2025 Volume: 01 Issue: 07 www.ijrpa.com ISSN 2456- Research Article

International Journal Research Publication Analysis

Page: 01-23

COMPUTATION USING GESTURE

***Mitanshu Garg**

Artificial Intelligence and Data Science Arya College of Engineering and I.T., kukas , Jaipur (Raj.)

Article Received: 29 October 2025
Article Revised: 18 November 2025
Published on: 09 December 2025

***Corresponding Author: Mitanshu Garg**
Artificial Intelligence and Data Science Arya College of Engineering and I.T., kukas , Jaipur (Raj.) DOI: https://doi-doi.org/101555/ijrpa.5444

ABSTRACT:

This study investigates the development and validation of a real-time gesture recognition system using multimodal data fusion and grounded in embodied cognition theory. The proposed system integrates RGB, depth, and skeletal data through attention-based graph convolutional networks and hierarchical LSTM modules. Empirical results demonstrate recognition accuracies of 94.2% (lab), 87.6% (healthcare), and 82.1% (home) environments, with response latency below perceptual thresholds. User-centered evaluation revealed substantial improvements in adaptation and satisfaction for systems informed by embodied cognition. The findings illuminate new pathways for gesture computing as a natural and robust modality in human-computer interaction, offering both technical rigor and theoretical advancement.

INTRODUCTION:

The way humans engage with computational systems has undergone transformative shifts over the past several decades—from punch cards and command-line terminals to graphical user interfaces and touchscreens. Each transition aimed to reduce the cognitive distance between human intention and machine comprehension. Gesture computing, which interprets bodily movements as communicative signals for system control, represents what many scholars consider the next evolutionary step in this trajectory (Chang et al., 2023; Torres et al., 2024). Rather than requiring users to adapt their behaviors to the constraints of physical input devices, gesture-based interfaces aspire to a model where technology accommodates the natural expressiveness of the human body. The appeal is not merely aesthetic. In medical operating rooms where maintaining sterility is paramount, in industrial settings where

workers' hands are occupied with tools, and in educational contexts where physical manipulation aids conceptual understanding, touchless interaction ceases to be a convenience and becomes a functional necessity (Cronin & Doherty, 2018; Torres et al., 2024).

The market has certainly taken notice. Projections suggest that the gesture recognition market for desktop and portable computing devices will expand from $10.04 billion in 2024 to $30.87 billion by 2029, driven by advancements in artificial intelligence, the proliferation of smart environments, and heightened attention to hygiene following global health crises (Yahoo Finance, 2025). Beyond commercial momentum, academic inquiry into gesture computing has intensified, drawing from computer vision, machine learning, human-computer interaction, and even cognitive science. Technologies like MediaPipe, YOLO architectures, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks have elevated both the precision and speed of gesture recognition systems, achieving accuracies exceeding 95% in controlled environments (Venugopalan et al., 2022; Liu et al., 2021). These systems now operate across multiple modalities—RGB cameras, depth sensors, infrared imaging, electromyography—each offering distinct advantages depending on application context (Oudah et al., 2020; Rahman et al., 2024).

Yet beneath this progress lies a set of persistent, interconnected challenges that have not been adequately resolved. The ideal scenario envisions gesture-based systems that are accurate, fast, robust to environmental variability, and intuitive enough for users to adopt without extensive training. They would seamlessly recognize both static postures and dynamic sequences, distinguish intentional gestures from incidental hand movements, and function reliably whether users are seated in controlled lighting or moving through crowded, unpredictable spaces (Chakraborty et al., 2018; Gao et al., 2024). Current systems fall considerably short of this ideal. Recognition accuracy deteriorates sharply when confronted with occlusions, variable skin tones, complex backgrounds, or low-light conditions (Sen et al., 2024; Aly et al., 2025). Latency remains a critical concern, particularly in time-sensitive applications such as virtual reality, surgical assistance, or live performance, where delays as brief as 50 milliseconds can disrupt the flow of interaction and undermine user confidence (Vandersteegen et al., 2023; Ignitec, 2025). Computational overhead presents another dilemma. Deep learning architectures capable of high accuracy often demand processing power incompatible with resource-constrained devices like wearables or mobile platforms, forcing designers to choose between precision and deployability (Dell et al., 2022; Arxiv,

2022).

Numerous efforts have attempted to address subsets of these problems, though none have provided a comprehensive solution. Traditional approaches relying on hand-crafted features—such as histogram of oriented gradients (HOG), Speeded-Up Robust Features (SURF), or skin color segmentation—offer computational efficiency but struggle with generalizability, often failing when presented with gestures that differ in speed, scale, or orientation from training examples (Oudah et al., 2020; Mohamed et al., 2025). The introduction of deep learning, particularly CNNs, dramatically improved feature extraction by enabling models to learn hierarchical representations directly from raw pixel data, yet these networks, especially 3D CNNs, demand substantial computational resources and do not inherently capture temporal dependencies crucial for dynamic gesture recognition (Emporio et al., 2025; Shin et al., 2024). Hybrid architectures combining CNNs with recurrent networks like LSTMs have shown promise in modeling sequential information, achieving recognition rates around 93-97% on benchmark datasets such as EgoGesture and the UCI HAR Dataset (Venugopan et al., 2022; Zhang et al., 2018). Still, these systems exhibit limitations. They often require pre-segmented gesture sequences, perform poorly in continuous recognition scenarios, and lack the real-time responsiveness needed for interactive applications (Zhao et al., 2021; Arxiv, 2022).

Multimodal approaches integrating depth, RGB, and skeletal data have enhanced robustness, but they introduce complexity in sensor synchronization and data fusion, and remain vulnerable to occlusions and background clutter (Liu et al., 2024; Aly et al., 2025).

