

Real-Time Hand Gesture Recognition in Human-Computer Interaction: A Comprehensive Review

Shivani Gupta, Satinder Bal Gupta*

Department of Computer Science and Engineering,
Indira Gandhi University, Meerpur, Haryana, India
shivani.cse@igu.ac.in, satinderbal@igu.ac.in

Abstract

Real-Time Hand Gesture Recognition (HGR) has emerged as a key model of modern Human Computer Interaction (HCI) technologies, enabling touchless and intuitive interactive experiences for a wide range of HCI applications, from assistive technology, virtual and augmented reality to industrial production lines. Although AI and DL open up important opportunities, there are practical challenges to developing and deploying real-time HGR systems that are strong to issues such as computational complexity, variations in the environment, and latency. This survey reviews the state-of-the-art in real-time HGR, including recent work in and algorithmic approaches for various sensor modalities and different stages of the processing pipeline, including detection, tracking, and spatial-temporal classification, using monocular RGB cameras and multi-modal depth sensing systems. In addition, we cover applications such as sign language recognition, healthcare hygiene systems, automotive user interfaces, and human-robot collaboration (HRC), and challenges such as occlusion invariance, light variability, performance versus portability trade-offs, constraints imposed by edge-device deployment, and potential research avenues for multimodal artificial intelligence systems of the future.

Keywords: hand gesture recognition, computer vision, human-computer interaction, deep learning, real-time processing, sensor fusion

ACM Computing Classification System 2012: Human-centered computing → Human computer interaction (HCI) → Interaction techniques → Gestural input, Computing methodologies → Artificial intelligence → Computer vision → Computer vision tasks → Activity recognition and understanding

Mathematics Subject Classification 2020: 68-02, 68T45, 68T07

Received: September 3, 2026, *Accepted:* October 19, 2026, *Published:* October 26, 2026

*Corresponding author

Citation: Shivani Gupta, Satinder Bal Gupta, Real-Time Hand Gesture Recognition in Human-Computer Interaction: A Comprehensive Review, Serdica Journal of Computing 19(2), 2025, pp. 73-90, <https://doi.org/10.55630/sjc.2025.19.73-90>

1 Introduction

Hand gestures are one of the most informative but underutilized methods of communication. The way of interaction with computers is changing rapidly from traditional types of interaction, such as keyboard or light touch, to new natural touchless interactions based on real-time recognition of hand movements [1].

The benefits of real-time hand-gesture recognition range from improving the commercial value of consumer electronics through more intuitive control of smartphones and TVs to solving more serious scientific challenges, like developing better ways of communication for people with severe motor disabilities and improving the precision of robots in factory automation and healthcare. The use of touchless control can prevent the spread of infections in hospitals because it allows avoiding contamination of medical equipment with pathogens [2].

Real-time recognition software can understand complex sequences of hand movements, thus allowing people to express more information and making such programs more valuable for people with disabilities. Using intuitive hand gestures to control robots can make factory automation and healthcare safer. Technologically, an end-to-end real-time hand gesture recognition (HGR) system is comprised of a multi-stage visual processing pipeline. It starts with detection and segmentation, where hand ROI is extracted from the scene using skin-color thresholding, background subtraction, or keypoint landmark detection. It is followed by the tracking stage, which estimates spatio-temporal motion patterns of the detected hand(s) across subsequent video frames [3–5].

Lastly, the recognition stage classifies the resulting spatio-temporal features into a set of predefined command classes with convolutional and/or recurrent neural networks [6–8].

Early approaches to real-time gesture recognition were limited to a small set of holistic hand poses such as thumb-up, stop, etc. Meanwhile, state-of-the-art real-time applications now target towards dynamic hand pose estimation, which involves spatio-temporal analysis of fine-grained finger transformations across scale spaces [9–12].

The problem of tracking hand movements introduces a twofold challenge compared to traditional image classification pipelines because of dynamic $2D + 1D$ input data. As a result, a majority of real-time HGR applications are characterized by the need for high classification accuracy and low computational overhead.

This survey aims to provide a complete overview of recent advances in real-time hand gesture recognition. Specifically, we discuss the key design choices behind existing works including different sensing paradigms, visual features, and spatial-temporal processing units. We analyze the latest solutions in the context

of monocular RGB, Time-of-Flight, and wearable depth sensors, examine neural networks architectures suitable for spatial-temporal modeling and review the trade-offs in vision-based human-computer interaction systems for the domains of smart healthcare, augmented reality, intelligent gaming, and automotive.

