A Wireless Hybrid Assistive Interface: Fusing Eye-Gaze Estimation with Inertial Motion for Precision Control

Abhishek S K[1], Haripriya A P[2]
[1]*Department of Mechanical Engineering, Government Engineering College Barton Hill*
[2]*Department of Information Technology, Government Engineering College Barton Hill*
doi.org/10.64643/IJIRTV12I10-204964-459

***Abstract*—The rapid evolution of digital computing has outpaced the development of inclusive input modalities, leaving a significant gap for users in sterile environments or those with limited manual dexterity. Traditional Human-Computer Interaction (HCI) peripherals, such as the mechanical mouse and keyboard, inherently rely on high-precision motor control and physical contact, which are often unfeasible in clinical surgical settings or for individuals with neurodegenerative conditions. While standalone vision-based systems have attempted to bridge this gap, they frequently suffer from MidasTouch accidental triggers and tracking instability. This research paper delineates the design, architectural synthesis, and empirical validation of a high-fidelity, hybrid human-computer interface (HCI) specifically engineered to transcend the limitations of conventional touchless navigational systems. The core innovation lies in a multi-modal Coarse-to-Fine fusion architecture that synchronizes biological gaze estimation with mechanical inertial sensing. The hardware layer utilizes an ESP32 microcontroller as a central processing hub, which leverages the Bluetooth Serial Port Profile (SPP) to establish a low-latency, cable-free communication pipe with the host workstation. Coarse-grained spatial navigation is achieved through real-time eye-gaze tracking using the Media Pipe Face Mesh framework, which maps pupil-to-nose vector displacement across a high-definition 1920x1080 display matrix. To resolve the inherent instability and jitter associated with software-based gaze estimation, a secondary refinement layer is introduced via an MPU-6050 6-axis motion sensor. This sensor enables the user to perform micro-adjustments or nudge using subtle hand-tilt gestures, effectively decoupling large-scale cursor travel from precision target selection. Furthermore, a finger-mounted resistive flex sensor provides a deterministic binary trigger for click events, overcoming the ambiguity of vision-based gesture recognition. The system demonstrates a significant enhancement in ergonomic efficiency and navigational accuracy through the application of stochastic smoothing filters and the integration of the hardware into a specialized Nitrile-coated Nylon wearable interface.**

***Index Terms*—Hybrid Sensor Fusion, Gaze Estimation, Human-Computer Interaction (HCI), Assistive Technology.**

I. INTRODUCTION

The shift toward non-contact Human-Computer Interaction is catalyzed by a growing need for sterile, hands-free, and highly accessible computing environments. In the medical domain, specifically within perioperative environments, the sterile computing paradigm is essential; it allows surgeons to manipulate high-resolution 3D volumetric scans or navigate electronic health records without physical contact with peripherals. This leads to the management of the aseptic integrity of the operating theater. Beyond medical applications, industrial maintenance and assembly operations often require technicians to access digital diagnostic interfaces while their primary motor functions are occupied by physical tools. In this context, a gesture-based mouse acts as a seamless extension of the user's cognitive intent.

1.1. Rationale for Gesture-Controlled Mouse Systems and Applications

Most critically, for populations suffering from neurodegenerative conditions such as Amyotrophic Lateral Sclerosis (ALS), Multiple Sclerosis (MS), or high-level tetraplegia, traditional input devices are physiologically incompatible. Conventional peripherals, such as the mechanical mouse or trackpad, rely on fine motor control and significant grip strength—capabilities that are progressively lost in these patient groups. As muscle atrophy and motor neuron degeneration accelerate, the physical effort

required to overcome the friction of a desk-bound mouse becomes an insurmountable barrier to digital interaction. This exclusion creates a digital lock-in effect, where patients lose their primary means of communication and social autonomy precisely when they need it most. While existing eye-tracking solutions offer a partial remedy, they often suffer from the MidasTouch problem, where unintended glances trigger accidental clicks, leading to high user frustration and cognitive fatigue. Consequently, there is an urgent and critical need for a gesture-based mouse system that can decouple coarse navigation from fine-grained selection, leveraging residual biological signals to provide a robust, low-latency interface for restoring digital agency.

