IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154165

AI-Driven Gesture Tracking: A Comprehensive Review of Techniques, Applications, and Future Directions

Huda Khan[1],Isha Kaushal[2],Gunjan Rani[3],Nikhil Kumar[4], Mr. K.S. Mishra[5]

Department of MCA, MIET[1-5]

Abstract: Gesture tracking, the technological ability to interpret human movements as commands, has become a cornerstone of modern Human-Computer Interaction (HCI). Propelled by advances in artificial intelligence (AI), particularly deep learning, gesture recognition systems have evolved from laboratory experiments to practical applications in robotics, manufacturing, healthcare, and consumer electronics. This paper provides a comprehensive review of AI techniques for gesture tracking, spanning the last decade of research. We systematically analyze the gesture recognition pipeline, from data acquisition methods (vision-based, sensor-based) to feature extraction and classification algorithms. The review contrasts traditional machine learning approaches like Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) with modern deep learning architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers. A significant focus is placed on training strategies, such as multi-modal fusion and ModDrop, which enhance robustness in real-world conditions. Furthermore, we explore key application domains—Human-Robot Interaction (HRI), Industry 5.0, Augmented Reality (AR), and drone control—highlighting how AI techniques are tailored to meet specific domain challenges. The review concludes by identifying persistent challenges, including occlusion, environmental variability, and the need for large, annotated datasets, and proposes future research directions towards more adaptive, multi-modal, and human-centric gesture recognition systems.

Keywords: Gesture Recognition, Human-Computer Interaction, Deep Learning, Computer Vision, Human-Robot Interaction, Industry 5.0, Multi-modal Fusion

1. INTRODUCTION

The quest to make technology more intuitive and accessible has long driven research in Human-Computer Interaction (HCI). Among various interaction modalities, gesture recognition holds a unique position. Gestures are a fundamental and natural part of human communication, used to emphasize speech, convey intent, and interact with our environment (Yasen & Jusoh, 2019). Translating this innate ability to the digital realm—enabling machines to see, interpret, and respond to human gestures—has been a persistent and evolving challenge.

In its infancy, gesture recognition relied on cumbersome hardware like data gloves and simple computer vision techniques constrained by controlled environments. However, the last decade has witnessed a paradigm shift, driven by the convergence of increased computational power, the proliferation of affordable depth sensors (e.g., Microsoft Kinect, Leap Motion), and most importantly, the revolution in artificial intelligence (AI). Machine learning, and particularly deep learning, has transformed the field, enabling systems that can understand complex, dynamic gestures with high accuracy, even in the presence of background clutter and varying illumination (Hussain et al., 2024).

This review aims to provide a holistic and structured overview of the AI techniques that power modern gesture tracking systems. By synthesizing findings from seminal and recent research, we chart the evolution of the field, analyze the state-of-the-art, and explore its most impactful applications. The scope of this review is defined by the following key objectives:

- To deconstruct the gesture recognition pipeline and analyze the AI techniques employed at each stage, from data acquisition to classification.
- To compare and contrast traditional machine learning approaches with modern deep learning architectures for gesture recognition.
- To investigate advanced training methodologies, particularly multi-modal fusion, that enhance system robustness.
- To explore how AI-driven gesture tracking is being applied across diverse domains, including robotics, industry, augmented reality, and drone control.
- To identify persistent challenges and propose future research directions that will shape the next generation of gesture-based interfaces.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154165

2. THE GESTURE RECOGNITION PIPELINE

Regardless of the final application, most AI-driven gesture recognition systems follow a structured pipeline. This process transforms raw data about human movement into a meaningful command. The core stages, as synthesized from the literature (Yasen & Jusoh, 2019; Hussain et al., 2024), are data acquisition, preprocessing and segmentation, feature extraction, and classification.

2.1 Data Acquisition

The first and most critical step is capturing the gesture itself. The choice of acquisition method fundamentally shapes the subsequent AI techniques that can be applied. Acquisition methods can be broadly divided into two categories: vision-based and sensor-based.

2.1.1 Vision-Based Methods

These methods use cameras to capture gesture data non-invasively. Standard webcams are the most accessible option; they capture color images processable through computer vision techniques, though they are highly susceptible to lighting conditions and background clutter, making robust segmentation challenging (Yasen & Jusoh, 2019). Depth cameras (RGB-D), such as the Microsoft Kinect and Intel RealSense, have been transformative by providing depth information alongside color, enabling 3D hand pose estimation and significantly improving feature extraction robustness (Hussain et al., 2024). Time-of-Flight (ToF) cameras provide high-resolution depth maps by measuring the time light takes to travel to a scene and back; they are effective for specific applications, such as in-car touchless interfaces (Yasen & Jusoh, 2019).