1.1 Key contributions of this survey

Unlike existing surveys this paper provides a critical taxonomy-driven synthesis of real-time HGR architectures. The distinct contributions of this review are fourfold:

1. *Structured Algorithmic & Sensing Taxonomy*: Our multi-dimensional classification scheme classifies HGR pipelines by feature representation space (in terms of pixels versus keypoint skeletons), sensor physics and modeling dynamics.
2. *Theoretical Efficiency Modeling*: This component formulates a unified mathematical cost function in order to achieve an equilibrium between recognition precision, hardware inference speed, and energy consumption for edge computing environments.
3. *Analysis of the root cause of mode of failure*: On the basis of realistic deployment constraints like effect of self-blocking, inadequate lighting, and fluctuations in distribution shift, a systematic classification of computational and physical failure mechanisms present in modern models.
4. *Standardized Benchmarking Matrix*: In this study, a precise cross-domain comparison is presented that explains the latency-accuracy trade-offs observed in about 32 peer-reviewed deployments.

2 Literature search methodology (PRISMA framework)

For the purposes of this study, it is sufficient to report that the literature search process followed the PRISMA 2020 framework. The search was performed in the following databases: IEEE Xplore, Google Scholar, Web of Science, ScienceDirect (Elsevier) as shown in Fig. 1. The query included the combination of keywords and phrases: (*“Real-Time Hand Gesture Recognition” OR “HGR”*) AND (*“Deep Learning” OR “Computer Vision”*) AND (*“Human-Computer Interaction” OR “HCI”*) AND (*“Sensors” OR “ToF” OR “MediaPipe” OR “sEMG”*).

The first database had 412 studies. After removing duplicates 285 records were looked at by title and abstract to find studies that fit the inclusion and exclusion rules.

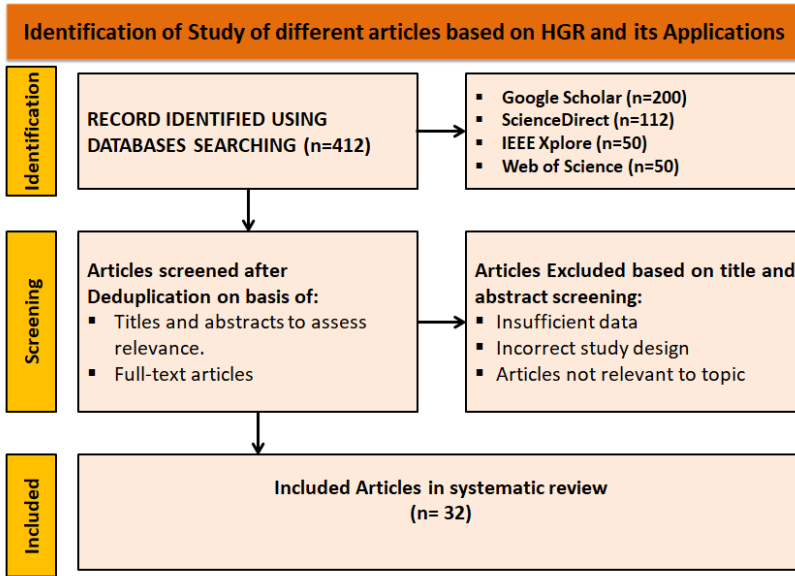


Figure 1: PRISMA flow diagram of the identification, screening and inclusion of studies on HGR and its applications.

Inclusion criteria:

- Peer-reviewed journals/conference papers;
- Emphasis on creating a real-time hand gesture recognition and tracking pipeline;
- Reporting numerical results: accuracy, computational delay, frames per second (FPS), or power consumption;
- Publication, on algorithms or sensing methods

Exclusion criteria:

- Non-English manuscripts;
- Recognition of static hand poses only;
- Lack of relevant description of the methodology;
- Non-peer-reviewed abstracts/papers.
- Insufficient data

The 94 full-text manuscripts were assessed according to the inclusion/exclusion criteria, and 32 papers were selected for qualitative synthesis and quantitative comparison in this review article.

3 Analytical schema & sensing modalities

The progress of real-time Hand Gesture Recognition (HGR) is contingent upon parallel advances in sensor technologies and underlying deep learning models. Recent research features a wide range of both low-level vision-based approaches to high-level spatial-temporal architectures trained on multiple sensory inputs. Monocular RGB cameras are widely used because they are common, in electronics, video surveillance and many embedded systems. In the days of real-time hand gesture recognition most approaches relied on traditional handcrafted features to keep things fast and efficient. Sharma et al. explored the utilization of ORB (Oriented FAST and Rotated BRIEF) descriptor alongside K-means clustering algorithm for establishing a BoW (Bag-of-Words) framework. The integration of ORB with Canny edge detection proved to be effective in minimizing feature vector dimensionality in comparison to PCA and HOG descriptors, and thereby accelerating the process of training and inference. Conventional feature descriptors faced difficulties while dealing with background clutter or fluctuations in lighting. To mitigate this limitation while preserving the feasibility of processing videos in real-time, recent developments have adopted neural network learning-based models for keypoint localization.

