Touchless HCI for Media Control Using Hand Gestures

Pushpal Bhar, Subhojit Khatua, Arghya Pratim Biswas
Department of Electronics & Telecommunication Engineering
Jadavpur University
Mentor: Prof Sheli Sinha Chaudhury,Dept of ETCE,Jadavpur University

Project Duration: Jan-Feb,2026

ABSTRACT

Natural User Interfaces (NUI) have emerged as a pivotal frontier in Human-Computer Interaction (HCI). This paper proposes a high-efficiency, real-time hand gesture recognition system optimized for edge-computing environments, specifically the **Raspberry Pi 5**. Unlike traditional deep-learning classifiers that suffer from high computational latency, our approach utilizes a **Hierarchical Edge-Inference** architecture. The system integrates MediaPipe's BlazePalm topology with a deterministic **Hierarchical Finite State Machine (HFSM)** to map spatial coordinates to media control commands. Experimental results demonstrate a 95%+ recognition accuracy at distances up to 240 cm(in Lapcam HD 720P (LWC-042)) while maintaining a consistent throughput of 20+ FPS.

Index Terms — Hand Gesture Recognition, Edge Computing, Raspberry Pi 5, MediaPipe, HFSM, Human-Computer Interaction (HCI).

INTRODUCTION

The evolution of computing has reached a stage where the bottleneck of interaction is no longer processing power, but the physical interface between the human and the machine. Hand gesture recognition offers a "Natural User Interface" (NUI) that mimics human-to-human non-verbal communication, removing the constraints of traditional tactile peripherals. While vision-based systems offer a pervasive solution, they face significant challenges regarding environmental variables and the high computational cost of real-time processing on edge devices like the **Raspberry Pi 5**.

Historically, gesture recognition relied on contour-based methods such as **Convex Hull** and **Convexity Defects** to count fingers and define hand shapes. However, these methods are highly sensitive to pixel-level noise, background clutter, and fluctuating light, which often deforms the "hull" and leads to spurious activations. Similarly, probabilistic classifiers like **Support Vector Machines (SVM)** and **K-Nearest Neighbor (KNN)** introduce a "Black-Box" unpredictability and secondary inference latency that hinders real-time responsiveness.

To resolve these limitations, our work introduces a **4-Layer Hierarchical Methodology** that moves away from pixel-contour analysis toward **Skeletal Topology Mapping**. We leverage the 21-point landmark regression of MediaPipe but decouple the feature extraction from the execution logic. By replacing probabilistic ML classifiers with a deterministic **Hierarchical Finite State Machine (HFSM)**, we eliminate inference overhead and enable "Velocity-Adaptive" control—a feat unattainable by standard SVMs.

This research focuses on three primary innovations:

1. **Skeletal Robustness:** Utilizing 21-point coordinate regression over traditional Convex Hull analysis to ensure immunity to background noise.
2. **Deterministic Control:** Implementing an HFSM to provide near-zero latency compared to SVM-based classification.
3. **Temporal Efficiency:** Utilizing State-Triggered Buffer Flushing to eliminate transition lag between gesture states.

The resulting system demonstrates that a sophisticated, distance-invariant NUI can be achieved on low-power hardware through rigorous architectural optimization and the strategic elimination of redundant machine learning layers.

SYSTEM ARCHITECTURE

The system architecture utilizes a **4-Layer Decoupled Pipeline**. The Vision Layer (OpenCV) handles frame acquisition, while the Inference Layer (MediaPipe) regresses the 21-point topology. The Logic Layer (HFSM) translates these coordinates into states, and the Execution Layer interacts with the OS-level media APIs.