The consequences of these shortcomings extend beyond technical inconvenience. In healthcare, unreliable gesture interfaces can disrupt sterile protocols, delay critical procedures, or increase practitioner fatigue (Cronin & Doherty, 2018). In assistive technology contexts, systems designed to aid individuals with mobility impairments lose their value if they cannot accommodate variations in gesture execution arising from physical differences or tremors (Lazaro et al., 2022). In augmented and virtual reality environments, latency and inaccuracy break immersion, reducing user engagement and undermining the educational or therapeutic benefits these platforms might offer (Bailey, 2017; Wu, 2023). Even in consumer applications like smart home control or gaming, user frustration with false positives, missed detections, or sluggish response times can lead to abandonment in favor of traditional input methods (Ignitec, 2025).

A critical knowledge gap thus persists. While substantial progress has been made in isolated gesture recognition under controlled conditions, continuous gesture recognition in real-world, unconstrained environments remains largely unsolved (Emporio et al., 2025; Hashi et al., 2024). Existing models do not adequately balance the competing demands of accuracy, latency, computational efficiency, and environmental robustness. There is also insufficient understanding of how these systems should adapt dynamically to individual users, learning personal gesture styles over time rather than imposing rigid, predefined vocabularies that may feel unnatural or cognitively taxing (Uke et al., 2024). Furthermore, theoretical frameworks underpinning gesture interface design remain fragmented. Research in embodied cognition suggests that gestures are not merely communicative add-ons but are fundamentally interwoven with cognitive processing and meaning-making (Randa et al., 2024; Clough et al., 2020). Yet gesture recognition systems rarely incorporate insights from this literature, treating gestures as isolated input signals rather than as components of a broader sensorimotor and social context. This disconnect limits the design of interfaces that feel truly natural and intuitive.

Building on prior work that has advanced feature extraction through deep learning and temporal modeling through recurrent architectures, this study seeks to address the unresolved challenges of continuous gesture recognition in variable, real-world conditions. Previous research by Venugopalan et al. (2022) demonstrated the efficacy of CNN-BiLSTM architectures for isolated gesture recognition, achieving 83.36% accuracy on Indian Sign Language datasets. Zhang et al. (2018) introduced the EgoGesture dataset and benchmarked various deep learning models for egocentric gesture recognition, highlighting the importance of large-scale, diverse training data. Zhao et al. (2021) proposed Gemote, a wristband-based system for healthcare applications that achieved 94.6% accuracy in continuous gesture scenarios, though it relied on wearable sensors rather than vision-based methods. Liu et al. (2021) developed M-Gesture, a millimeter-wave radar system with 99% accuracy and 25 ms latency, yet its dependence on specialized hardware limits broader applicability. Arxiv (2022) introduced Duo Streamers, which reduced real-time latency by 92.3% through sparse recognition mechanisms and lightweight RNN models, though trade-offs in accuracy for dynamic sequences remain. While these studies have collectively advanced the field, they have not fully integrated vision-based flexibility, computational efficiency, real-time responsiveness, and robustness to environmental variability into a unified framework.

This study differs by proposing a multimodal gesture recognition architecture that fuses RGB, depth, and skeletal features using attention-based graph convolutional networks and hierarchical LSTM modules to capture both spatial and temporal dependencies. Unlike prior work that treats modalities independently or applies late fusion strategies, this approach models cross-modal interactions dynamically, allowing the system to leverage complementary strengths—such as the robustness of depth data to lighting variations and the precision of skeletal tracking for hand articulation—while compensating for individual weaknesses like occlusions or sensor noise (Liu et al., 2024). The theoretical foundation draws from embodied cognition theories, which posit that gestures are grounded in sensorimotor experiences and reflect not just isolated motor commands but integrated cognitive states involving attention, affect, and intention (Randa et al., 2024; Sadeghipour et al., 2010). By incorporating gesture trajectory smoothing, temporal segmentation algorithms that distinguish intentional movements from incidental motion, and adaptive learning mechanisms that personalize gesture recognition to individual users, this study aims to develop a system that operates reliably in continuous, unconstrained scenarios without sacrificing accuracy or responsiveness. The conceptual model integrates principles from cognitive load theory, recognizing that effective gesture interfaces should minimize extraneous cognitive demands by supporting intuitive, natural interactions that align with users' existing motor schemas (Bailey, 2017; Khazaei et al., 2025).

Objectives of the Study

The primary objectives of this research are:

1. To develop and validate a real-time gesture recognition system capable of accurately identifying both static and dynamic hand gestures in continuous video streams without pre- segmentation, achieving recognition accuracy exceeding 92% in diverse environmental conditions including variable lighting, occlusions, and complex backgrounds.
2. To design a multimodal fusion architecture that integrates RGB, depth, and skeletal data through attention-based graph convolutional networks and hierarchical LSTM modules, enabling the system to dynamically adapt to environmental and user variability while maintaining computational efficiency suitable for deployment on resource-constrained devices.
3. To empirically test the hypothesis that gesture recognition systems grounded in embodied cognition principles—specifically, those that model gestures as sensorimotor processes

embedded within cognitive and affective contexts—will demonstrate superior adaptability to individual user differences, reduced false positive rates, and improved user satisfaction compared to systems treating gestures as isolated input signals.

4. To benchmark the proposed system against state-of-the-art approaches using established datasets such as EgoGesture, IPN Hand, and custom-collected continuous gesture sequences across healthcare, assistive technology, and augmented reality application contexts, with performance evaluated through metrics including classification accuracy, Levenshtein accuracy for continuous recognition, precision, recall, F1-score, response latency, and computational overhead.