2.1.2 Sensor-Based Methods

These methods require the user to wear or hold devices that directly measure motion. Data gloves employ flex sensors, accelerometers, and gyroscopes to capture fine-grained finger movement and coarse hand motion. This multi-sensor fusion enables high-precision gesture capture in low-light or cluttered environments where vision-based systems fail. Infrared controllers, such as the Leap Motion Controller, track hands and fingers with high precision within a limited field of view and are noted for high repeatability. Surface Electromyography (sEMG) sensors measure the electrical activity produced by muscles during contraction, allowing gesture recognition based on muscle activation patterns even before visible movement occurs—Yasen & Jusoh (2019) identified sEMG as the most prominent acquisition tool in their systematic review. At the frontier of HCI, EEG headsets measure brain activity for direct brain-computer interfaces (IEEE, 2023).

2.2 Preprocessing and Segmentation

Once data is acquired, it must be cleaned and relevant parts isolated. Preprocessing involves techniques such as image resizing to reduce computational load, applying Gaussian or median filters to remove noise, and edge detection (e.g., Canny, Sobel operators) to identify sharp boundaries (Hussain et al., 2024). Segmentation isolates the gesture from the background using skin-color detection, edge-based contour methods, or depth-based thresholding for RGB-D and ToF cameras—the last of these is highly robust to visual clutter (Hussain et al., 2024).

2.3 Feature Extraction

Feature extraction represents raw pixel or sensor data as a compact set of meaningful descriptors. Global features describe the overall shape and motion of the hand—its area, orientation, velocity, and trajectory—and are computationally efficient but sensitive to occlusion. Local features focus on distinctive points or regions, such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), Speeded-Up Robust Features (SURF), and Local Binary Patterns (LBP), offering greater robustness to occlusion and viewpoint changes. Ahmed & Aly (as cited in Hussain et al., 2024) reported that combining LBP with PCA achieved 99.97% accuracy on their dataset.

3. AI CLASSIFICATION TECHNIQUES

The extracted features are fed into a classifier—the core AI component—which learns to map feature sets to specific gesture labels. This section reviews the evolution from classical machine learning to deep learning.

3.1 Classical Machine Learning Approaches

Before the deep learning era, gesture recognition was dominated by classical algorithms requiring manual feature engineering.

3.1.1 Support Vector Machines (SVMs)

SVMs find an optimal hyperplane to separate different classes in a high-dimensional feature space. They are effective for static gesture recognition and are considered to provide moderate accuracy (Hussain et al., 2024).

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154165

3.1.2 Hidden Markov Models (HMMs)

HMMs are statistical models well-suited for sequential data, making them a popular choice for dynamic gesture recognition. They model a gesture as a sequence of hidden states and handle temporal variations effectively (Hussain et al., 2024).

3.1.3 Artificial Neural Networks (ANNs)

Even before the rise of deep learning, shallow ANNs were widely applied. Yasen & Jusoh (2019) found ANNs to be the most applied classifier in their reviewed studies from 2016–2018.

3.2 Deep Learning Architectures

Deep learning automates feature engineering. By using many-layered neural networks, these models learn hierarchical features directly from raw data—from simple edges to complex hand configurations.

3.2.1 Convolutional Neural Networks (CNNs)

CNNs are the de facto standard for processing image data. They use convolutional and pooling layers to learn spatial hierarchies of features, taking raw images or depth maps as input and directly outputting a gesture class. Hussain et al. (2024) rate CNNs as providing "high" accuracy, surpassing traditional methods. In manufacturing, Qin et al. (2023, as cited in Bertolotti et al., 2025) successfully applied skeletal data-based CNN classifiers for action recognition on assembly lines.

3.2.2 Recurrent Neural Networks (RNNs) and LSTMs

RNNs are designed for sequential data, making them ideal for dynamic gestures. Long Short-Term Memory (LSTM) networks address the vanishing gradient problem, enabling learning of long-term dependencies. In AR applications, RNNs are crucial for interpreting the flow of a user's movement (Carter, 2023).