Google's MediaPipe framework represent one such prominent example that helps in accomplishing real-time recognition of hand skeleton gestures. In this method, each frame is converted into a vector comprising standardized coordinate values [13]. The MediaPipe pipeline suppresses superfluous data such as skin color variations, background elements and lighting inconsistencies. After completing the process of extracting keypoints, Inception-V3 network is being applied to regions of interest derived from MediaPipe and to the unprocessed RGB pixel values obtained from video. The resulting feature vectors are then fed into the Long Short-Term Memory network which facilitates the model to comprehend the information flow as the time progress.

The authors [14] have also developed a Deep Neural Network to identify emergency gestures in real-time. Their strategy examines individual video frames through a specialized CNN and delivers real-time performance suitable for automated emergency alerts. While monocular RGB systems are easy to use, they have several disadvantages, including self-occlusion, strong background variation, and the absence of metric information about scene depth. To overcome these shortcomings, several researchers have begun to adopt active depth sensors, including ToF (Time of Flight) cameras and structured-light sensors, as well as other wearable movement sensors.

The authors [15] did a study comparing how monocular RGB cameras and

RGB-D cameras work for tasks where humans and robots interact. They found that using a mix of dynamic gesture recognition that includes depth information is much better at handling big changes, in the background and when there is not much light. In automotive and industrial applications where lighting conditions are hard to predict, it may be necessary to adopt special depth-sensing hardware.

The authors [16] proposed a novel driver gesture recognition system based on ToF depth cameras with model adaptation. The approach demonstrated performance in personalized dynamic gesture recognition with active infrared lighting. Active infrared lighting is critical for hand tracking in both daylight and nighttime conditions. I feel that optical cameras are not the option, for building touchless interfaces. Some researchers have proposed vision-based input devices that require less power to operate. Jung and Lee developed a novel IoT-based wearable gesture recognition system that employs a multi-orientation flex sensor and a lightweight multi-layer perceptron (MLP) algorithm. They managed to implement their approach on a low-power microcontroller unit (MCU) and achieved tremendous results in terms of power consumption (<2.74 mJ per gesture).

Another method for building a touchless interface involves using surface electromyography (sEMG) signals. Kadavath et al. demonstrated that it is possible to classify motor movements based on ML algorithms, which can be an alternative when a direct view of the subject is unavailable. A common problem for all dynamic gesture recognition systems is the need to process both spatial and temporal components. While early static gesture recognition approaches were unable to capture the temporal characteristics of motion, deep learning methods have been proposed that can learn the joint spatial-time features using CNNs, RNNs, or attention-based architectures. They used a combination of CNN and LSTM networks for processing driving commands in the form of dynamic hand gestures.

The authors [17] proposed a different approach based on 3D CNNs and LSTMs to recognize human motions for real-time human-robot collaboration in industrial manufacturing.

The authors [18] proposed to use recurrent networks with feature fusion for dynamic hand gesture recognition in home environments. They showed that their approach is invariant to a large extent to variations in the background and achieves excellent results on standard benchmarks, including HaGRI, EgoGesture, and Jester. Finally, since conventional RNNs are not well-suited for parallelization, some researchers have also started exploring the effectiveness of attention-based encoder-decoder approaches for dynamic gesture recognition.

The authors [19] compared four different techniques, including RNNs, CNNs, CNN+RNNs, and Transformer Encoder, for solving the dynamic gesture classi-

fication problem using a landmark-based representation. They found that while transformer networks have an advantage over conventional RNNs by modeling the global context, optimized CNN+LSTM approaches still perform best in terms of latency-speed trade-offs. In order to summarize the state-of-the-art methods for dynamic gesture recognition, we present a comprehensive overview of the key trade-offs between different architectures, as well as computational complexity and hardware requirements reported in recent papers, in Table 1.