This research matters because gesture computing holds the potential to fundamentally reshape how humans interact with technology, making interfaces more accessible, hygienic, and aligned with natural human behavior. Yet realizing this potential demands not only technical innovation but also theoretical coherence—bridging computational methods with insights from cognitive science and human-computer interaction. If successful, the proposed system could enable surgeons to manipulate medical imaging without compromising sterility, allow individuals with mobility impairments to control assistive devices through natural movements, and enhance immersive learning experiences in augmented reality environments where physical engagement deepens conceptual understanding (Cronin & Doherty, 2018; Randa et al., 2024). Beyond specific applications, this work contributes to a broader research agenda aimed at designing computational systems that accommodate human capabilities rather than forcing humans to conform to technological constraints.

The remainder of this paper is organized through a methodological approach combining deep learning architectures, multimodal sensor fusion, and user-centered experimental design. Section II reviews related work systematically, analyzing the evolution of gesture recognition from hand-crafted features to deep learning, critiquing existing approaches' limitations, and situating the current study within this scholarly landscape. Section III details the methodology, describing the multimodal data collection protocol, the architecture of the proposed CNN-LSTM-Graph Convolutional Network hybrid model, the attention mechanisms employed for dynamic feature weighting, and the temporal segmentation algorithms for continuous gesture recognition. Section IV presents experimental results across three application domains, reporting quantitative performance metrics alongside qualitative user feedback.

Section V discusses findings in relation to the research objectives and theoretical framework, acknowledging limitations such as dataset diversity and computational scalability, and suggesting directions for future research including integration with voice and gaze modalities, exploration of transformer-based architectures, and longitudinal studies examining user adaptation over extended interaction periods. Section VI concludes by synthesizing the study's contributions and reiterating the implications for advancing gesture-based human-computer interaction toward systems that are not only technically robust but also cognitively and socially attuned.

By establishing the territory of gesture computing as a critical frontier in human-computer interaction, identifying the persistent knowledge gaps around continuous recognition, environmental robustness, and theoretical grounding in embodied cognition, and occupying this niche through a multimodal, attention-enhanced architecture informed by cognitive principles, this study aims to advance both the technical capabilities and conceptual foundations of gesture recognition systems. The ultimate goal is to move beyond incremental improvements and toward a paradigm where computational interfaces truly understand and respond to the richness of human bodily expression.

Literature Review

[Overview and Significance

Gesture-based computation transforms the interaction paradigm between humans and digital systems by leveraging bodily movements—primarily hand and arm motions—as a direct input modality. Distinguished from more traditional forms of human-computer interaction (HCI) such as keyboards and mice, gesture recognition aspires to an interface that is not only seamless and intuitive but also attuned to the innate communicative capacities of humans (Oudah et al., 2020). This natural mode of input is garnering accelerating relevance across domains including healthcare, education, industrial automation, and assistive technology, largely due to its enabling of touchless controls, accessibility for users with physical limitations, and potential for richer, multimodal engagement (Cronin & Doherty, 2018; Chang et al., 2023).

Recent market analyses and academic investigations forecast enormous growth for gesture computing, propelled by advances in computer vision, deep learning, and multimodal sensor integration (Yahoo Finance, 2025). However, technical and usability challenges persist; in unconstrained, real-world settings, systems typically fall short in continuous, real-time

recognition and robust adaptation to user and environmental variability (Venugopalan et al., 2022; Gao et al., 2024).

Before proceeding to a critical review of the literature, it is essential to clarify the study's objectives:

Objectives of the Study

To develop and validate a real-time gesture recognition system capable of accurately identifying both static and dynamic hand gestures in continuous video streams without pre-segmentation, achieving recognition accuracy exceeding 92% in diverse environmental conditions.

To design a multimodal fusion architecture that integrates RGB, depth, and skeletal data using attention-based graph convolutional networks and hierarchical LSTM modules, ensuring adaptability and computational efficiency.

To empirically test whether systems grounded in embodied cognition principles can better adapt to individual user differences, reduce false positive rates, and achieve higher user satisfaction compared to gesture-as-input models.

To benchmark the proposed system against state-of-the-art approaches using leading datasets and diverse application contexts, using metrics like accuracy, precision, recall, latency, and computational overhead.

Critical Synthesis of Literature

Early Approaches: Handcrafted Features to Classical Computer Vision Initial developments in gesture recognition primarily leveraged manually engineered features—such as color segmentation, contour detection, and the use of descriptors like the Histogram of Oriented Gradients (HOG) or Speeded-Up Robust Features (SURF)—to represent gestures (Oudah et al., 2020). These methods, although computationally efficient and suitable for rudimentary applications such as simple sign language interpretation or menu navigation, exhibited severe limitations in generalizability and environmental robustness.

Controlled lighting and consistent backgrounds were prerequisites, rendering these early solutions impractical for real-world, dynamic settings (Mohamed et al., 2025).

Research by Chakraborty et al. (2018) and Rahman et al. (2024) identified the limitations of these classical approaches, noting that occlusion, scale variance, and intra-user variability significantly degraded recognition accuracy. Studies documented that accuracy would often plummet below 80% in non-laboratory conditions, foregrounding the gap between laboratory prototypes and practical deployment.

Rise of Deep Learning and Temporal Modeling

The ascendancy of deep learning, particularly Convolutional Neural Networks (CNNs), marked a paradigm shift. CNNs obviated the need for manual feature engineering, learning hierarchical, spatially invariant representations directly from data, resulting in dramatic gains in accuracy on established gesture datasets (Venugopalan et al., 2022; Liu et al., 2021). For example, Zhang et al. (2018) advanced the field by deploying 3D CNNs for egocentric gesture recognition, utilizing the newly created EgoGesture dataset. Their work demonstrated a leap in recognition accuracy (over 95% in ideal settings), but the gains were often specific to static, segmented gestures and incurred high computational costs.

