Vision-Based Multimodal Teleoperation Interface for Firefighting Robots

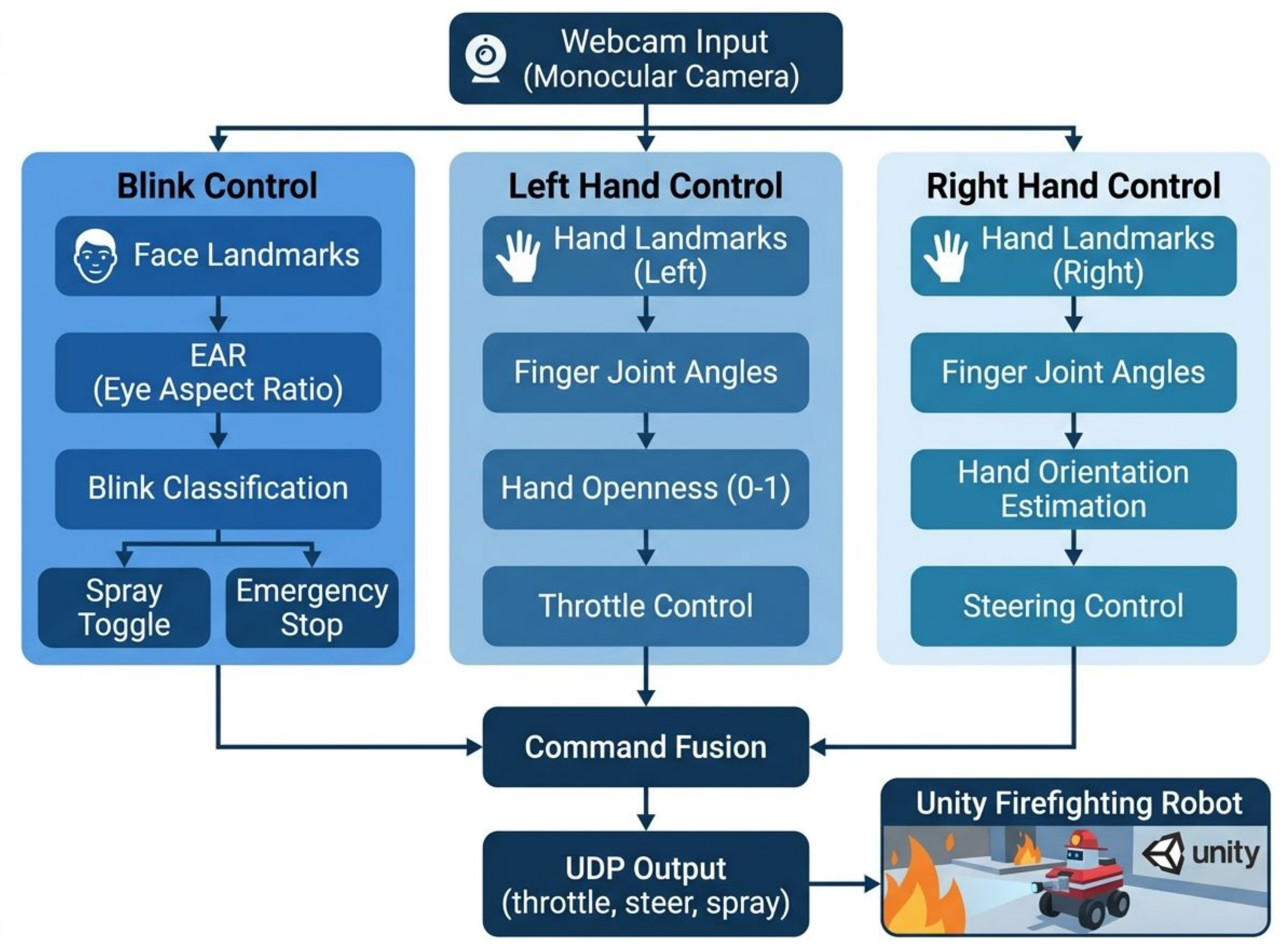
Using Hand Gestures and Eye Blinks

HeeJune Song, Jian Baek

Introduction

Robotic teleoperation is widely used in hazardous environments such as firefighting and disaster response, but conventional handheld controllers can be difficult to operate in high-risk situations. Vision-based interfaces provide a natural alternative, yet gesture-only systems often struggle to support both continuous motion control and discrete command triggering. To address this limitation, we propose a **vision-based multimodal teleoperation interface** using **hand gestures and eye blinks**. In the proposed system, left-hand finger joint angles and palm direction generate throttle commands, right-hand orientation generates steering commands, and blink patterns trigger auxiliary commands. The resulting signals are transmitted via UDP to a Unity-based firefighting robot simulator.