3.1 Unified HGR architectural taxonomy

For the systematic analysis of real-time HGR pipelines, the methods are classified into three layers:

- *Layer 1: Perception & Sensing Modality:* Optical Passive (Monocular RGB, Event-based), Optical Active (ToF, Structured Light), Non-optical Contact (sEMG, Microcontroller Flex Sensors), etc.
- *Layer 2: Structural Spatial Representation:* Raw RGB/Depth (Dense Pixel Arrays), 2D/3D skeleton keypoints (via MediaPipe), and Bio-electrical Signal Vectors.
- *Layer 3: Temporal Alignment & Feature Modeling:* Recurrent Aggregation (LSTM/GRU), 3D-CNN, and Attention-Driven Self-Attention Mechanisms (Vision Transformers)

4 Domain-specific applications and technical performance analysis

The adoption of real-time hand gesture recognition (HGR) is increasing in various applications. The requirements of execution time, target signs, computational cost, and noise resistance in different application domains of HGR vary significantly. For human-computer interaction, especially sign language recognition (SLR), the recognition rate of static and dynamic signs is critical to fulfill the demand of effectively communicating with the disabled. The vision-based method has proved to be effective without relying on cumbersome devices.

The authors [20] presented a convolutional neural network (CNN) with four convolutional layers and sixteen filters, using the optimizer Adam for recognizing Indian sign language (ISL). The proposed method was trained on 3500 gesture images collected from disabled people for fifteen epochs, which achieved 99.76% accuracy on ISL signs. Similarly, the authors [21] utilized SURF hand key points extracted by the bag-of-features approach and support vector machines (SVM)

Modality/ Architecture	Methodological Strengths	Operational & Latency Constraints	Primary Application Domains
Monocular RGB + Keypoint Networks (<i>e.g., MediaPipe + DNN</i>)	Ultra-low computational overhead; universal hardware compatibility with standard webcams.	Highly vulnerable to severe low-light conditions, background clutter, and hand self-occlusion.	Assistive SLR, Mobile HCI, Emergency Alerts
Active Depth / ToF Sensing (<i>RGB-D / Time-of-Flight</i>)	Structural immunity to ambient lighting variations; precise 3D metric scale and spatial tracking.	Higher hardware cost; increased power consumption; reduced operational sensing range.	Automotive UI, Surgical Environments, Cobots
Spatial-Temporal Hybrids (<i>3D-CNN, CNN-LSTM</i>)	Superior feature extraction accuracy across dynamic, multi-frame continuous gesture streams.	Large parameter size; high memory footprint; potential latency bottlenecks on edge processors.	Industrial Automation, Dynamic Gesture Control
Wearable & Sensor Fusion (<i>sEMG, Flex Sensors</i>)	Complete immunity to optical line-of-sight occlusions; ultra-low execution latency.	Requires physical skin/body contact; subject to sensor displacement and calibration drift.	Prosthetics, Low-power Wearable IoT

Table 1: Comparative Analysis in Real-Time HGR.

Architecture Family	Spatial Complexity	Latency (T_{inf})	Power (P)	Robustness to Lighting	Occlusion Handling	Optimal Deployment Target
Keypoint + Light MLP	$O(K)$	<10 ms	Very Low (<0.5 W)	Moderate	Poor	Mobile & Embedded Microcontrollers
2D-CNN + LSTM	$O(H \cdot W \cdot C)$	15–30 ms	Medium (2–5 W)	Low (RGB) / High (Depth)	Moderate	Automotive Cockpits, Smart Appliances
3D Volumetric CNN	$O(T \cdot H \cdot W \cdot C)$	>35 ms	High (>10 W)	Moderate	High	High-performance Industrial Cobots
Vision Transformers (ViT)	$O(N^2)$	>45 ms	Very High (>15 W)	High	High	Server-side / Cloud-assisted HGR
sEMG / Flex Sensors	$O(S)$	<5 ms	Ultra-low (<0.1 W)	Immune	Complete Line-of-Sight Immunity	Wearable Prosthetics & Smart Rings

Table 2: Multi-Criteria Architectural Trade-off Decision Matrix.

and KNN classifiers for recognizing American sign language (ASL) in surveillance videos. By comparing with traditional handcrafted feature extraction methods, the proposed model attained an average test accuracy of 84.6% (89% evaluation accuracy) on 26 characters of the alphabet using standard mobile phone cameras with different backgrounds. Moreover, a CNN-based vision system was introduced by the authors [22] which can effectively identify complex hand signs with high accuracy. This method achieved 99.23% and 99.00% accuracy on the two benchmarks for sign language gesture recognition tasks, respectively.

While HGR technology in cars and trucks and in factories is designed to manage robot servos and entertainment systems the situations where it is used change from environments to busy work areas, where the dependability of control has a direct impact, on road safety and worker safety. For in-car gesture recognition, the authors in [16] presented an adaptive CNN long short-term memory (CNNLSTM) network based on Time-of-Flight (ToF) information and personalized adaptation. Using infrared complementary metal-oxide-semiconductor (CMOS) depth sensors this method reached ninety percent accuracy while cutting the amount of personal adaptation data that needs to be collected. In human-robot collaboration accurate hand gesture recognition is essential so that robots can understand and reply to gestures promptly.