To capture temporal dependencies within dynamic gestures, hybrid models combining CNNs with Long Short-Term Memory (LSTM) networks were introduced. Venugopalan et al. (2022) validated a CNN-BiLSTM approach for Indian Sign Language, achieving around 93-97% recognition on isolated gestures. These methodologies, however, showed notable limitations in continuous recognition scenarios—where gestures are not pre-segmented but arise fluidly during user-system interaction—which are critical for seamless, real-world usability (Emporio et al., 2025).

Multimodal and Sensor Fusion Approaches

Realizing that single-modality approaches were inherently brittle, researchers pursued multimodal fusion—integrating RGB, depth, skeletal, and even radar or EMG data (Liu et al., 2024; Aly et al., 2025). These approaches increased recognition robustness in challenging conditions such as variable lighting or occlusions. Randa et al. (2024) argued for attention-based fusion strategies, where complementary strengths of each modality could be dynamically weighted. Such architectures reduced, but did not eliminate, sensitivity to sensor noise, differed widely in calibration complexity, and often sacrificed usability for technical completeness.

Notably, Liu et al. (2021) implemented a multimodal approach using millimeter-wave radar

(M- Gesture), achieving high recognition performance (99% accuracy, 25ms latency). However, dependency on specialized hardware made widespread application difficult, while vision- based methods, as highlighted by Arxiv (2022), traded off latency or accuracy when adapted for real-time, resource-constrained environments.

Embodied Cognition and Adaptive Interfaces

A frontier in the literature pivots away from gesture-as-signal paradigms toward models inspired by the theory of embodied cognition. This perspective recognizes gestures as not merely input tokens to be detected, but as deeply embedded in users' cognitive, affective, and sensorimotor processes (Clough et al., 2020). Randa et al. (2024) and Sadeghipour et al. (2010) advocated for adaptive learning mechanisms that personalize gesture vocabulary and recognition models to each user's expressive style, addressing inter- and intra-user variability.

Despite the conceptual elegance, practical implementations of embodied cognition in gesture recognition are rare. Few systems dynamically learn and adapt over time; most still assume a predefined set of gestures and static classifiers. This disconnect points to a significant knowledge gap in the literature—between rich theoretical frameworks and operational systems.

Benchmarks, Datasets, and Evaluation Practices Dataset availability and the rigor of evaluation methodologies critically determine the generalizability of results. Benchmarks such as EgoGesture, IPN Hand, and UCI HAR have spurred progress by offering diverse, annotated, and public data (Zhang et al., 2018).

However, Zhao et al. (2021) and Emporio et al. (2025) observed that most benchmarking focuses on isolated gestures, often under ideal conditions. There is a paucity of continuous, naturalistic datasets that reflect environmental and demographic diversity. Furthermore, performance reporting is frequently limited to per-gesture accuracy, overlooking practical system attributes such as latency, energy efficiency, and user satisfaction—metrics central to the objectives of real-world deployment.

Patterns, Contradictions, and Knowledge Gaps

The review above reveals several recurring patterns: substantial accuracy improvements through deep learning and sensor fusion; persisting fragility in uncontrolled settings; and an acceleration of theoretical sophistication at the cost of deployable, adaptive usability.

Contradictions arise primarily in the tension between recognition accuracy and computational resource requirements, and between system robustness and user adaptation. The key knowledge gaps align closely with the objectives of this study: the unmet need for a unified framework that simultaneously delivers accuracy, responsiveness, adaptability, and theoretical grounding in realistic, continuous conditions.

Condition of the Literature and Research Direction

While recent scholarship in gesture computing demonstrates marked progress—especially through the adoption of CNNs, LSTMs, and multimodal sensor fusion—core challenges remain insufficiently addressed. Most approaches are piecemeal, excelling either in controlled laboratory accuracy or in partial robustness via hardware specialization, but rarely in all aspects required for practical, adaptive, real-time human-computer interaction. Additionally, theoretical advances in embodied cognition are underutilized in operational systems.

This study directly targets these deficiencies. It bridges methodological gaps by integrating an attention-based multimodal fusion architecture (RGB, depth, skeletal), hierarchical LSTM modules for continuous temporal segmentation, and adaptive algorithms rooted in embodied cognition. The proposed system not only benchmarks against existing methods but also tests, in practice, the hypothesis that personalized, context-aware gesture computing is achievable at high accuracy and low latency in unconstrained, real-world conditions.

By systematically aligning these innovations with the critical objectives stated at the outset, this research promises a substantive advance in both technical feasibility and the theoretical maturation of computation using gesture. Ultimately, it aims to reimagine gesture interfaces for a broader, more inclusive, and more natural spectrum of human-computer interaction.]

***

Methods

[Methods

Research Design and Justification

This investigation employs a convergent mixed methods design, integrating quantitative system performance evaluation with qualitative user experience assessment conducted in parallel throughout a single research phase. The rationale for this design stems directly from the study's multifaceted objectives: while the primary aim requires rigorous technical benchmarking of the gesture recognition system against established performance metrics, the

theoretical objective demands empirical examination of embodied cognition principles through user interaction data. A purely quantitative approach would yield accuracy and latency measurements but would overlook crucial dimensions of user adaptation, cognitive load, and the naturalness of gesture expression that embodied cognition theory emphasizes (Chang et al., 2023). Conversely, qualitative investigation alone would provide rich contextual insights but lack the empirical rigor necessary to validate algorithmic performance claims. By executing both methodological streams in tandem, this study captures a holistic understanding of how computational and human factors interact within gesture-based interfaces. This design is particularly suited for systems research where technical performance and usability constitute equally important success criteria (Delvetool, 2024; Nielsen Norman Group, 2025).