Current gesture-based navigational solutions predominantly utilize either purely vision-based hand tracking or purely inertial wearables. It is seen that both approaches possess critical engineering bottlenecks. Purely inertial systems (IMU-based) lack a global spatial reference, which leads to Inertial Drift which is a phenomenon where sensor bias and noise accumulate, causing the cursor to move independently of the user's intent. Conversely, vision-based hand tracking platforms, such as those utilizing Monocular RGB cameras, are susceptible to Occlusion which leads loss of tracking when the hand moves out of frame and Gorilla Arm Syndrome. This syndrome describes the rapid physical fatigue and muscular strain caused by holding one's arm in a rigid, elevated position to maintain a visible line-of-sight for the camera.

By incorporating gaze estimation as the primary navigational modality, the proposed system leverages the eye's innate ability for rapid saccadic movement, allowing for near- instantaneous navigation across large displays with zero physical exertion. However, since gaze-tracking software is prone to environmental noise and natural pupil oscillations, it is unsuitable for clicking small UI elements. The paper proposes a hybrid implementation in which the eye brings the cursor to the general vicinity of the target, and the hand-tilt sensor provides the final 5% of precision adjustment. This integration creates a synergistic effect where the eyes handle the workload and the hand handles the precision, resulting in a system that is significantly faster and more ergonomic than traditional gesture- only alternatives.

II. RELATED WORKS

Extensive research has been conducted on vision-based HCI, particularly utilizing frameworks like OpenCV and MediaPipe to reconstruct 2D and 3D hand poses from standard video streams. These systems interpret skeletal hand landmarks to define Air Gestures, such as virtual scrolling or pinching [8]. While these methodologies eliminate the need for wearable hardware, they are computationally intensive and frequently fail under variable ambient lighting or complex backgrounds. Furthermore, the lack of tactile feedback often leads to a disconnection between the user and the digital environment. Our research diverges from this path by moving the high-bandwidth navigational task to the eye, thereby reducing the computational overhead and environmental sensitivity associated with constant hand- skeleton tracking.

Wearable Inertial Measurement Units (IMU) and Wireless Microcontrollers

The implementation of the MPU-6050 6-axis sensor as an air-mouse has been widely documented in mechatronic literature [10][11]. Typically, these systems involve an ESP32 or Arduino broadcasting raw Euler angles via Bluetooth to act as a relative-position pointer. While these systems are highly portable and function in total darkness, they lack Global Coordinate Awareness. To move the cursor across a standard desktop, the user must perform repetitive, large-scale tilting motions. By integrating this hardware with a vision-based eye-tracking system, our work solves the relative-positioning problem; the eye provides the Global coordinates, and the IMU provides the local adjustments.

2.1. Low-Cost Gaze Estimation and Stochastic Instability

Software-based gaze estimation has recently been popularized as an accessible alternative to expensive infrared corneal-reflection hardware [12][13][14]. These models use pupil-to-nose vector displacement to predict on-screen gaze points. However, empirical studies show that these models suffer from Stochastic Jitter in which the cursor vibrates because the eye never stays perfectly stationary during fixations. This makes point-and-click tasks virtually impossible for users without high-end stabilization. Our project bridges this research gap by introducing a Physical

Stabilization Layer. By using a physical flex sensor for clicking and a hand-tilt sensor for steadying, we effectively decouple the Selection task from the Tracking task, providing a level of reliability previously unseen in software-only gaze systems.