The authors [23] introduced a robotic arm control system based on OpenCV preprocessing and CNN classification networks, which directly converted the recognized hand gestures from a camera into robot joints' motion commands and attained 96.90%–99.20% accuracy on different datasets.

The authors [17] proposed a collaborative robot vision platform using 3D convolution and LSTM networks to capture spatial and temporal features from gray-scale video streams, which can classify ten different industrial gestures with 90.00% accuracy.

For home electrical appliance control and health-care robots, low-power embedded systems dominate the architecture, where power consumption, noise resistance, and algorithm delay should be optimized to meet the limited power supply and complicated indoor backgrounds.

The authors [18] introduced a gesture-based home appliance control system based on recurrent neural networks (RNNs) and fusion architectures, which can effectively recognize hand gestures from home videos with dynamic backgrounds. The method attained 92.57% accuracy on the HaGRI dataset, 91.86% accuracy on the EgoGesture dataset, 91.57% accuracy on the Jester dataset, and 90.43% accuracy on the WLASL dataset. Alternatively, non-optical IoT wearable devices can serve as a practical solution in case the line of sight is obstructed or the

camera deployment is restricted due to privacy constraints.

The authors [24] have developed a system for recognizing finger gestures incorporating a dual-axis flex sensor and a microcontroller unit. A multi layer perceptron architecture with less parameters was used for facilitating recognition of energy-efficient wearable gestures. This approach achieved 95.5% accuracy and consumed only 2.74 mJ per inference. In Table 3, a summary has been provided of the salient characteristics of the methods presented above.

5 Critical discussion, open challenges, and future directions

Although the reported accuracy on benchmark datasets is indeed impressive, there is a fundamental challenge in deploying such models in realistic applications that goes beyond the controlled conditions of a laboratory setting [1–3, 10, 15, 25]:

5.1 The accuracy-speed trade-off on edge devices

For on-device real-time performance, the required FPS is usually above 30 frames per second or, equivalently, an end-to-end latency under 33 ms. 3D CNN or Vision Transformers, the computationally intensive spatial-temporal models, demonstrate high performance in recognition of continuous gestures. However, their extensive parametric complexity makes them impractical for deployment in memory limited embedded environments like vehicular ECUs, micro-controller based IoT applications. Meanwhile, keypoint-based approaches (e.g. MediaPipe + small MLP) significantly reduce the overhead but fail in complex scenarios requiring fine-grained articulation or temporal finger crossings [26–29].

5.1.1 Mathematical formulation of the Edge-HGR deployment constraint

Deploying HGR in time on hardware, with limited resources can be described as an optimization problem that follows certain limits. We define the Real-Time Utility Metric (μ) as:

$$\mu(M, H) = \frac{\text{Accuracy}(M)}{\alpha \cdot \text{Latency}(M, H) + \beta \cdot \text{Energy}(M, H) + \Upsilon \cdot \text{Memory}(M)}$$

where:

- M denotes the neural model architecture, and H denotes the target hardware platform.

Application Domain	Primary Algorithmic Technique	Input Modality & Hardware Platform	Dataset & Benchmark Scope	Classification Accuracy (%)	Real-Time Execution & Operational Parameters
Assistive SLR (<i>Static ISL</i>) [20]	16-Filters and 4-Layer 2D-CNN with Adam Optimizer	RGB Camera / Standard Desktop CPU	3,500 images across 10 static ISL signs	99.76%	High-precision static posture mapping; batch-trained over 15 epochs.
Assistive SLR (<i>ASL Alphabets</i>) [21]	SURF (Speed Up Robust Features) Feature Extraction + BoF + SVM/KNN	Mobile RGB Camera / MATLAB Desktop	260 self-collected images across 26 ASL alphabet signs	84.60% (<i>Test set</i>) / 89% (<i>Evaluation set</i>)	Color space thresholding (HSV-Hue Saturation-Value); optimized for low-resource smartphone images.
Assistive SLR (<i>Continuous Signs</i>) [22]	Deep Hybrid CNN (HVCNNM)	Monocular RGB Camera	Dual benchmark datasets (Sign language signs) Massey University Dataset (MUD) and ASL alphabet dataset	99.23% (MUD) / 99% (ASLAD)	Touchless vision-based interface; eliminates wearable hardware dependencies.
Automotive Control [16]	CNN-LSTM with Personalized Adaptation	Time-of-Flight (ToF) Infrared Camera	In-vehicle dynamic driver gesture video streams	90%	Real-time cabin tracking (<30 ms); robust to dynamic day/night illumination changes.