System Overview



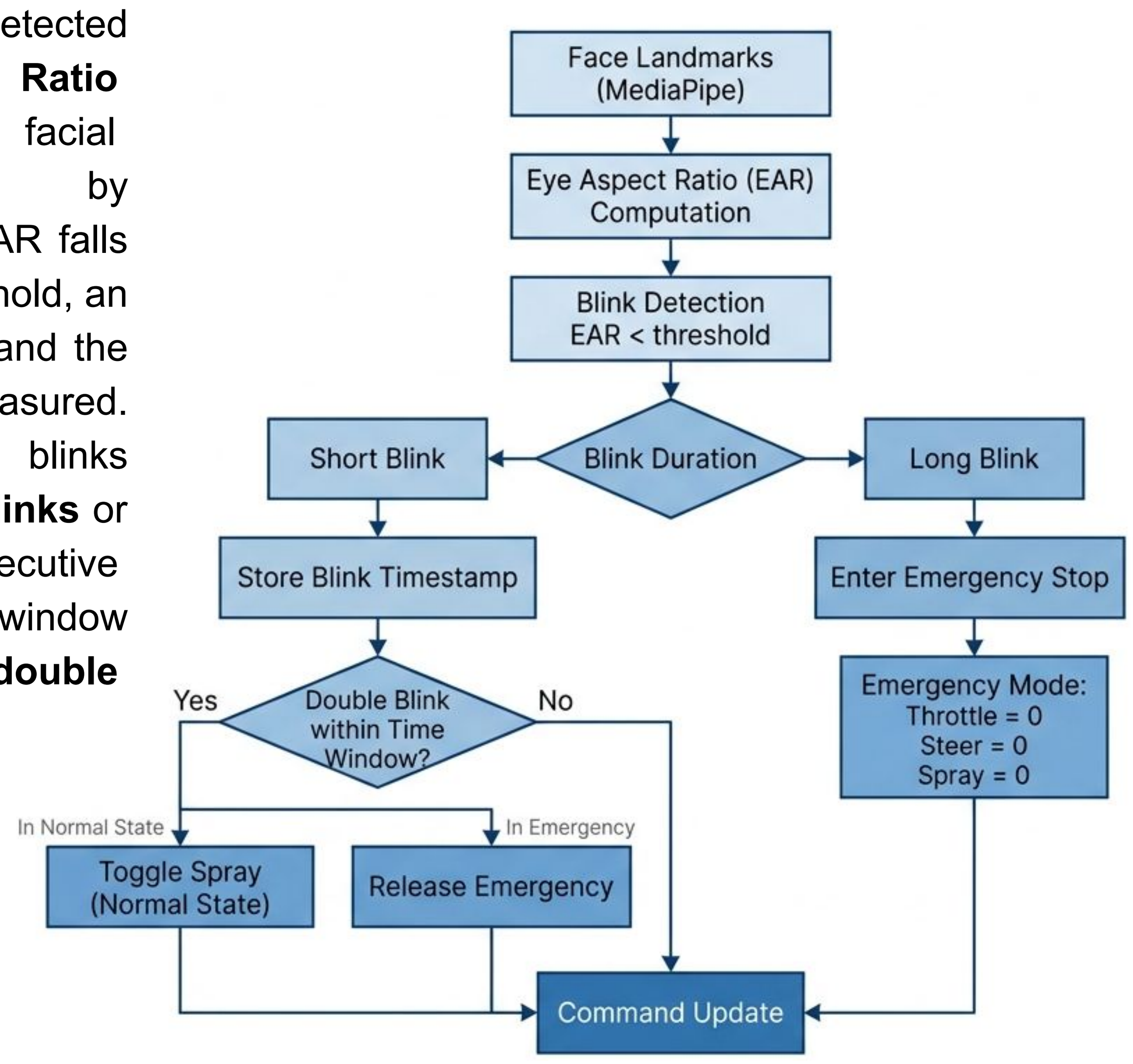
The proposed system consists of **perception, command generation, and robot control** modules. A monocular webcam captures the user's hand gestures and facial expressions, and MediaPipe extracts hand and face landmarks in real time.
From these landmarks, the system computes **finger joint angles**, **palm orientation**, and **eye aspect ratio (EAR)** to generate throttle, steering, and blink-triggered auxiliary commands. The final control tuple **(throttle, steer, spray)** is transmitted via UDP to a Unity-based.