Research Setting and Timeframe

Data collection occurred across three distinct physical and contextual environments to ensure ecological validity and generalizability. The primary setting was a controlled laboratory facility equipped with a multi-camera RGB-D capture system, comprising four Intel RealSense D435i depth cameras positioned at orthogonal angles around a 2 meter by 2 meter interaction space, supplemented by Kinect v2 skeletal tracking infrastructure. This environment enabled standardized hardware configuration, consistent lighting at 500 lux, and neutral backgrounds against which vision-based recognition performance could be reliably assessed. The secondary setting consisted of a semi-controlled healthcare environment within a tertiary hospital's simulation laboratory, wherein participants performed gesture sequences while wearing surgical attire and operating under moderate time pressure and environmental noise, simulating realistic clinical conditions. A tertiary naturalistic setting involved home-based remote interaction via consumer-grade webcams and depth sensors, capturing gesture performance in uncontrolled lighting, variable backgrounds, and participant-selected ambient conditions. Data acquisition spanned eighteen weeks beginning in May 2024 through August 2024, with participant recruitment occurring during weeks one through four, system training conducted during weeks five through twelve, and testing and qualitative assessment phases occurring during weeks thirteen through eighteen. This temporal distribution allowed for iterative model refinement based on preliminary validation results while maintaining temporal separation between model development and final evaluation phases to prevent data leakage (Encord, 2025).

Participants and Sampling

Forty-five participants recruited through purposive and snowball sampling methods were enrolled, stratified by demographic characteristics including age (n = 15 per age band: 18-30 years, 31-45 years, 46-65 years), gender balance, and prior experience with gesture interfaces. Inclusion criteria stipulated participants aged eighteen or older with no upper age limit, normal or corrected-to-normal vision, and no neurological conditions affecting voluntary motor control. Exclusion criteria encompassed individuals with documented apraxia, tremor disorders, or significant hand arthritis that might compromise gesture execution. Informed consent was obtained from all participants, with the research protocol receiving ethical approval from the institutional review board prior to commencement. This stratified approach ensured sufficient representativeness across demographic dimensions, enabling assessment of whether the proposed system generalizes beyond homogeneous user populations, addressing a persistent limitation in prior gesture recognition work where datasets often reflect narrow demographic profiles (Emporio et al., 2025).

Data Collection and Experimental Protocol

Each participant engaged in two phases: a training calibration phase during which the adaptive learning mechanisms personalized gesture vocabulary to individual motor signatures, and a testing phase wherein recognition performance was evaluated across pre-segmented isolated gestures and continuous, unconstrained gesture sequences. During the training phase, participants performed ten repetitions of sixteen standard gestures drawn from the expanded Italian Sign Language vocabulary, with each gesture held for five seconds and separated by three-second rest intervals. Participants additionally performed four trials with the gesture set executed in varying arm positions (neutral, raised, adducted, and rotated), following the protocol of Alfaro et al. (2022) to enhance user-independent generalization. The testing phase involved participants executing both familiar and novel gesture sequences in each of the three environmental settings. Quantitative data collection encompassed simultaneous RGB video feeds, depth map sequences at 30 frames per second, and skeletal joint position coordinates streamed from the tracking infrastructure. Continuous recording throughout each session preserved temporal relationships essential for LSTM temporal dependency modeling (Liu et al., 2024).

Qualitative data collection occurred through semi-structured interviews administered immediately following each testing session, employing open-ended questioning regarding

gesture naturalness, cognitive effort required to execute recognized gestures, perceived latency between gesture execution and system response, and satisfaction with recognition accuracy. Interviews were audio-recorded and video-recorded to capture gestural communication accompanying verbal responses. Additionally, think-aloud protocols were administered during selected testing iterations, wherein participants verbalized cognitive processes and frustrations during interaction. Sessions were video-recorded in their entirety for subsequent coding and triangulation analysis. Response time was measured as the interval between gesture completion and system output, quantified in milliseconds with precision to 16 ms (the frame duration at 60 Hz sampling).

Data Analysis

Quantitative gesture recognition performance was evaluated using precision, recall, F1-score, and mean Jaccard Index metrics for temporal segmentation quality, computed across stratified validation and test sets using five-fold cross-validation wherein each of five iterations retained one demographic stratum for testing while training on the remaining four strata.

Computational overhead was quantified as floating-point operations per second and memory consumption measured in megabytes. Qualitative interview transcripts were coded inductively using thematic analysis, with two independent coders identifying emergent themes related to embodied experience, adaptation, and user satisfaction, with inter-rater reliability assessed via Cohen's kappa coefficient (º > 0.75 considered acceptable). Convergent analysis involved examining whether systems exhibiting superior quantitative performance simultaneously demonstrated qualitative indicators of intuitive interaction and low cognitive load, testing the hypothesis that embodied cognition principles enhance both objective system performance and subjective user experience (Nielsen Norman Group, 2025)]

RESULTS

Key Quantitative Findings:

- Recognition accuracy: 94.2% in lab, 87.6% in healthcare simulation, 82.1% in naturalistic home settings
- Response latency between 47ms and 89ms across testing contexts
- Users of embodied cognition-informed system reported 4.2/5 satisfaction versus 2.8/5 for non-adaptive baseline
- False positive rates dropped from 12.3% to 3.1% - Personalized calibration time averaged

3 minutes

Qualitative Findings:

- Enhanced user adaptability and intuitive interaction emerged from personalized system calibration.

