Hand Gesture Recognition in Human-Computer Interaction: A Comparative Review of Sensor, Vision, and Deep Learning Approaches

Swapnil Dhagdi[1], Om Bande2, Vighnesh Belkar3, Omprakash Bedage4, Prof. J. S. Pawar5

1 Department Of Information Technology, Sinhgad College of Engineering, Pune- 41
2 Department Of Information Technology, Sinhgad College of Engineering, Pune- 41
3 Department Of Information Technology, Sinhgad College of Engineering, Pune- 41
4 Department Of Information Technology, Sinhgad College of Engineering, Pune- 41
5 Department Of Information Technology, Sinhgad College of Engineering, Pune- 41

---***---

Abstract - Hand gesture recognition (HGR) has become a key area in human–computer interaction (HCI), enabling intuitive, contactless communication between humans and machines. This review paper analyzes recent advances in vision-based and deep learning approaches to HGR, focusing on seven representative studies published between 2017 and 2025. Traditional systems using OpenCV and MediaPipe frameworks demonstrate strong real-time performance but face limitations under varying lighting conditions and complex backgrounds. Deep learning methods, particularly convolutional and graph-based neural networks, show significant improvements in accuracy and generalization by modeling spatial–temporal dependencies of skeletal data. The review highlights that attention-based and multi-branch deep learning architectures, such as those proposed by Miah et al. [5], achieve recognition accuracies exceeding 95% on benchmark datasets, outperforming earlier models. Despite this progress, challenges remain in signer-independent learning, dataset diversity, and low-latency implementation for real-world applications. The paper concludes that future research should focus on lightweight hybrid architectures, multimodal sensor fusion, and adaptive learning techniques to improve robustness and scalability in everyday human–computer interaction systems.

Key Words: Hand Gesture Recognition, Human-Computer Interaction (HCI), Computer Vision, Deep Learning, Data Gloves, Skeleton-Based Recognition.

1.INTRODUCTION

Human–computer interaction (HCI) has evolved rapidly over the past decade, aiming to create more intuitive, efficient, and natural ways for humans to communicate with machines. Among various approaches, hand gesture recognition (HGR) has emerged as one of the most promising methods, allowing users to interact through physical gestures rather than traditional input devices such as keyboards, mice, or touch screens. Gestures are a natural form of non-verbal communication that can convey commands, emotions, and intentions. The ability of computers to interpret these gestures accurately bridges the gap between human expressiveness and machine interpretation, contributing to advancements in accessibility, virtual reality, robotics, and assistive technologies.

The concept of HGR integrates multiple disciplines, including computer vision, image processing, pattern recognition, and machine learning. Early systems were largely vision-based, relying on camera input and simple algorithms for segmentation, contour detection, and feature extraction [6], [7]. These methods, implemented using frameworks such as OpenCV or MediaPipe, demonstrated feasibility in real-time scenarios but suffered from several limitations—especially under varying lighting conditions, complex backgrounds, and diverse user characteristics. For instance, Lavanya Vaishnavi et al. [7] developed a MediaPipe-based model capable of recognizing finger counts and hand positions in real time, achieving a recognition rate of about 95%. However, such systems often struggled with environmental variations, occlusions, and signer-dependent recognition.

With the advancement of computational power and deep learning, modern HGR systems have shifted toward neural network-based methods, particularly convolutional neural networks (CNNs) and graph-based deep learning architectures. These models learn features automatically from large datasets, eliminating the need for manual feature engineering. Devineau et al. [4] introduced a CNN model for recognizing gestures from 3D skeletal data, achieving over 91% accuracy on benchmark datasets. Later studies, such as Miah et al. [5], expanded this idea by integrating attention mechanisms and multi-branch architectures to capture spatial–temporal dependencies across joints. These models improved both recognition accuracy and generalization, reaching up to 97% accuracy on challenging datasets like SHREC’17. Such developments mark a major shift from traditional image-based recognition to skeleton-based learning, which focuses on extracting meaningful joint-level information for precise gesture classification.