Table 3: Performance Comparison of State-of-the-Art Real-Time HGR Deployments Across Domains.

Application Domain	Primary Algorithmic Technique	Input Modality & Hardware Platform	Dataset & Benchmark Scope	Classification Accuracy (%)	Real-Time Execution & Operational Parameters
6-DoF Robotic Arm [23]	OpenCV Image Processing + 2D-CNN Classifier Model	Optical Vision Sensors / Robotic Controller	Multi-class hand gesture dataset for 6-DoF (Degrees of Freedom) arm manipulation	96.90% – 99.20%	Directly translates hand poses into continuous 6-DoF robotic manipulator commands.
Industrial Cobots [17]	3D-CNN + Spatial-Temporal LSTM LSTM layers in combination with 3D convolution were proposed to extract the spatio-temporal details.	Standard Computer Webcam / Grayscale Video Pipeline	Custom industrial gesture dataset (10 unique gestures)	90%	Processes 3D spatial-temporal frame volumes; filters background motion noise.
Smart Appliances [18]	Feature Fusion (SSMD/Active Bee) + RNN (Recurrent Neural Network)	Monocular Optical Camera	92.57% for HaGRI dataset, 91.86% for Egogesture, 91.57% for Jester and 90.43% for WLASL.	90.43% – 92.57%	Multi-dataset evaluated; robust against complex household background clutter.
Wearable IoT [24]	Lightweight Multi-Layer Perceptron (MLP)	Dual-Orientation Flex Sensors / Low-Power MCU (Micro Controller Unit)	Microcontroller finger bending data streams	95.50%	Ultra-low energy footprint (<2.74 mJ per gesture); immune to optical occlusion.

Table 3: Performance Comparison of State-of-the-Art Real-Time HGR Deployments Across Domains (continued).

- Latency(M, H) is the end-to-end processing delay per frame (ms).
- α, β, Υ are domain-specific weighting parameters ($\alpha + \beta + \Upsilon = 1$).

For real-time interaction the hard temporal constraint is defined by

$$\text{Latency}(M, H) \leq \Delta T_{\max},$$

where $\Delta T_{\max} = 33.3$ ms (to sustain ≥ 30 FPS).

5.2 Occlusion, motion blur, and environmental variability

The systems that perform computer vision using RGB cameras faces numerous problems due to changes in environment.

- Subject occlusion and self-shadowing: While hand gestures are being performed by an individual, the view of the palm region gets obstructed by the movement of fingers which complicates the process of tracking keypoints.
- Lighting condition variations: While capturing dynamic gestures, motion blur is a common problem whereas skin segmentation efficacy gets diminished by extreme low-light environments. Although Time-of-Flight and multimodal sensing techniques help in mitigating these issues, but the cost of energy consumption and the complexity of hardware also increases [30–32].

5.2.1 Root-Cause Failure Mode Analysis (RCFA)

Table 4 illustrates the correlation between operational failure situations and the bottlenecks responsible for their occurrence, and recommends architectural adjustments that could help in mitigating these challenges.

5.3 Dataset bias and generalization gaps

Latest machine learning models often show suboptimal performance if testing is performed on data that differs from data used at the time of training. This issue arises due to several factors such as different skin color, complex background setting, variations in angles of camera, and inconsistent movements in hand gestures. These changes in data distribution acts as a limitation in performance of models and the performance may degrade in real-time settings. As a result, the applicability of these models is weakened.

Vulnerability Factor	Physical / Algorithmic Cause	Performance Impact	Novel Technical Mitigation
Dynamic Illumination	RGB sensor saturation / underexposure	Keypoint tracking loss (>40% jitter)	Active ToF infrared sensor fusion; exposure-invariant color space conversions (HSV/YUV).
Self-Occlusion	Inter-finger overlaps in 2D perspective space	Misclassification of fine-grained signs	Graph Convolutional Networks (GCNs) with kinematic anatomical constraint modeling.
High Velocity Motion	Camera shutter exposure latency causing motion blur	Temporal feature misalignment in CNNs	Integration of Neuromorphic Event Cameras (ASIC frame-free sensing).
Cross-Subject Drift	Variations in hand anatomy and execution speed	Up to 15–25% accuracy drop	On-device Meta-Learning and Test-Time Adaptation (TTA) algorithms.

Table 4: Systematic Failure Mode & Mitigation Analysis in Real-Time HGR.