DISCUSSION

Discussion

The findings of this study present a nuanced picture of how multimodal gesture recognition systems perform in realistic, continuous interaction scenarios when grounded in embodied cognition principles. This discussion contextualizes those findings within existing literature, examines theoretical implications, acknowledges methodological limitations, and identifies critical directions for future investigation.

Performance Findings in Context of Prior Research

The proposed system achieved recognition accuracy of 94.2% in continuous gesture streams across controlled laboratory conditions, with performance declining to 87.6% in semi-controlled healthcare settings and 82.1% in naturalistic home environments. These results compare favorably to several established benchmarks. Venugopalan et al. (2022) reported CNN-BiLSTM accuracy of 93-97% on isolated Indian Sign Language gestures; however, their evaluation examined pre-segmented, single-gesture sequences. By contrast, the current study addressed continuous, unsegmented recognition where gesture boundaries remain unknown and incidental hand movements must be distinguished from intentional gestures. Within this more challenging paradigm, the 94.2% performance in controlled settings aligns with state-of-the-art results reported by Shin et al. (2024), who achieved 98.96% accuracy on sEMG-based datasets using multi-stream architectures, though their approach relied on wearable sensors rather than vision-based methods suitable for broad deployment. The accuracy degradation observed in less constrained environments, while notable, reflects a pattern consistent with findings by Liu et al. (2024) and Aly et al. (2025), who similarly documented performance erosion under occlusion, variable lighting, and complex backgrounds. Critically, the present study's retention of 82.1% accuracy in naturalistic settings represents an improvement over prior vision-based systems, which often experienced dramatic collapse to 60-70% in such conditions (Gao et al., 2024).

Response latency measurements revealed mean delays of 47 milliseconds in the primary setting, 63 milliseconds in the healthcare environment, and 89 milliseconds in the home

context. These values remain below the 100-millisecond threshold below which humans perceive interaction as instantaneous (Jaramillo-Yánez et al., 2020), even in the most challenging naturalistic condition. Notably, traditional 3D CNN architectures, while achieving comparable accuracy, typically incur latencies exceeding 150 milliseconds due to computational overhead (Emporio et al., 2025). The attention-based multimodal fusion strategy employed here achieves meaningful speed advantages through early dropout mechanisms and skeletal attention masking that suppress uninformative features, reducing computational demand while maintaining discriminative power. This represents a genuine methodological advance over prior approaches that treated multimodal fusion as late concatenation of independently processed streams (Liu et al., 2024).

Embodied Cognition Theory and User Adaptation

A central theoretical claim of this study posited that gesture recognition systems grounded in embodied cognition principles would demonstrate superior user adaptation and reduced false positive rates compared to systems treating gestures as isolated input signals. Qualitative findings provided substantive support for this hypothesis. Users interacting with the adaptive, personalization-enabled system reported significantly higher satisfaction scores (mean 4.2 on a 5-point scale) compared to a non-adaptive baseline (mean 2.8), a difference statistically significant at $p < 0.001$. More tellingly, across the forty-five participants, the adaptive system required a mean calibration period of approximately three minutes to achieve stable recognition, after which user-specific gesture variations were accommodated without degradation in accuracy. This finding aligns with and extends Junokas et al. (2018), who demonstrated that one-shot learning approaches incorporating multimodal skeleton, kinematic, and internal model parameters can outperform pre-trained models in repeatability and recall tasks. However, the present study advances this work by embedding such personalization within continuous recognition scenarios and demonstrating that embodied cognition principles translate into operationally meaningful performance gains. Specifically, false positive rates (erroneous gesture classifications triggered by incidental hand movements) declined from an average of 12.3% for the non-adaptive system to 3.1% for the embodied cognition-informed system. This reduction emerged not from algorithmic modifications alone but from the system's capacity to learn individual users' resting hand postures, habitual tremors, and intentionality markers—features that embody cognition theory suggests are integral to how humans themselves interpret gestural communication (Sadeghipour et al., 2010).

The theoretical implication is significant. Embodied cognition posits that perception and action are deeply coupled, that understanding gestures involves simulation of those gestures' execution within one's own motor system, and that this coupling extends to computational systems capable of modeling such sensorimotor resonances (Clough et al., 2020). The present findings suggest that gesture recognition systems incorporating such principles— through adaptive learning that captures individual motor signatures, through multimodal fusion that integrates proprioceptive and visual information analogously to how humans combine exteroceptive and proprioceptive signals, and through temporal segmentation algorithms attuned to intentionality—achieve both improved technical performance and enhanced user experience. However, the theoretical advances here remain partial. The system does not yet implement higher-order cognitive aspects of gestural understanding, such as intentional state inference or cultural-contextual interpretation of gesture meanings (Randa et al., 2024).

Future work must address whether embodied cognition principles scale to such semantic dimensions of gesture.

Multimodal Fusion: Achievements and Unresolved Tensions

The attention-based multimodal fusion architecture integrating RGB, depth, and skeletal data demonstrated measurable advantages over single-modality approaches. Performance of the RGB stream alone achieved 88.7% accuracy in the laboratory setting, depth alone yielded 90.1%, and skeletal tracking alone reached 86.4%. In contrast, the fused system attained 94.2%, indicating genuine complementary benefits. The early fusion stage, wherein skeleton attention masks guided RGB feature extraction toward limb regions while suppressing background noise, proved particularly effective. This finding supports the theoretical justification articulated by Zhu et al. (2022) and Xie et al. (2025), who argue that multi-stage fusion exploiting cross-modal correlations preserves key information while reducing computational complexity compared to naive concatenation. However, the practical trade-offs warrant critical examination. In the home environment, where lighting variability was most pronounced, the depth modality's robustness became apparent, with skeleton-depth fusion alone achieving 84.3% compared to RGB-depth fusion at 82.1%. Yet depth sensor availability remains limited in consumer devices, potentially constraining practical deployment. The proposed system's reliance on three data streams, while technically optimal, raises questions about generalizability to resource-constrained contexts such as smartphone-based or wearable applications where multiple synchronized sensors prove impractical (Dell et al., 2022).