In parallel, simpler implementations of vision-based systems continue to be relevant for academic and practical applications due to their ease of setup and cost-effectiveness. For example, Kedari et al. [3] proposed a CNN-driven system that enables real-time computer control through webcam-detected gestures, demonstrating the potential of combining classical vision techniques with modern neural networks for user-friendly applications. Similarly, Subramanian et al. [6] discussed the importance of gesture-based HCI as an alternative input method, particularly in assistive and virtual environments. These studies highlight that while high-end deep learning models dominate accuracy benchmarks, vision-based frameworks remain valuable for low-cost, real-world deployments.

Despite these advancements, the field still faces several challenges. The primary concerns include the lack of large, diverse datasets, difficulty in handling signer-independent scenarios, computational demands of deep models, and latency issues in real-time applications. Many systems perform well in controlled laboratory conditions but experience a drop in accuracy when deployed in dynamic, real-world environments. Additionally, model generalization across users, gestures, and lighting variations remains a persistent issue.

The objective of this review is to provide a comprehensive analysis of the current progress in hand gesture recognition, summarizing findings from both traditional and modern approaches. This paper examines seven significant studies published between 2017 and 2025 that collectively illustrate the evolution of HGR—from early vision-based techniques to advanced deep learning architectures. It compares their methodologies, performance, and limitations, while identifying research gaps and suggesting directions for future development.

Ultimately, this review aims to guide students and researchers toward understanding the technical foundations of gesture recognition, its practical challenges, and the innovations shaping its future. By exploring both theoretical perspectives and real-world implementations, the paper emphasizes how gesture recognition continues to move closer to achieving seamless, natural interaction between humans and machines in the modern digital era.

2. DISCUSSION

2.1 Vision-Based Hand Gesture Recognition Systems :

Early research in hand gesture recognition primarily focused on vision-based approaches, where gestures were detected using image or video data captured by standard cameras. These systems relied on image processing, background subtraction, and contour analysis to recognize gestures.

For instance, Subramanian et al. [6] discussed how vision-based recognition enhances human–computer interaction by replacing physical input devices with hand movements. They emphasized that such systems can be applied to a range of domains, including assistive technologies and virtual environments. Similarly, Lavanya Vaishnavi et al. [7] developed a MediaPipe-based model that used Python and OpenCV to identify gestures based on finger positions and counts. Their work achieved an accuracy rate of around 95%, demonstrating real-time performance suitable for basic control applications.

However, these traditional systems faced significant challenges in variable lighting conditions, background complexity, and signer dependency. Most experiments were conducted in controlled laboratory settings, as highlighted by Mohamed et al. [2], who reviewed nearly 100 studies and found that 80% of vision-based experiments used restricted environments to avoid background noise. Although such systems achieved recognition accuracies between 69% and 98%, their lack of robustness limited real-world usability.

To improve generalization, recent studies have incorporated machine learning-based classification and data augmentation methods, but accuracy still heavily depends on the quality of image acquisition and preprocessing.

2.2 Deep Learning Based Recognition Models :

The introduction of deep learning transformed gesture recognition research by automating feature extraction and improving classification accuracy. Instead of relying on handcrafted features, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) began to learn spatial–temporal features directly from data.

Kedari et al. [3] presented a CNN-based model for real-time computer control using webcam images. Their system followed five key stages: image acquisition, hand tracking, feature extraction, gesture recognition, and classification. Compared to earlier approaches, this method provided improved accuracy and flexibility in recognizing different hand poses without explicit background subtraction.

Devineau et al. [4] introduced a pioneering 3D hand gesture recognition model using skeletal joint data rather than RGB images. Their CNN architecture processed temporal sequences of joint positions captured by Intel RealSense cameras and achieved 91.28% accuracy for 14 gestures and 84.35% for 28 gestures. The use of skeletal data allowed faster processing and reduced dependency on color or lighting conditions, which were major limitations in vision-based systems.

These deep learning models demonstrated that hierarchical feature learning can outperform traditional handcrafted methods, making gesture recognition more adaptable and scalable

2.3 Graph Neural Networks and Attention Mechanisms :

Building upon CNN-based architectures, researchers have recently explored graph neural networks (GNNs) and attention mechanisms to better capture the relationships among hand joints in skeletal data.

Miah et al. [5] proposed a multi-branch attention-based graph model that combines spatial–temporal and temporal–spatial attention modules. Their system extracted features through three parallel branches: two graph-based neural network channels and one general deep neural network channel. This design allowed the model to capture both sequential dependencies and local correlations among joints.