3.2.3 Transformers

Originally developed for natural language processing, Transformers have recently achieved state-of-the-art results in computer vision. Their core innovation, the attention mechanism, allows the model to weigh the importance of different parts of the input. For gesture recognition, a Transformer can focus on the most informative joints (e.g., fingertips) at each timestep. Carter (2023) highlights Transformer models as a key advancement for improving user experience in complex AR environments.

4. ADVANCED TRAINING STRATEGIES

4.1 Multi-Modal Fusion and ModDrop

Human communication is inherently multi-modal. To create truly intuitive HRI, systems must fuse information from multiple sensors. The ModDrop technique, introduced by Neverova et al. (2015), is a pioneering approach: during training, data from one or more modalities (e.g., video, depth, audio) is randomly dropped out. This forces the network to learn robust, correlated features across modalities without becoming overly reliant on any single one, resulting in a system that performs meaningfully even if a sensor fails or data is noisy. Neverova et al. (2015) demonstrated this approach by winning the ChaLearn 2014 Looking at People Challenge, showing that fusing multiple modalities at several spatial and temporal scales significantly increases recognition rates.

4.2 Static vs. Dynamic Gesture Recognition

A comprehensive system must handle both static poses (e.g., a peace sign) and dynamic movements (e.g., a hand wave). Static gesture recognition typically relies on CNNs to analyze the spatial configuration of the hand in a single frame. Dynamic gesture recognition requires a temporal model, such as an RNN-LSTM or a 3D CNN (which convolves across both space and time), to analyze the gesture across a sequence of frames (Hussain et al., 2024).

4.3 Real-Time Processing Constraints

For applications like drone control or AR, latency is critical: the entire pipeline—from image capture to command execution—must complete in milliseconds. This imposes significant constraints on algorithm choice. A highly accurate but slow Transformer model might be unsuitable, whereas a faster, optimized CNN may be more appropriate. Research on the Tello EDU drone (IEEE, 2023) explicitly focuses on techniques lightweight enough to run on embedded systems.

5. APPLICATION DOMAINS

5.1 Human-Robot Interaction (HRI)

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154165

This is perhaps the most active area of research. Gestures provide an intuitive way for humans to command and collaborate with robots. Hussain et al. (2024) provide a comprehensive review highlighting how vision-based systems allow robots to perceive and interpret human actions—from simple directional commands to complex collaborative tasks like object handovers. The integration of deep learning enables robots to move beyond simple command execution towards a nuanced understanding of human behavior.

5.2 Industry 5.0

The manufacturing floor is being transformed by gesture recognition. Within the Industry 5.0 paradigm, which emphasizes human-centricity, AI creates collaborative and safe environments. Gesture-based interfaces enable workers to interact with Manufacturing Execution Systems (MES) without breaking focus—requesting materials, reporting issues, or logging task completion via simple hand signals visible to ceiling-mounted cameras, eliminating walk-time to terminals and reducing physical strain. By reducing repetitive movements like typing or scanning, gesture systems lower ergonomic injury risk (Bertolotti et al., 2025). Combined with AR, gesture recognition supports interactive training modules where a virtual assistant guides a worker through assembly tasks, verifying each step via the trainee's gestures (Pilati et al., 2020, as cited in Bertolotti et al., 2025).

5.3 Augmented and Virtual Reality (AR/VR)

AR and VR environments demand natural interaction, as holding a controller breaks the illusion of immersion. Carter (2023) explores how deep learning, particularly CNNs and RNNs, enables free-hand interaction in AR. Users can manipulate virtual objects, navigate menus, or interact with virtual characters using natural gestures, creating a more intuitive and compelling user experience for gaming, design, and training simulations.

5.4 Drone Control and Automotive Interfaces

Gestures offer a hands-free way to control complex machinery. The review on Tello EDU drone techniques (IEEE, 2023) catalogues methods for using hand, face, and eye gestures to pilot a drone—with applications in search and rescue and entertainment. Similarly, ToF cameras inside vehicles can enable drivers to control infotainment systems or answer calls with simple hand movements, reducing distraction (Yasen & Jusoh, 2019).

6. CHALLENGES AND OPEN PROBLEMS

6.1 Occlusion

Hands are dexterous and can easily occlude themselves or be occluded by other objects. Robustly tracking a hand through self-occlusion remains a difficult open problem (Hussain et al., 2024; Yasen & Jusoh, 2019).