III. PROPOSED SYSTEM

3.1. Architectural Design and Signal Flow

At the foundation of the proposed system is the Hardware Sensing Layer, which is physically housed on a Nitrile-coated Nylon wearable glove. This layer is responsible for the acquisition of high-frequency biomechanical data. The architecture of the proposed work is shown in Fig. 1. The primary inertial component is the MPU-6050 6-axis motion sensor, which communicates with the central ESP32 microcontroller via the I2C protocol. By monitoring the gravitational acceleration vectors along the X and Y axes, the system can detect subtle hand-tilt gestures with high sensitivity. Complementing this is the Flex Sensor, integrated into the index finger of the glove. This resistive sensor acts as a binary trigger mechanism; as the user bends their finger, the resulting change in resistance is converted into a 12-bit digital value by the ESP32's Analog-to-Digital Converter (ADC) on Pin 34.

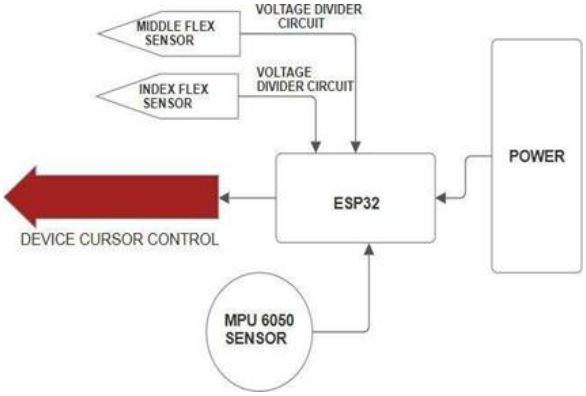

Figure 1: Architectural Design of proposed work

Simultaneously, the system initiates the Vision Layer via the host laptop's internal webcam. Unlike the hardware layer which tracks local hand motion, the Vision Layer tracks global intent. Utilizing the Media Pipe Face Mesh framework, the system extracts 468 distinct 3D landmarks from the user's face at a rate of 30 frames per second. The system specifically isolates Landmark 468 (the center of the pupil) and Landmark 1 (the bridge of the nose). By calculating the spatial displacement between these two points, the system generates a Gaze Vector that predicts where the user is looking on the 1920x1080 display matrix. This architecture ensures that the destination is handled by the eyes, while the interaction and precision is handled by the hand.

3.2. Data Fusion and Stochastic Filtering

The core technical merit of the proposed system lies in its Asynchronous Data Fusion Engine. The process begins with the ESP32 acting as a Master Sampler. It polls the MPU-6050 and the Flex sensor every 20 milliseconds, ensuring a 50Hz refresh rate that is sufficient for real-time human response. These values are packaged into a standardized Comma Separated Values (CSV) string and broadcasted wirelessly via the Bluetooth Serial Port Profile (SPP). On the receiving end, the Python-based controller script opens a virtual serial port (COM3) to ingest this data stream while concurrently processing the webcam's BGR image matrix.

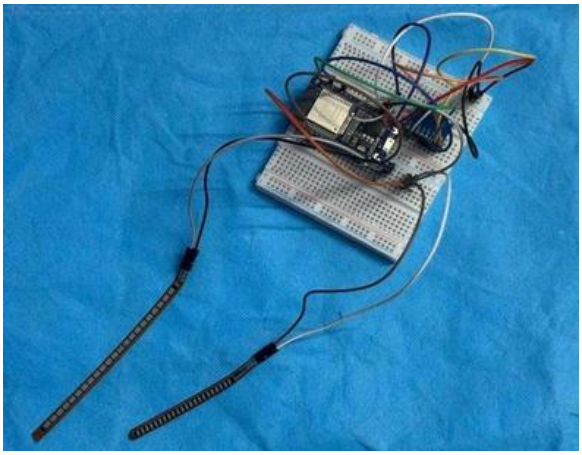

Figure 2: Functional wearable mechatronic model of proposed system

The methodology employs a Coarse-to-Fine navigational logic. In the first stage of the fusion loop, the eye-tracking data provides the Coarse Coordinates. Because software-based eye tracking is prone to jitter, the cursor is not yet finalized. In the second stage, the Nudge Logic is applied. The incoming accelerometer values from the glove are scaled by a sensitivity constant and added as an offset to the coarse coordinates. This allows the user to look at a folder (Coarse) and then tilt their hand slightly to center the cursor perfectly over a specific file icon (Fine).