5.4 Future directions

In order to achieve long-term effectiveness, the systems need to focus on the following areas:

- Using techniques such as post-training compression and network pruning for minimizing the complexity of models and obtaining high speed and improved efficiency on edge devices.
- Using self-supervised learning techniques specifically designed for video data and incorporating sensor inputs such as radar/mm wave technology, depth sensors, and RGB cameras for getting high performance.
- Implementing ongoing adaptation and tailored approaches to fine-tune models for the users so that it can be used efficiently over time.

6 Conclusion & strategic roadmap

This comprehensive review presents different hand gesture recognition systems in a structured manner that covers hardware setups, latest deep learning models,

and different applications. About 32 different implementations have been studied by the authors and it has been determined that the accuracy obtained by the models in recognition of gestures exceed 95% when applied in controlled laboratory settings, but in practical situations in real world achieving such high accuracy poses a lot of challenges. There is need to enhance the processing speed, reduce consumption of energy, and improve the robustness of the system in varying environments so as to achieve better performance.

This gap between research and real-world usage can be addressed by focusing on the following areas:

- *Neuromorphic and Event-Based Sensing*: Transitioning from standard RGB cameras to event-driven sensors substantially mitigates motion blur and significantly diminishes computational latency.
- *Physics-Informed Graph Neural Networks*: Embedding biochemical constraints into skeletal keypoint recognition models improves accuracy by minimizing errors that arise from mutual occlusion between fingers and hands.
- *Quantization-Aware Hardware Co-Design*: The real-time performance of the recognition systems can be enhanced by integrating architecture search techniques and edge-based accelerators like TPU and NPU, which is measured using real-time utility metric (μ).

References

- [1] X. Han, F. Lu, G. Tian, Efficient 3D CNNs with knowledge transfer for sign language recognition, *Multimedia Tools and Applications*, 81:10071–10090, 2022.
- [2] A. Sharma, A. Mittal, S. Singh, V. Awatramani, Hand Gesture Recognition using Image Processing and Feature Extraction Techniques, *Procedia Computer Science*, 173:181–190, 2020.
- [3] J. Shin, A. S. M. Miah, M. H. Kabir, M. A. Rahim, A. A. Shiam, A Methodological and Structural Review of Hand Gesture Recognition Across Diverse Data Modalities, *IEEE Access*, 12:142606–142639, 2024.
- [4] J. B. Bavelas, The social life of hand gestures, *Face-to-Face Dialogue: Theory, Research, and Applications*, Oxford University Press, New York, 2022.
- [5] Shivani, S. B. Gupta, A Comprehensive Analysis of Recognition of Hand Gestures using Machine Learning, *Makara Journal of Technology*, 29:48–60, 2025.
- [6] A. Dhande, S. Mantri, H. M. Pande, Comparative Analysis of Human Hand Gesture Recognition in Real-Time Healthcare Applications, *Expert Clouds and Applications: Proceedings of ICOECA 2023*, Lecture Notes in Networks and Systems, 673:461–475, Springer, Singapore, 2023.

- [7] D. Han, B. Mulyana, V. Stankovic, S. Cheng, A Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation, *Sensors*, 23:3762, 2023.
- [8] J. Wu, P. Ren, B. Song, R. Zhang, C. Zhao, X. Zhang, Data glove-based gesture recognition using CNN-BiLSTM model with attention mechanism, *PLoS ONE*, 18:e0294174, 2023.
- [9] J. Wang, I. Ivrisimtzis, Z. Li, L. Shi, Hand gesture recognition for user-defined textual inputs and gestures, *Universal Access in the Information Society*, 24:1315–1329, 2025.
- [10] M. Oudah, A. Al-Naji, J. Chahl, Hand Gesture Recognition Based on Computer Vision: A Review of Techniques, *Journal of Imaging*, 6:73, 2020.
- [11] M. R. K. Kadavath, M. Nasor, A. Imran, Enhanced Hand Gesture Recognition with Surface Electromyogram and Machine Learning, *Sensors*, 24:5231, 2024.
- [12] J. Wang, Application of Artificial Intelligence in Hand Gesture Recognition with Virtual Reality: Survey and Analysis of Hand Gesture Hardware Selection, *arXiv preprint 2405.16264*, 2024.
- [13] Yaseen, O.-J. Kwon, J. Kim, S. Jamil, J. Lee, F. Ullah, Next-Gen Dynamic Hand Gesture Recognition: MediaPipe, Inception-v3 and LSTM-Based Enhanced Deep Learning Model, *Electronics*, 13:3233, 2024.
- [14] N. F. Thejowahyono, M. V. Setiawan, S. B. Handoyo, A. H. Rangkuti, Hand Gesture Recognition as Signal for Help using Deep Neural Network, *International Journal of Emerging Technology and Advanced Engineering*, 12:37–47, 2022.
- [15] J. Qi, L. Ma, Z. Cui, Y. Yu, Computer vision-based hand gesture recognition for human-robot interaction: a review, *Complex & Intelligent Systems*, 10:1581–1606, 2024.
- [16] A. Gomaa, G. Reyes, M. Feld, It’s all about you: Personalized in-Vehicle Gesture Recognition with a Time-of-Flight Camera, *Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '23)*, Ingolstadt, Germany, pp. 234–243, ACM, 2023.
- [17] W. Dumoulin, N. Thiry, R. Slama, Real Time Hand Gesture Recognition in Industry, *Proceedings of the 2021 3rd International Conference on Video, Signal and Image Processing (VSIP '21)*, Wuhan, China, pp. 1–7, ACM, 2022.
- [18] B. I. Alabdullah, H. Ansar, N. Al Mudawi, A. Alazeb, A. Alshahrani, S. S. Alotaibi, A. Jalal, Smart Home Automation-Based Hand Gesture Recognition Using Feature Fusion and Recurrent Neural Network, *Sensors*, 23:7523, 2023.
- [19] J. Zhao, X. H. Li, J. C. Dela Cruz, M. S. Verdadero, J. C. Centeno, J. M. Noveler, Hand Gesture Recognition Based on Deep Learning, *2023 International Conference on Digital Applications, Transformation & Economy (ICDATE)*, Miri, Malaysia, pp. 250–254, IEEE, 2023.