6.2 Environmental Variability

Changes in lighting, cluttered backgrounds, and varying skin tones can dramatically affect the performance of vision-based systems. Building systems that generalize across these conditions remains challenging (Yasen & Jusoh, 2019).

6.3 Data Scarcity

Deep learning models are data-hungry. Creating large, well-annotated datasets of gestures that capture the full range of human variability is expensive and time-consuming, often leading to overfitting in practical deployments (Yasen & Jusoh, 2019).

6.4 Computational Cost

State-of-the-art models like Transformers require significant computational resources, making deployment on low-power edge devices (e.g., mobile phones, drones) challenging without architectural optimization.

7. FUTURE RESEARCH DIRECTIONS

7.1 Advanced Multi-Modal Fusion

Building on work like ModDrop, future systems will need to seamlessly fuse vision and audio as well as gaze, physiological signals, and contextual data to achieve a deeper understanding of user intent.

7.2 Few-Shot and Zero-Shot Learning

To overcome data scarcity, research is needed into models that can recognize new gestures from only one or a few examples (few-shot learning), or understand a gesture they have never been trained on by relating it to known concepts (zero-shot learning) (Wang et al., 2019, as cited in Carter, 2023).

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154165

7.3 Self-Supervised Learning

These techniques aim to learn rich representations from unlabeled data. For example, a model could be trained to synchronize video of a gesture with its associated sound, learning powerful features without any manual annotation.

7.4 Efficient On-Device Deployment

Developing lightweight, efficient versions of deep learning models through techniques such as pruning, quantization, and knowledge distillation will be key to deploying advanced gesture recognition on ubiquitous devices.

7.5 Explainable AI (XAI)

As gesture-based systems are deployed in critical applications like surgery or industrial robotics, understanding why a system made a particular decision becomes crucial for trust and safety. XAI techniques can help visualize which parts of the input led to a classification, aiding in debugging and verification.

8. CONCLUSION

The field of gesture tracking has been fundamentally reshaped by artificial intelligence. What was once a niche area of computer vision is now a diverse and mature field, powering applications that range from collaborative robots on the factory floor to immersive gaming experiences in virtual worlds. This review has traced the journey from manual feature extraction with SVMs and HMMs to the automatic, hierarchical feature learning of CNNs, RNNs, and Transformers. We have seen how advanced training strategies like multi-modal fusion and ModDrop are building systems robust enough for real-world uncertainty.

The integration of AI has moved gesture recognition from the laboratory into daily life. In Industry 5.0, it is creating safer, more human-centric workplaces. In robotics, it is enabling more natural collaboration. In AR, it is dissolving the barrier between the user and the digital world. Yet, challenges of occlusion, environmental variability, and data efficiency remain. The future of the field lies in developing AI that is not only more accurate, but also more adaptive, efficient, and explainable. By continuing to draw inspiration from the richness and complexity of human communication, researchers will create gesture-based interfaces that are not just tools, but true extensions of human intent.

REFERENCES

[1]. Yasen, M., & Jusoh, S. (2019). A systematic review on hand gesture recognition techniques, challenges and applications. *PeerJ Computer Science*, 5, e218.

[2]. Hussain, S., Saeed, K., Baimagambetov, A., Rab, S., & Saad, M. (2024). Advancements in gesture recognition techniques and machine learning for enhanced human-robot interaction: A comprehensive review. *arXiv preprint arXiv:2409.06503*.

[3]. Bertolotti, F., et al. (2025). Implementing an AI-driven gesture recognition system in MES for enhanced efficiency and human-centric operations in Industry 5.0. *IFAC-PapersOnLine*.

[4]. Neverova, N., Wolf, C., Taylor, G. W., & Nebout, F. (2015). ModDrop: Adaptive multi-modal gesture recognition. *arXiv preprint arXiv:1501.00102*.

[5]. Carter, E. (2023). Deep learning in human-computer interaction: Improving gesture recognition for augmented reality. *Journal of Artificial Intelligence Research (JAIR)*.

[6]. IEEE. (2023). Artificial intelligence-based human gesture tracking control techniques of Tello EDU quadrotor drone. *2023 IEEE Conference*.

[7]. Sajwan, V., Gandhi, A. B., Chopra, N. K., Praveen, S., Thapliyal, M., Jaleel, U., Pandey, S. D., Kumar, R., & Negi, D. Technological transformations: A comprehensive review of AI, IoT, VR, AR, and emerging technologies in diverse industries.