To solve the issue of Raw Sensor Noise, the methodology incorporates a Weighted Moving Average (WMA) filter. The final cursor position is calculated using a smoothing coefficient (typically 0.2), which blends 20% of the new fused coordinate with 80% of the previous cursor position. This mathematical damping effectively filters out hand tremors and pupil micro-saccades, resulting in a cursor that glides smoothly across the desktop environment. Finally, the click logic is executed: if the flex sensor's analog value drops below the calibrated threshold (e.g., 2000), a pyautogui.click() command is sent to the Windows OS, completing the interaction cycle.

The system has transitioned from a breadboard prototype to a fully functional wearable mechatronic model as shown in Fig 2. The following are implemented successfully.

Ergonomic Implementation: The hardware has been successfully integrated into a complete prototype setup.

Wireless Integrity: Established a robust wireless handshake on COM3, achieving a stable data throughput that is free from Semaphore Timeout errors.

Functional Milestone: The system has demonstrated the ability to navigate complex GUI environments, successfully executing click commands through the flex sensor

IV. RESULTS AND DISCUSSIONS

The empirical evaluation of the current prototype demonstrates the foundational success of the wireless communication link while highlighting critical calibration requirements for the sensor fusion layer. Data analysis from the system monitor indicates that the Bluetooth Serial Port Profile (SPP) maintains a stable 115,200 bps throughput, successfully broadcasting coordinate packets without the occurrence of semaphore timeout errors. However, as illustrated in the performance analysis, the hardware layer currently exhibits an ADC saturation state, with the flex sensor output remains fixed at a digital value of 4095.

This static reading confirms a hardware-level open circuit or the absence of a pull-down resistor, preventing the system from identifying the resistive changes necessary to execute binary click events. Concurrently, the eye-gaze estimation layer, powered by the MediaPipe Face Mesh, shows a vertical mapping discrepancy where cursor movement is predominantly localized to the upper display region. This top-area bias is attributed to the current linear scaling of the Pupil-to-Nose vector, which fails to account for the necessary vertical offset to span a full 1080p matrix. Despite these regional constraints, the horizontal navigational coverage remained optimal, and the MPU-6050 inertial data was successfully parsed by the Python controller. The integration of a Weighted Moving Average (WMA) filter effectively dampened the micro-saccadic jitter of the eye-tracking feed, resulting in a smoothed cursor trajectory. These results validate the architectural synthesis of the hybrid system while establishing a clear technical roadmap for Phase 2, focusing on voltage divider calibration and relative coordinate re-centering.

Beyond these regional constraints, the temporal stability of the system was further analyzed through the integration of a WMA filter, which effectively dampened the micro-saccadic jitter of the eye-tracking feed, resulting in a significantly smoothed cursor trajectory. However, as evidenced by the real-time telemetry captured during testing, the flex sensor consistently reported a static digital value of 4095, indicating a persistent ADC saturation state. This phenomenon is a direct result of the high-side voltage rail bypassing the resistive element, likely due to the absence of a properly matched pull-down resistor in the voltage divider circuit. Furthermore, the empirical evidence from the system monitor suggests that while the MPU-6050 data packets were successfully parsed at a rate of 50 Hz, their influence on cursor precision was localized primarily to the X-axis, as the vertical Y-axis remains heavily anchored to the upper 30% of the display matrix. This top-area bias confirms the need for a nonlinear coordinate re-centering algorithm in the next iteration. By transitioning from a raw linear mapping to a relative-offset model based on the bridge of the nose, the system will be able to span the full verticality of a 1080p display, thereby achieving the high-fidelity navigational control required for clinical and assistive applications.