Experimental Setup

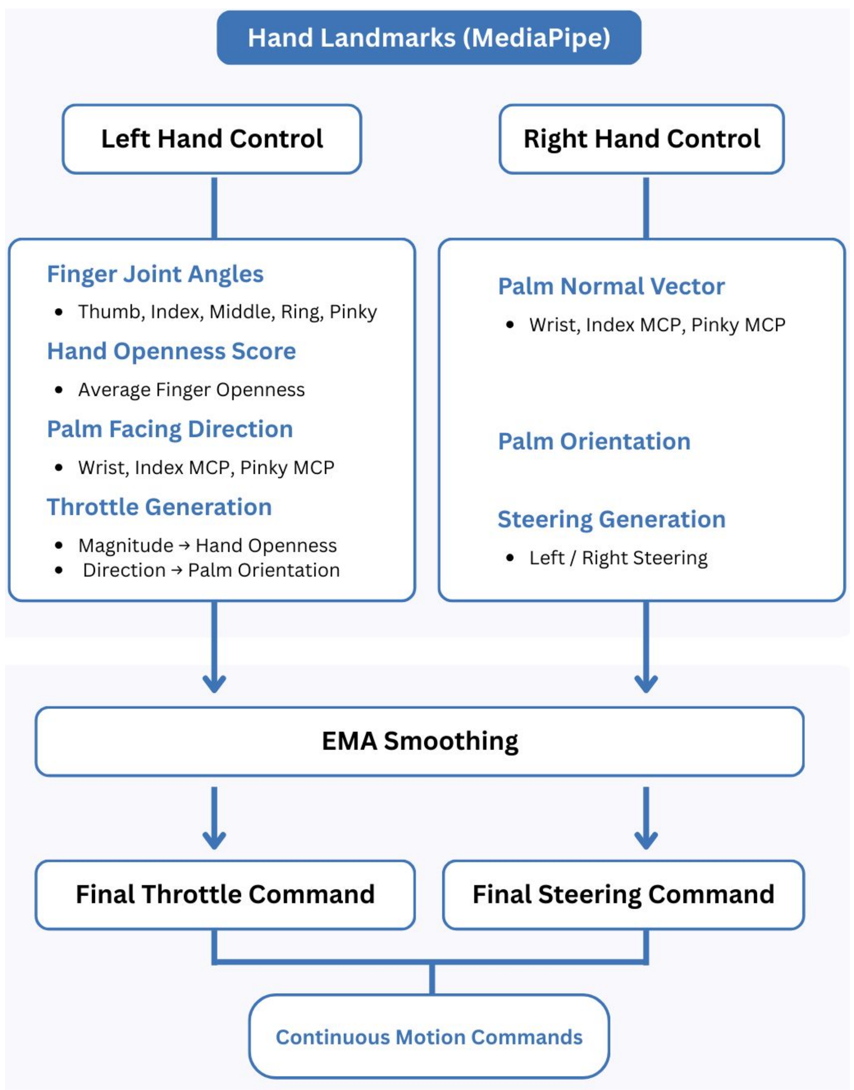
Monocular webcam	MediaPipe-based hand and face landmark tracking
Python + OpenCV control pipeline	
Unity-based firefighting robot simulation	UDP communication

Control Strategy

Eye blink events are detected using the **Eye Aspect Ratio (EAR)** computed from facial landmarks extracted by MediaPipe. When the EAR falls below a predefined threshold, an eye closure is detected and the blink duration is measured. Based on the duration, blinks are classified as **short blinks** or **long blinks**. Two consecutive short blinks within a time window are interpreted as a **double blink command**.