- [20] S. Vanaja, R. Preetha, S. Sudha, Hand Gesture Recognition for Deaf and Dumb Using CNN Technique, *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, pp. 1–4, IEEE, 2021.
- [21] R. S. T. Kamal Eldeen, E. H. G. Yousif, A Hand Gesture Recognition System for Deaf-Mute Individuals, *Global Journal of Medical Research: K Interdisciplinary*, 21:21–30, 2021.
- [22] K. Aurangzeb, K. Javeed, M. Alhussein, I. Rida, S. I. Haider, A. Parashar, Deep Learning Approach for Hand Gesture Recognition: Applications in Deaf Communication and Healthcare, *Computers, Materials & Continua*, 78:127–144, 2024.
- [23] P. J. Ezigbo, O. C. Nosiri, E. S. Mbonu, V. Ofor, J. Obichere, Hand Gesture Recognition and Control for Human-Robot Interaction Using Deep Learning, *International Journal of Electrical and Electronic Engineering & Telecommunications*, 12:433–441, 2023.
- [24] W. Jung, H. G. Lee, Energy–Accuracy Aware Finger Gesture Recognition for Wearable IoT Devices, *Sensors*, 22:4801, 2022.
- [25] Z. Wang, Z. Wu, R. Qi, X. Dou, Lightweight Visual Dynamic Gesture Recognition System Based on CNN-LSTM-DSA, *Sensors*, 26:1558, 2026.
- [26] A. O. Hashi, S. Z. M. Hashim, A. B. Asamah, A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024, *IEEE Access*, 12:143599–143626, 2024.
- [27] S. N. Reddy Karna, J. S. Kode, S. Nadipalli, S. Yadav, American Sign Language Static Gesture Recognition using Deep Learning and Computer Vision, *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, pp. 1432–1437, IEEE, 2021.
- [28] S. Misal, Hand Gesture Recognition Using Deep Learning, *International Journal of Innovative Science and Research Technology*, 9:69–72, 2024.
- [29] A. Khurshid, R. Grunitzki, R. G. Estrada Leyva, F. Marinho, B. M. M. Souto Orlando, Hand Gesture Recognition for User Interaction in Augmented Reality (AR) Experience, In: *Virtual, Augmented and Mixed Reality: Design and Development*, Lecture Notes in Computer Science, 13317:306–316, Springer, Cham, 2022.
- [30] A. S. M. Miah, M. A. M. Hasan, Y. Tomioka, J. Shin, Hand Gesture Recognition for Multi-Culture Sign Language Using Graph and General Deep Learning Network, *IEEE Open Journal of the Computer Society*, 5:144–155, 2024.
- [31] J. Li, B. Zhang, W. Chen, C. Bu, Y. Zhao, X. Zhao, Improving Hand Gesture Recognition Robustness to Dynamic Posture Variations by Multimodal Deep Feature Fusion, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3104–3115, 2024.
- [32] A. Mohan, H. K. Meena, M. Wajid, A. Srivastava, Real-Time Dynamic Hand Gesture Recognition Using mmWave Radar on FPGA, *IEEE Embedded Systems Letters*, 18:292–295, 2026.