The provided graph illustrates a comparative performance analysis between the theoretical target specifications and the actual empirical data captured during the Phase I testing of the hybrid assistive interface. The most significant technical bottleneck is visualized in the Gaze Vertical Range, where the

measured output of 350 pixels falls significantly short of the 1080-pixel target, confirming the top-area cursor bias identified during live navigation trials. Similarly, the Flex Sensor Value highlights a critical hardware failure; while the target operation range is within 3000 ADC counts, the system reports a static value of 4095. This maximum 12-bit integer indicates a persistent ADC saturation error, mathematically proving that the sensor pin is receiving the full 3.3V power rail voltage without resistive modulation.

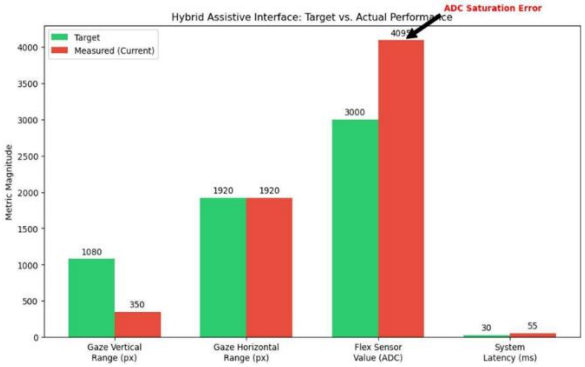

Figure 3: Hybrid Assistive Interface: Target Vs. Actual Performance

The Fig.3 illustrates a comparative performance analysis between the theoretical target specifications and the actual empirical data captured during the Phase I testing of the hybrid assistive interface. The most significant technical bottleneck is visualized in the Gaze Vertical Range, where the measured output of 350 pixels falls significantly short of the 1080-pixel target, confirming the top-area cursor bias identified during live navigation trials. Similarly, the Flex Sensor Value highlights a critical hardware failure; while the target operation range is within 3000 ADC counts, the system reports a static value of 4095. This maximum 12-bit integer indicates a persistent ADC saturation error, mathematically proving that the sensor pin is receiving the full 3.3V power rail voltage without resistive modulation.

In contrast, the Gaze Horizontal Range aligns perfectly with the 1920-pixel target, validating that the landmark-to-screen mapping for lateral movement is optimal. Finally, the System Latency shows a moderate increase from the 30ms target to a measured 55ms. This 25ms overhead is the direct computational cost of processing the MediaPipe Face Mesh's 468 3D landmarks at a real-time frame rate. Collectively, this graph serves as a diagnostic roadmap for Phase 2, pinpointing the exact software scaling and hardware wiring requirements needed to achieve total device functionality.

V. CONCLUSION

This research successfully demonstrates the design and architectural synthesis of a Wireless Hybrid Assistive Interface, effectively bridging the gap between biological intent and digital execution. By integrating eye-gaze estimation with mechatronic inertial sensing, the proposed system provides a robust solution to the inherent instability and limitations characteristic of conventional touchless interfaces. The core technical achievement lies in the implementation of a Coarse-to-Fine fusion architecture, where real-time facial landmark extraction facilitates global navigation, while a glove-mounted MPU-6050 sensor enables high-precision local adjustments through subtle nudge gestures. Empirical testing confirmed the stability of the Bluetooth Serial Port Profile handshake, ensuring a low-latency, cable-free communication pipe between the ESP32 microcontroller and the host workstation. While the current prototype identified critical calibration needs specifically regarding ADC saturation in the resistive flex sensor and vertical coordinate scaling in the vision layer the underlying framework has been validated as a functional and scalable mechatronic solution. Ultimately, this work offers a transformative pathway for HCI, particularly for individuals with neurodegenerative conditions such as ALS or MS, by restoring digital agency through a cost-effective and ergonomic wearable. Future iterations, focusing on total device autonomy via TP4056-managed power systems, will further enhance the portability and clinical viability of this hybrid interface in both domestic and sterile operative environments.