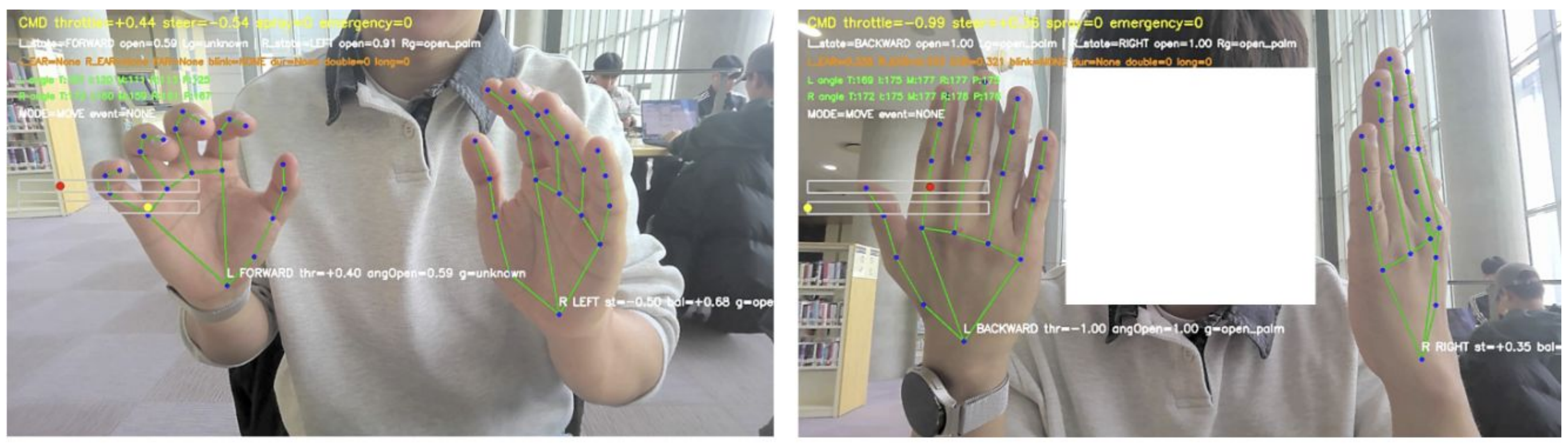


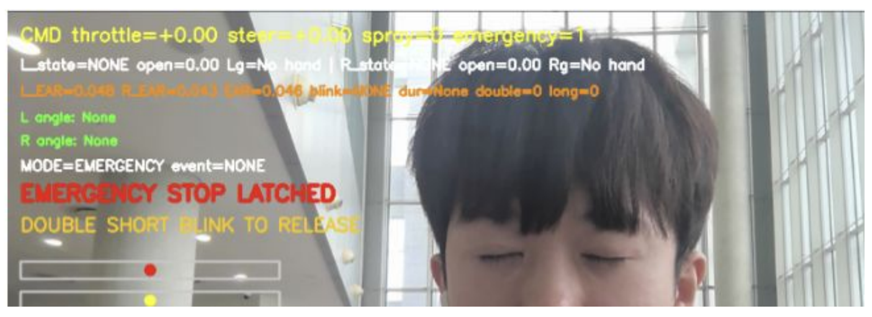
In normal operation, a double short blink toggles the **water spray function**. A long blink triggers an **emergency stop**, which immediately sets throttle and steering to zero and disables spray output. While the system is in the emergency state, a double short blink releases the emergency lock and restores normal robot control.



Hand gesture recognition was performed using the **21 hand landmark coordinates** extracted by **MediaPipe Hands**. Finger joint angles, finger openness, palm orientation, and hand tilt were used to generate continuous steering and throttle commands.
The left hand controlled throttle, where acceleration magnitude depended on average finger openness and forward/backward motion was determined by hand tilt relative to the camera. **The right hand controlled steering**, where left/right turning was identified using palm edge orientation and the palm normal vector. This geometric feature–based control **enabled intuitive and smooth vehicle operation** through natural hand movements.

Results





The **Vision-based multimodal interface** successfully enabled **simultaneous recognition of hand gestures and eye states** using a single RGB camera.
Based on the 21-point hand landmarks, **finger openness, palm orientation, and joint angles** were used to generate **continuous steering and throttle commands.**
The EAR module reliably distinguished **short blinks and long eye-closures**, which accurately triggered **spray toggle and emergency stop events**. The recognized control signals were transmitted to the Unity simulation via **UDP in real time**, resulting in stable vehicle motion and water spraying behavior with negligible latency.
Experimental observations indicate that users could **intuitively control the vehicle without physical controllers**, demonstrating the **feasibility of contactless multimodal robot interaction**.

Contributions

This study presents a **vision-based multimodal teleoperation interface** for robot control that combines hand gestures and eye blinks. Continuous motion commands are generated using **geometry-based hand control**, where finger joint angles and palm orientation derived from hand landmarks are mapped to throttle and steering signals. In addition, **EAR-based blink detection** is utilized for discrete command triggering such as spray activation and emergency control. The proposed system enables **real-time robot teleoperation via UDP communication** with a Unity-based firefighting robot simulator.

Discussion

The results demonstrate that integrating **hand gestures with blink-based commands** provides an intuitive multimodal interface for robot teleoperation. By separating **continuous motion control** from **discrete command triggering**, the proposed method improves usability and enables stable robot operation using simple visual inputs. Furthermore, the use of **geometric features extracted from hand landmarks** allows smooth throttle and steering control without physical controllers.
Future work will focus on enhancing robustness under challenging conditions such as **lighting variations, occlusion, and unstable camera viewpoints.**

Conclusion

This work proposes a **vision-based multimodal control interface** for firefighting robots using hand gesture recognition and blink detection. Finger joint angle–based hand openness and palm orientation are mapped to continuous motion commands, while blink patterns are used to trigger auxiliary functions such as spray activation and emergency release. Experimental validation shows that the system enables **real-time robot control using only a monocular webcam**, suggesting that multimodal vision interfaces can provide an effective and intuitive solution for teleoperation in hazardous environments.