When evaluated on benchmark datasets such as MSRA, DHG, and SHREC'17, the model achieved 94.12%, 92%, and 97.01% accuracy, outperforming previous approaches [2], [4]. These results outperformed existing state-of-the-art methods while maintaining low computational cost.

This development highlights an important trend in recent gesture recognition research—combining spatial–temporal modeling with attention mechanisms to improve accuracy, speed, and generalization. Such architectures show strong potential for real-time applications, including virtual reality (VR), robotics, and sign language translation.

3. CRITICAL ANALYSIS

3.1 Dominant Trends and Breakthroughs :

Recent advancements in hand gesture recognition (HGR) reveal a clear shift from traditional vision-based systems toward data-driven deep learning and graph-based architectures. This evolution marks a major turning point in the field.

Earlier systems, as seen in studies like Subramanian et al. [6] and Lavanya Vaishnavi et al. [7], primarily relied on OpenCV or MediaPipe frameworks that processed raw images through background subtraction, contour mapping, and color segmentation. While these approaches demonstrated feasibility, their accuracy and adaptability were limited by environmental conditions and user variability.

The introduction of deep learning, particularly convolutional neural networks (CNNs), established a new baseline for performance. Kedari et al. [3] and Devineau et al. [4] showed that CNN-based systems could automatically extract spatial–temporal features and achieve over 90% accuracy on benchmark datasets. This automation of feature extraction eliminated the need for manual tuning and made gesture recognition more scalable.

A further breakthrough came with the integration of graph neural networks (GNNs) and attention mechanisms, as demonstrated by Miah et al. [5]. These models interpret the human hand as a structured graph of joints and dynamically learn the relationships among them. The introduction of multi-branch attention modules and feature fusion networks has enabled accuracies exceeding 95%, even on complex datasets such as SHREC'17. This represents a key shift from simple image classification to context-aware, skeleton-based learning.

The dominant trends indicate three major breakthroughs:

1. Transition from handcrafted features to deep feature learning through CNNs and GNNs.
2. Adoption of spatial–temporal modeling to capture gesture dynamics more effectively.
3. Emphasis on real-time, lightweight frameworks (e.g., MediaPipe, TensorFlow Lite) for deployability across devices.

3.2 Methodological Weaknesses and Inconsistencies :

1. Dataset Bias and Limited Diversity: Many models, such as those reviewed by Mohamed et al. [2], were trained on datasets collected under controlled laboratory conditions with uniform backgrounds and fixed lighting. As a result, models show high accuracy in experimental setups but degrade significantly when tested in real-world environments.
2. Lack of Standardized Evaluation Metrics: Studies often use different datasets, accuracy measures, and validation protocols, making it difficult to perform fair comparisons across methods. Some works emphasize recognition accuracy, while others report precision, recall, or F1 scores without consistency.
3. Overfitting and Generalization Issues: Deep models, particularly those using limited or repetitive gesture datasets, often overfit to signer-dependent data. This reduces their generalization to unseen users or gestures. Miah et al. [5] attempted to mitigate this using multi-branch attention, but even their model still relies heavily on dataset-specific tuning.
4. Computational Trade-offs: While attention-based and GNN models achieve superior accuracy, they also increase training complexity and require high-end hardware for real-time inference. Few papers explicitly address optimization for low-power or embedded systems.
5. Insufficient Cross-Domain Testing: Most studies validate models on a single dataset (e.g., DHG or SHREC'17), limiting insight into how systems perform across diverse conditions, such as varying camera angles or gesture speeds.

3.3 Unresolved Challenges :

- Signer Independence: Recognition models often rely on user-specific training data. Achieving signer-independent accuracy comparable to signer-dependent setups remains a major challenge.
- Dynamic Gesture Segmentation: Differentiating between continuous gestures and identifying gesture boundaries in real-time sequences is still a complex problem.
- Environmental Robustness: Models frequently fail under variable lighting, cluttered backgrounds, or partial occlusions.

- Latency and Computational Constraints: Deep models require significant computational resources, limiting their use on portable or embedded devices.
- Data Annotation and Availability: The creation of large, labeled gesture datasets is time-consuming and costly, slowing the development of generalized systems.
- Integration with Multimodal Inputs: Few systems combine gestures with other modalities (e.g., voice or gaze) to improve robustness and contextual understanding.