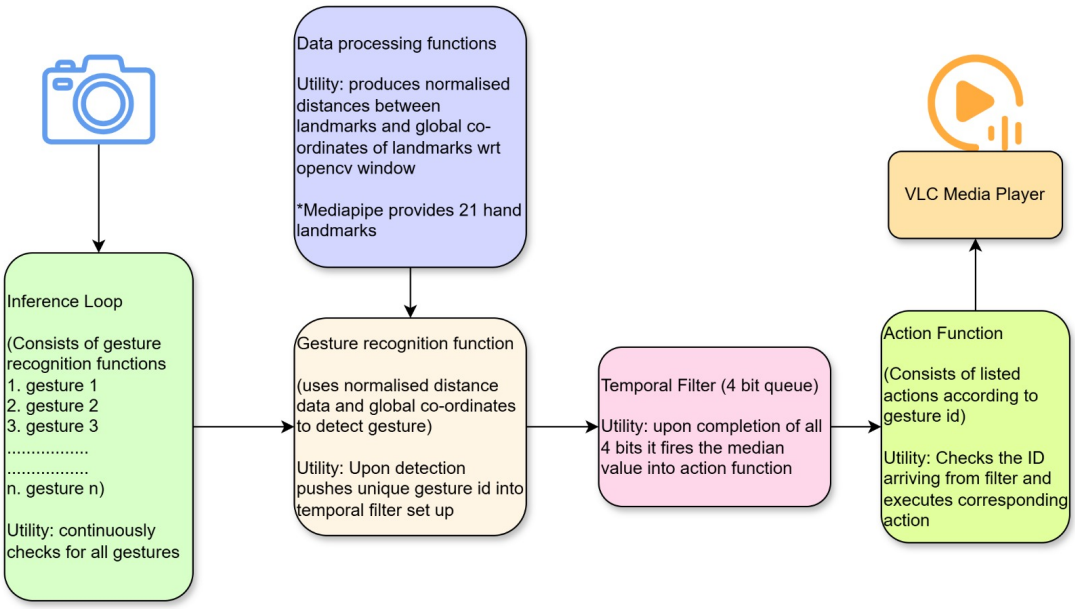

Figure 1: System Architecture

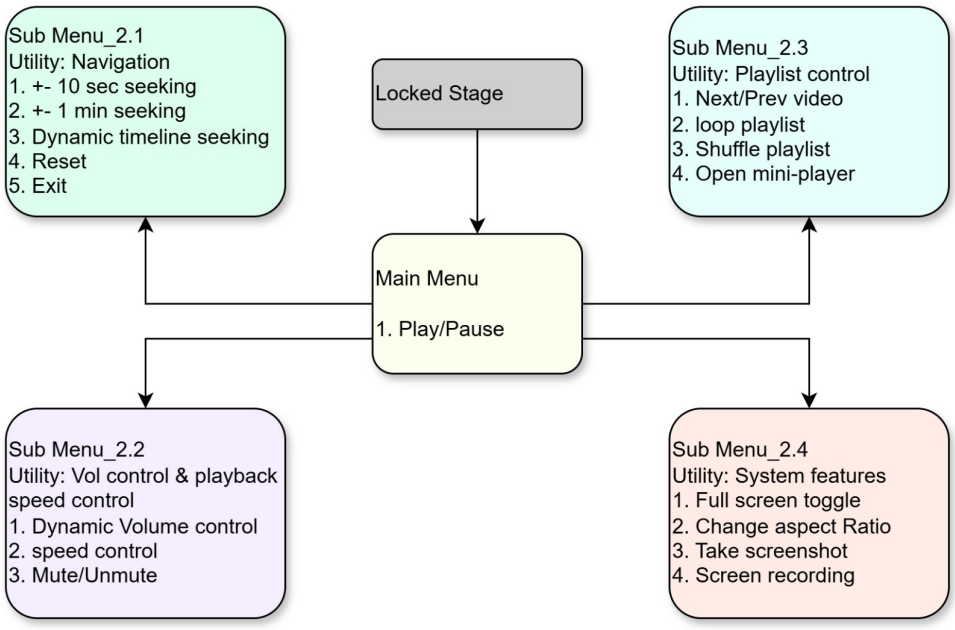

Figure 2: Navigation And Feature Chart

TECHNICAL METHODOLOGY: HIERARCHICAL EDGE-INFERENCE & DETERMINISTIC CONTROL

The system architecture is engineered to resolve the "Computational Bottleneck" typically found when running real-time computer vision on resource-constrained edge devices like the **Raspberry Pi 5**. The methodology is structured into four high-order layers: Signal Conditioning, Feature Extraction, State-Machine Logic, and Asynchronous Execution.

