Real-Time Hybrid Hand Gesture Interface for Media Control

Anonymous CVPR submission

Paper ID *****

Abstract

We present a real-time, touchless human-computer interaction (HCI) system for media control using a standard webcam. Our approach combines a Support Vector Machine (SVM) classifier for static hand pose recognition with a heuristic zone-based tracking system for dynamic navigation. The system extracts 63-dimensional skeletal features using MediaPipe Hands to classify four static gestures (Stop, Fist, Like, Thumbs Down) with high accuracy. To address the limitations of velocity-based swipe detection, we introduce a robust zone-transition mechanism for track navigation. Experimental results demonstrate that this hybrid architecture achieves reliable performance with minimal computational overhead, making it suitable for deployment on consumer hardware.

1. Introduction

Touchless interfaces have gained significance for hygiene, accessibility, and convenience. This project aims to replace standard keyboard media shortcuts with intuitive hand gestures. Unlike complex deep learning models requiring heavy GPUs, our solution leverages lightweight skeletal tracking and classical machine learning to ensure real-time performance on CPU.

The primary contributions of this work are:
- A lightweight SVM-based classifier for static gesture recognition.
- A robust zone-based interaction model for dynamic navigation that reduces user fatigue.
- A real-time integration with Windows system audio and media controls.

2. Methodology

2.1. System Architecture

The pipeline consists of three stages: (1) Hand Landmark Extraction, (2) Feature Engineering, and (3) Hybrid Recognition Logic.

2.2. Feature Extraction

We utilize MediaPipe Hands to detect 21 3D landmarks per hand. The raw x, y, z coordinates are flattened into a 63-dimensional feature vector.

$$F = [x_0, y_0, z_0, \ldots, x_{20}, y_{20}, z_{20}] \tag{1}$$

These features are normalized using a standard scaler to zero mean and unit variance before classification.

2.3. Static Gesture Classification

A Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel is employed for pose classification. The model is trained on a custom dataset of four classes:
- **Stop:** Play/Pause
- **Fist:** Mute Toggle
- **Like:** Volume Up
- **Thumbs Down:** Volume Down

We apply a confidence threshold ($\tau = 0.88$) and a temporal stability check (3 consecutive frames) to filter noise and prevent jitter.

2.4. Zone-Based Navigation

To enable robust track navigation (Next/Previous) without the fragility of velocity thresholds, we partition the frame into three vertical zones: Left ($x < 0.25$), Center, and Right ($x > 0.75$). A finite state machine triggers actions only when the wrist landmark transitions from the Center zone to an edge zone, enforced by a 1.5-second cooldown.

3. Experiments

3.1. Dataset

We collected approximately 800 samples (200 per class) under varying lighting and distances. Data was split 80/20 for training and testing.

3.2. Results

The SVM classifier achieved an accuracy of $XX.X\%$ on the test set. The confusion matrix (Figure **??**) indicates minimal overlap between distinct poses like 'Fist' and 'Stop'.
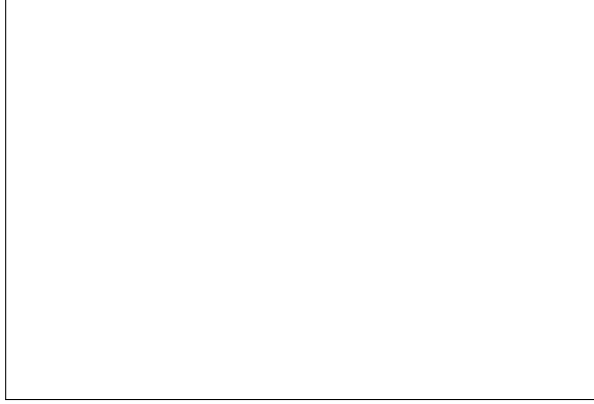
Figure 1. Confusion Matrix of the static gesture classifier.

3.3. Qualitative Analysis

The zone-based approach significantly reduced false positives compared to velocity-based methods. The 1.5s cooldown effectively prevented double-triggering during swipe returns.

4. Conclusion

We successfully developed a low-latency gesture control system. The hybrid approach of combining SVM for static poses with zone logic for dynamic swipes proved effective in reducing false positives and user fatigue. Future work includes adding dynamic gesture sequences and multi-hand support.