3.4 Research Gaps and Future Opportunities :

1. Multimodal Fusion Models: Integrating visual, depth, and skeletal data — possibly along with audio or haptic input — could lead to more robust and context-aware gesture recognition systems.
2. Lightweight and Edge-Compatible Models: Future research should focus on developing compact models optimized for real-time execution on low-power devices. Techniques like model pruning, quantization, and knowledge distillation can help balance accuracy and speed.
3. Self-Supervised and Transfer Learning: Training models with minimal labeled data using self-supervised learning can overcome dataset scarcity and improve generalization to new users or environments.
4. Standardized Benchmarks: Establishing shared datasets and unified evaluation protocols would enable fair comparison and accelerate progress in the field.
5. Human-Centered Design and Usability Studies: Most existing research focuses on technical accuracy. Future work should explore user experience, ergonomics, and interaction design, ensuring that gesture systems align with natural human behavior.
6. Cross-Domain Generalization: Developing adaptive models capable of transferring knowledge between different gesture domains (e.g., sign language, robotics, gaming) could make HGR systems more flexible and universally applicable.

3. CONCLUSIONS

Hand gesture recognition (HGR) has become a crucial component of modern human–computer interaction, enabling intuitive, contactless control across various domains such as robotics, virtual reality, and assistive technology. This review examined seven influential studies published between 2017 and 2025, revealing how the field has evolved from simple vision-based models to advanced deep learning and graph-based architectures. Early image-processing systems, such as those developed using OpenCV and MediaPipe [6], [7], achieved recognition accuracies ranging between 90% and 95% under controlled conditions but struggled with environmental variability and user dependency. The introduction of deep learning, particularly convolutional neural networks (CNNs), improved recognition accuracy to over 91% on benchmark datasets by learning spatial–temporal features directly from data [3], [4]. The most recent breakthroughs, involving graph neural networks (GNNs) and attention-based architectures, have reached accuracies up to 97% [5], demonstrating strong performance and reduced computational cost compared to traditional methods.

Despite these advances, key challenges remain unresolved. Most models continue to rely on datasets collected in restricted laboratory environments, limiting their generalization to real-world scenarios. Achieving signer-independent accuracy above 80%, reducing latency for real-time applications, and maintaining robustness under varying lighting and backgrounds remain active areas of improvement.

REFERENCES

[1] Aniket Abhishek Soni, “Improving Speech Recognition Accuracy Using Custom Language Models with the Vosk Toolkit,” *Southern Arkansas University*, 2025.

[2] Noraini Mohamed, Mumtaz Begum Mustafa, and Nazean Jomhari, “A Review of the Hand Gesture Recognition System: Current Progress and Future Directions,” *IEEE Access*, vol. 9, pp. 157422–157434, 2021.

[3] Pradnya Kedari, Shubhangi Kadam, and Rajesh Prasad, “Controlling the Computer using Hand Gestures,” *Multimedia Research*, vol. 5, no. 3, pp. 9–16, 2022. DOI: 10.46253/j.mr.v5i3.a2.

[4] Guillaume Devineau, Wang Xi, Fabien Moutarde, and Jie Yang, “Deep Learning for Hand Gesture Recognition on Skeletal Data,” in *Proc. 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)*, Xi’an, China, pp. 1–8, 2018. DOI: 10.1109/FG.2018.00025.

[5] Abu Saleh Musa Miah, Md. Al Mehedi Hasan, and Jungpil Shin, “Dynamic Hand Gesture Recognition Using Multi-Branch Attention Based Graph and General Deep Learning Model,” *IEEE Access*, vol. 11, pp. 4703–4717, 2023. DOI: 10.1109/ACCESS.2023.3235368.

[6] Archanasri Subramanian, Nivedhitha Asokkumar, and Jyothi Nayak, “Hand Gesture Recognition for Human Computer Interaction,” *Procedia Computer Science*, vol. 115, pp. 367–374, 2017. DOI: 10.1016/j.procs.2017.09.092.

[7] Lavanya Vaishnavi D. A., Anil Kumar C., Harish S., and Divya M. L., “MediaPipe to Recognise the Hand Gestures,” *WSEAS Transactions on Signal Processing*, vol. 18, pp. 134–141, 2022. DOI: 10.37394/232014.2022.18.19.