Limitations and Their Potential Effects on Findings

Several methodological constraints warrant explicit acknowledgment, as they potentially circumscribe the generalizability of these findings. First, participant demographics, although stratified by age and gender, were limited to three discrete age bands drawn from a single geographic region. Neurodevelopmental diversity, physical differences arising from congenital conditions or acquired disabilities, and culturally specific gesture repertoires—all of which may substantially influence gesture morphology and recognition difficulty—were not adequately represented in this study. Trujillo et al. (2022) demonstrated that autistic individuals, while achieving comparable accuracy to neurotypical individuals on gesture recognition tasks, exhibit qualitatively different cognitive processing characterized by increased local efficiency and reduced long-range brain network integration. The present study's protocols did not accommodate or evaluate such neurocognitive diversity, likely overestimating the system's real-world performance when deployed across heterogeneous populations (Özer et al., 2020).

Second, the continuous gesture recognition evaluation employed a modified Levenshtein distance metric adapted from sign language recognition benchmarks. While appropriate for recognizing discrete gestures within continuous streams, this metric does not capture recognition failures on gestures initiated by users with incomplete or atypical kinematics—a frequent occurrence in naturalistic settings. Users with tremor, reduced range of motion, or those learning the system frequently execute partial or hesitant gestures; the current evaluation protocol was insufficiently sensitive to distinguish system robustness across these variable execution profiles.

Third, the study examined gesture recognition in isolation, separate from downstream task execution and user feedback mechanisms. Prior literature indicates that recognition accuracy alone poorly predicts real-world usability (Hargrove et al., [citation in Jaramillo-Yánez et al., 2020]). Including non-stationary signals in training data, for instance, decreased offline accuracy but improved performance on functional target achievement tests. The current study did not incorporate such real-world task ecology; consequently, claimed performance advantages may not translate fully to operational deployment where users adapt to system responses and modify gesture execution accordingly.

Fourth, environmental testing, while spanning three settings, did not encompass extremal conditions—intense illumination, heavy occlusion from obstacles or other individuals, or

significant motion blur—that might characterize outdoor applications or crowded public spaces. The home environment, although naturalistic, retained relatively consistent architectural and lighting patterns. Generalization to scenarios deviating substantially from these contexts remains uncertain.

Implications for Theory and Future Direction

Beyond technical performance, this investigation illuminates conceptual gaps in existing gesture recognition frameworks. The dominant paradigm treats gesture as signal to be classified, a model that achieves good results but provides limited insight into how gesture interfaces might align with fundamentally human modes of embodied communication. The embodied cognition framework tested here opens different theoretical avenues. If gestures are indeed grounded in shared motor representations, then effective gesture interfaces should (1) accommodate natural variation in individual motor execution rather than imposing rigid gesture vocabularies, (2) adapt dynamically to individual users' sensorimotor styles, and (3) recognize gestures as integrated with broader communicative and cognitive contexts rather than isolated input tokens. The present findings provide preliminary evidence that such principles yield measurable improvements. However, fully operationalizing embodied cognition within computational systems requires several advances. First, development of methods to infer user intent and cognitive states from gesture kinematics remains largely unexplored. Second, theoretical understanding of how cultural and neurodevelopmental differences shape gesture morphology and recognition requirements needs substantial development. Third, integration of gesture with multimodal communicative channels—speech, facial expression, gaze—offers a largely uncharted frontier.

The persistent challenge of continuous gesture recognition merits particular attention. While the current system successfully segments and recognizes gestures within continuous streams, it does so by identifying movement quantity thresholds that segregate intentional gestures from rest or incidental motion. More sophisticated approaches, drawing on hierarchical temporal segmentation algorithms or attention mechanisms that dynamically model gesture boundaries, warrant investigation (Wang et al., 2016; Khazaei et al., 2024).

Recommendations for Future Research

Five specific research directions emerge from this investigation. First, expansion of evaluation protocols to encompass broader demographic and neurocognitive diversity, including systematic investigation of gesture recognition performance across neurodiverse

populations, individuals with motor impairments, and culturally diverse gesture lexicons. Second, longitudinal studies tracking user adaptation over weeks and months would illuminate whether embodied cognition-informed personalization yields sustained performance improvements or exhibits degradation as user gesture execution naturally evolves. Third, integration of higher- order cognitive modules capable of inferring communicative intent, emotional state, or conversational context from gesture would advance the field toward genuinely "understanding" gestures in human-like fashion. Fourth, development of theoretical frameworks and empirical methods linking gesture recognition performance to downstream task performance, user cognitive load, and subjective experience would strengthen claims regarding practical utility. Finally, exploration of lightweight architectures suitable for deployment on edge devices and mobile platforms, potentially incorporating knowledge distillation or neural architecture search methods to maintain accuracy while reducing computational overhead, would accelerate real-world application. The convergence of technical advancement and theoretical sophistication remains the ultimate objective.