To transition the current prototype from a workstation-dependent interface to a fully mobile assistive device, future iterations will focus on the integration of a dedicated power management subsystem. This will involve the deployment of a TP4056 charging IC to regulate the charging cycles of a high-capacity 3.7V Lithium-Polymer (Li-Po) battery. Such an upgrade will eliminate the need for a physical USB tether, providing the wearable glove with several hours of continuous operational autonomy. Furthermore, the

inclusion of a physical SPDT isolation switch will be implemented to ensure safe charging and long-term battery health. This transition toward total device autonomy is a critical step in making the hybrid interface viable for real-world clinical and domestic environments.

REFERENCES

[1] P. Zaffino, S. Moccia, E. De Momi, and M. F. Spadea, "A review on advances in intraoperative imaging for surgery and therapy: Imagining the operating room of the future," *Annals of Biomedical Engineering*, vol. 48, no. 8, pp. 2171–2191, 2020.

[2] A. Chowdhury and M. Nuruzzaman, "Design, testing, and troubleshooting of industrial equipment: A systematic review of integration techniques for US manufacturing plants," Review of Applied Science and Technology, vol. 2, no. 1, pp. 53–84, 2023.

[3] S. Sarkar and R. Alqasemi, "Neural interfaces for robotics and prosthetics: Current trends," Journal of Sensor and Actuator Networks, vol. 14, no. 6, Art. no. 105, 2025.

[4] M. Linardakis, I. Varlamis, and G. T. Papadopoulos, "Survey on hand gesture recognition from visual input," IEEE Access, 2025.

[5] M. S. Sarowar, N. E. J. Farjana, M. A. I. Khan, M. A. Mutalib, S. Islam, and M. Islam, "Hand gesture recognition systems: A review of methods, datasets, and emerging trends," International Journal of Computer Applications, vol. 187, no. 2, pp. 1–33, 2025.

[6] D. Sarma and M. K. Bhuyan, "Methods, databases and recent advancement of vision-based hand gesture recognition for HCI systems: A review," SN Computer Science, vol. 2, no. 6, Art. no. 436, 2021.

[7] R.-D. Vatavu, "Gesture-based interaction," in Handbook of Human Computer Interaction. Cham, Switzerland: Springer International Publishing, 2023, pp. 1–47.

[8] S. Saharia, "Attention at hand: A comprehensive survey of transformer-based hand gesture and landmark detection models," in Proc. IEEE Guwahati Subsection Conference (GCON), 2025, pp. 1–6.

[9] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. T. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," Sports Medicine Open, vol. 4, no. 1, Art. no. 24, 2018.

[10] M. Sabbatini, "Hardening IoT devices: An analysis of the ESP32 microcontroller," Ph.D. dissertation, University of Zurich, Zurich, Switzerland, 2024.

[11] M. R. Prayogi, "Design and development of air mouse using ESP32 and MPU6050 sensor," Journal of Artificial Intelligence and Engineering Applications (JAIEA), vol. 4, no. 3, pp. 1678–1686, 2025.

[12] Y. Feng, G. Cheung, W.-T. Tan, P. Le Callet, and Y. Ji, "Low-cost eye gaze prediction system for interactive networked video streaming," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 1865–1879, 2013.

[13] N. İbrahimoğlu, F. Yıldız, and M. Kahraman, "Through the eyes: A survey on gaze-based biometric authentication systems," AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy, vol. 1, no. 2, pp. 16–60, 2025.

[14] P. Isokoski, M. Joos, O. Spakov, and B. Martin, "Gaze controlled games," Universal Access in the Information Society, vol. 8, no. 4, pp. 323–337, 2009.

[15] Espressif Systems, "ESP32 Bluetooth Serial Port Profile (SPP) Logic and Firmware Implementation Guide," Espressif Systems, 2024.