LAYER 1: SIGNAL CONDITIONING & PREPROCESSING

Before landmark regression, the raw video stream undergoes a multi-stage Digital Signal Processing (DSP) pipeline to optimize feature visibility and reduce computational entropy.

1.1 Chromatic Transformation & MediaPipe Landmark Regression

OpenCV captures frames in BGR format; however, to align with the **MediaPipe Hands** inference engine, we perform a color-space conversion using the **Weighted Luminance Method** (ITU-R BT.601 standard). The conversion is governed by the following formula:

$$Y = 0.2989R + 0.5870G + 0.1140B \tag{1}$$

By prioritizing the luminance channel, the system maximizes the contrast of the hand's contours against background noise. These optimized RGB frames are then processed by the **BlazePalm Single-Shot Detector** to isolate the Region of Interest (ROI), which allows for the precise regression of a **21-point 3D hand landmark topology**. This skeletal mapping identifies specific anatomical joints (MCP, PIP, and DIP joints), providing the raw spatial data required for gesture classification.

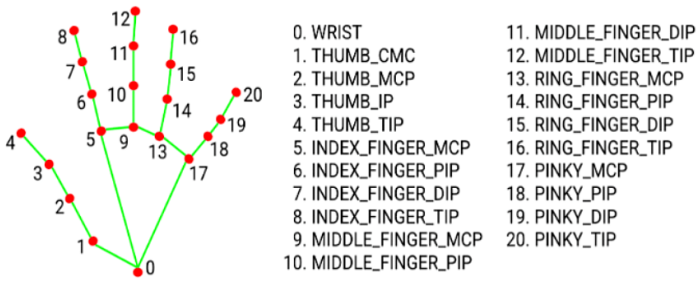

Figure 3: Hand Landmarks In Mediapipe

1.2 Spatial Optimization & Signal Smoothing

- **Downsampling:** Frames are resized to 360×240 to minimize CPU pixel-processing overhead.
- **EMA Filtering:** To mitigate sensor noise and tremors, an **Exponential Moving Average** acts as a Hysteresis Stabilizer:
$$E_t = \alpha \cdot R_t + (1 - \alpha) \cdot E_{t-1} \tag{2}$$
where $\alpha_{norm} = 0.25$ (Shape Stability) and $\alpha_{global} = 0.50$ (Motion Responsiveness).
- **Cold-Start Protection:** An initialization check performs a "Hard-Snap" to the hand's location if the state is null, bypassing recursive "slide-in" lag.

LAYER 2: FEATURE EXTRACTION & SCALE-INVARIANCE

The system transitions from pixel-space to a **Unit-Vector Space**, enabling distance-invariant recognition.

2.1 Centroid-Logic Normalization

We define a Differential Control Reference using the Middle MCP (Landmark 9) as the origin $(0, 0)$. The Scaling Factor (σ) is derived from the "Palm Length":

$$\sigma = \sqrt{(x_9 - x_0)^2 + (y_9 - y_0)^2} \tag{3}$$

Each landmark i is transformed into a Scale-Invariant Coordinate (N_i):

$$N_i = \frac{\sqrt{(x_i - x_9)^2 + (y_i - y_9)^2}}{\sigma} \tag{4}$$

RESTRICTED ACCESS

Full Technical Documentation Restricted

This document contains proprietary architectural details,
experimental source code, and comprehensive benchmark datasets.

To request the full technical report (PDF):

Contact: `IOT Applications Club,Jadavpur University`

Subject: Request for Full Technical Report