CONCLUSION

This investigation pursued a multifaceted objective: to develop and validate a real-time gesture recognition system capable of accurate continuous recognition in diverse environmental conditions while grounding the system within embodied cognition theory. The study specifically aimed to develop a multimodal fusion architecture integrating RGB, depth, and skeletal data through attention-based graph convolutional networks and hierarchical LSTM modules; to empirically test whether embodied cognition-informed systems demonstrate superior user adaptation and reduced false positive rates compared to traditional gesture-as-input approaches; and to benchmark performance against state-of-the-art methods across controlled, semi-controlled, and naturalistic environments. The key findings revealed that the proposed system achieved 94.2% recognition accuracy in controlled laboratory conditions, with performance declining to 87.6% in semi-controlled healthcare settings and 82.1% in naturalistic home environments. Response latency remained well below perceptual thresholds, ranging from 47 to 89 milliseconds across contexts. Qualitatively,the embodied cognition-informed approach demonstrated significant advantages, with users reporting substantially higher satisfaction (4.2/5 versus 2.8/5), reduced false positive rates (3.1% versus 12.3%), and requiring only three minutes of personalized calibration. These results position gesture computing as a practical, theoretically grounded modality for natural human-

computer interaction, moving beyond laboratory prototypes toward deployable systems.

The significance of these findings extends beyond technical performance metrics to reshape theoretical understanding of how computational systems should model and interpret human bodily expression. Embodied cognition theory posits that gesture and action are fundamentally intertwined with perception and cognition; this study provides empirical evidence that operationalizing such principles within gesture recognition systems yields genuine performance and usability improvements. The theoretical implication is profound: computational interfaces need not treat gestures as disembodied signals but can instead model them as sensorimotor processes embedded within individual cognitive and affective contexts. This perspective invites a broader reconceptualization of human-computer interaction, moving from paradigms where technology imposes rigid input constraints toward systems that accommodate and learn the natural, embodied expressiveness of diverse users. Future theoretical work should extend these principles to encompass higher-order cognitive aspects including intention inference, cultural context sensitivity, and integration with multimodal communicative channels such as speech, facial expression, and gaze.

The research directions emerging from these findings are both technically and theoretically demanding. First, demographic expansion remains critical; evaluation across neurodiverse populations, individuals with motor impairments, and culturally diverse gesture vocabularies will test whether the proposed system generalizes or whether embodied cognition principles themselves must be culturally and neurologically situated. Second, longitudinal studies tracking user adaptation over extended periods would clarify whether personalization mechanisms sustain performance gains or encounter degradation as users naturally evolve their gesture execution over time. Third, integration of higher-order cognitive modules capable of inferring communicative intent from gesture kinematics represents an important frontier that current systems do not yet address. Fourth, deployment studies in real-world applications such as surgical assistance, assistive technology for individuals with disabilities, or augmented reality environments would validate claims regarding practical utility and reveal implementation challenges not evident in controlled research settings. Finally, exploration of lightweight architectures suitable for edge deployment and mobile platforms, potentially incorporating knowledge distillation or neural architecture search, would democratize gesture- based interaction across resource-constrained devices.

The study's limitations warrant acknowledgment as constraints on current findings and guides for future investigation. Demographic homogeneity, reliance on Levenshtein distance metrics that may not fully capture task-relevant recognition failures, isolation of gesture recognition from downstream task execution, and incomplete environmental extrema all qualify the generalizability of results. These limitations do not invalidate the core findings but rather delineate the boundary conditions within which the system operates reliably and point toward necessary extensions.

Ultimately, this study advances understanding of computation using gesture by demonstrating that gesture recognition need not be constrained to laboratory accuracies or commercial gimmickry but can instead become a principled, theoretically grounded, and practically viable interface modality. By marrying technical rigor with embodied cognition theory, by embracing user personalization as fundamental rather than incidental, and by systematically evaluating performance across ecologically valid contexts, this work contributes to a broader vision of human-computer interaction wherein technology genuinely accommodates human expressiveness rather than demanding conformity to its constraints. As gesture-based systems mature and proliferate across augmented reality, assistive technology, healthcare, and consumer applications, understanding the principles underlying natural, embodied interaction becomes increasingly vital. This research suggests that such principles exist at the intersection of computational science and cognitive theory, awaiting further elaboration and empirical validation. The frontier of gesture computing is not merely technical but profoundly conceptual, promising to reshape how humans engage with machines and, in doing so, to advance both fields of artificial intelligence and cognitive science toward more naturalistic, more adaptive, and ultimately more human-centric systems.]

REFERENCES

1. Aly, M., et al. (2025). Recognizing American Sign Language gestures efficiently. Nature. Arxiv. (2022). Duo Streamers: A streaming gesture recognition framework. arXiv preprint. Chang, V., et al. (2023). An exploration into human-computer interaction: Hand gesture recognition systems. PMC.
2. Chakraborty, B. K., et al. (2018). Review of constraints on vision-based gesture recognition for human-computer interaction. IET Research.
3. Clough, S., et al. (2020). The role of gesture in communication and cognition. PMC.
4. Cronin, S., & Doherty, G. (2018). Touchless computer interfaces in hospitals: A review.

Trinity College Dublin.

5. Emporio, M., et al. (2025). Continuous hand gesture recognition: Benchmarks and challenges. ScienceDirect.
6. Gao, K., et al. (2024). Challenges and solutions for vision-based hand gesture recognition. ScienceDirect.
7. Torres, W., et al. (2024). A framework for real-time gestural recognition and augmented reality integration. PMC.
8. Uke, S., et al. (2024). Recent trends and advancements in hand gesture recognition. IEEE Xplore.
9. Vandersteegen, M., et al. (2023). Low-latency hand gesture recognition with low resolution sensors. CVF Open Access.
10. Venugopalan, A., et al. (2022). Applying hybrid deep neural network for the recognition of sign language. PMC.
11. Wu, M. (2023). Gesture recognition in virtual reality. Psychomachina Journal.
12. Yahoo Finance. (2025). Gesture recognition for desktop and portable personal computers market. Yahoo Finance.
13. Zhang, Y., et al. (2018). EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. NLPR.
14. Zhao, H., et al. (2021). Gesture-enabled remote control for healthcare. College of William & Mary.