recognition consistency (ZainEldin et al., 2024). These insights point to the need for incorporating adaptive learning capabilities that allow the system to personalize

10.48047/jocaaa.2024.34.01.22

its response based on user-specific gesture patterns. It also calls for further refinement in gesture set design to ensure accessibility and inclusivity for all users (Casheekar et al., 2024).

Limitations and future directions

While the results validate the proposed system's capabilities, certain limitations remain. The system's reduced performance in variable lighting and among differently-abled users signals the need for further enhancement through techniques like transfer learning on specialized datasets, multi-camera systems, or depth and EMG sensor integration. Furthermore, expanding the gesture vocabulary and incorporating continuous gesture detection in natural conversation scenarios would greatly improve the system's applicability in real-world HCI.

the AI-powered gesture recognition framework demonstrates strong potential for enhancing natural communication between humans and machines. The hybrid CNN-LSTM model, coupled with robust system design and hardware adaptability, lays the foundation for scalable deployment in real-world environments. However, to fully realize its potential, future iterations must focus on personalization, environmental adaptability, and inclusive design to support all user demographics equally.

Conclusion

This study successfully demonstrates the development and effectiveness of an AI-powered gesture recognition system designed to enhance natural human-machine interaction. By leveraging a hybrid CNN-LSTM architecture and multimodal input data, the system achieved high accuracy, responsiveness, and usability across various environmental conditions and user groups. The results validate the proposed framework's robustness, particularly in real-time applications involving diverse gestures and hardware platforms. Furthermore, the system's scalability and adaptability make it suitable for integration into smart devices, assistive technologies, and interactive robotics. However, challenges remain in optimizing performance under dynamic lighting conditions and ensuring inclusivity for differently-abled users. Addressing these limitations through adaptive learning and sensor fusion can further enhance the system's reliability and accessibility. Overall, the study establishes a strong foundation for future advancements in AI-driven gesture-based communication interfaces, contributing meaningfully to the evolution of more intuitive, inclusive, and human-centric machine interaction systems.

10.48047/jocaaa.2024.34.01.22

References

Al-Shayeb, I. E., Abro, G. E. M., Khan, F. S., Boudville, R., & Abdallah, A. M. (2024, August). Integrating AI-Driven Robust Control Algorithm with 3D Hand Gesture Recognition to Track an Underactuated Quadrotor Unmanned Aerial Vehicle (QUAV). In *2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 70-75). IEEE.

Awad, A. I., Babu, A., Barka, E., & Shuaib, K. (2024). AI-powered biometrics for Internet of Things security: A review and future vision. *Journal of Information Security and Applications*, *82*, 103748.

Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K. S., & Srinivasan, K. (2024). A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Computer Science Review*, *52*, 100632.

Devi, D. S. R., BhagyaSri, O. U. C., Sravanthi, R., Chaitrika, S. L., Priyanka, M. N., Swarna, M., & Srilekha, M. (2024). AI-Enhanced Cursor Navigator. *R. and Chaitrika, SL and Priyanka, MN and Swarna, M. and Srilekha, M., AI-Enhanced Cursor Navigator (May 10, 2024)*.

Dubey, P., Bhagat, P. N., Anjum, G. H., & Akhter, P. (2024). Signs Unveiled: The Power and Promise of AI-Based Sign Recognition Systems. In *AI in the Social and Business World: A Comprehensive Approach* (pp. 118-138). Bentham Science Publishers.

Guo, K., Orban, M., Lu, J., Al-Quraishi, M. S., Yang, H., & Elsamanty, M. (2023). Empowering hand rehabilitation with ai-powered gesture recognition: A study of an semg-based system. *Bioengineering*, *10*(5), 557.

Harshini, P. J., Rahman, M. A., Raj, J. A., & Durgadevi, P. (2024, May). Sign Language Recognition System for Seamless Human-AI Interaction. In *International Research Conference on Computing Technologies for Sustainable Development* (pp. 124-144). Cham: Springer Nature Switzerland.

Jixuan, W. (2024, July). Research on Key Technologies of Human-Computer Interaction in Artificial Intelligence-Based Immersive Virtual Reality. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)* (pp. 1384-1390). IEEE.

10.48047/jocaaa.2024.34.01.22

Kotti, J., Padmaja, B., & Deepa, D. (2024). Enhancing Gesture-Controlled Virtual Mouse and Virtual Keyboard Using AI Techniques. *Journal of Mobile Multimedia*, *20*(2), 473-493.

Mukherjee, D., Gupta, K., & Najjaran, H. (2022, August). An ai-powered hierarchical communication framework for robust human-robot collaboration in industrial settings. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1321-1326). IEEE.

Nama, P. (2023). AI-Powered Mobile Applications: Revolutionizing User Interaction Through Intelligent Features and Context-Aware Services. *Journal of Emerging Technologies and Innovative Research*, *10*(01), g611-g620.

Patil, N. (2024). Gesture Voice: Revolutionizing Human-Computer Interaction with an AI-Driven Virtual Mouse System|. *662b3ece15f5a428a6035b02*, *15*(3).

Sadiq, S., & Saraswathi, S. (2024). Enhance the AI Virtual System Accuracy with Novel Hand Gesture Recognition Algorithm Comparing to Convolutional Neural Network. In *E3S Web of Conferences* (Vol. 491, p. 04022). EDP Sciences.

Srikanth, K., Venuthurumilli, M., Manaswini, N. V., & Somayajulu, M. V. N. N. S. S. R. K. S. (2023, March). Creating An Artificial Intelligence-Powered Image Classification Model For Specially Abled Persons. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 2275-2280). IEEE.

Wang, H., Ding, Q., Luo, Y., Wu, Z., Yu, J., Chen, H., ... & Wu, J. (2024). High-performance hydrogel sensors enabled multimodal and accurate human–machine interaction system for active rehabilitation. *Advanced Materials*, *36*(11), 2309868.

Xu, C., Solomon, S. A., & Gao, W. (2023). Artificial intelligence-powered electronic skin. *Nature machine intelligence*, *5*(12), 1344-1355.

ZainEldin, H., Gamel, S. A., Talaat, F. M., Aljohani, M., Baghdadi, N. A., Malki, A., ... & Elhosseini, M. A. (2024). Silent no more: a comprehensive review of artificial intelligence, deep learning, and machine learning in facilitating deaf and mute communication. *Artificial Intelligence Review*, *57*(7), 188.

Zhou, H., Tawk, C., & Alici, G. (2024). A multipurpose human–machine interface via 3D-printed pressure-based force myography. *IEEE Transactions on Industrial Informatics*.

10.48047/jocaaa.2024.34.01.22