Modifly

A Multimodal Interaction Simulator for Drone Operation via Gesture Recognition and Voice Commands

Master's Degree in Computer Science

Pasquale Celani (1839634)

Academic Year 2024/2025

SAPIENZA
UNIVERSITÀ DI ROMA

Introduction

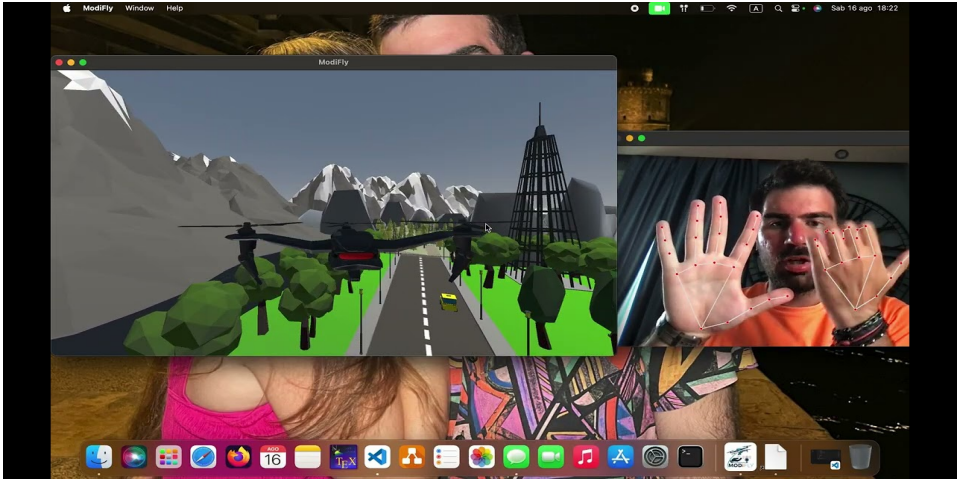
Spider Web operation

Motivation

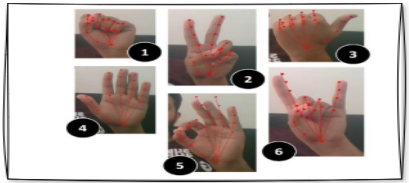
Traditional drone controls, similar to those in consumer electronics, have significant limitations that must be addressed for future applications:

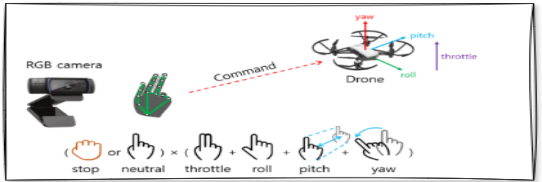
- Higher production costs;
- Physical Fragility;
- Risk of Loss;
- **Unnatural Interaction**.

My contribution



Related works



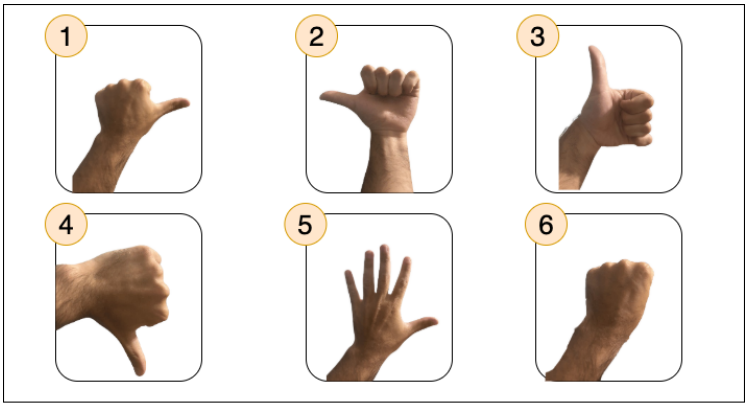


Interaction design & Modality roles

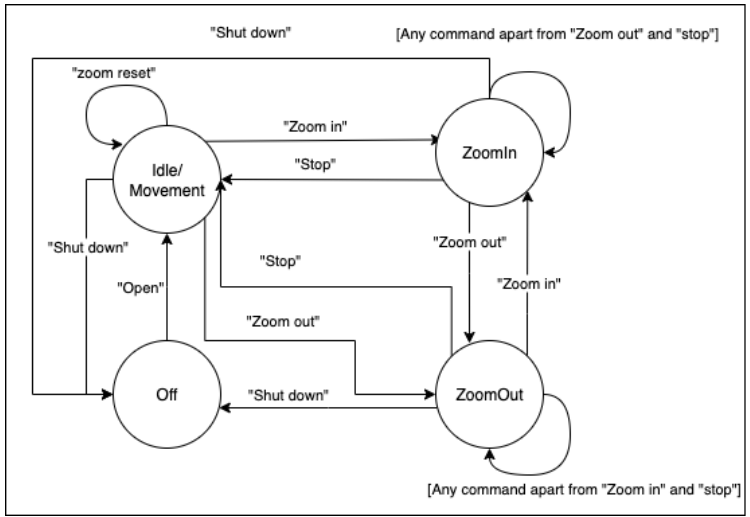
Drawing on the work of Bernsen and Dybkjær [2009] and the D'Ulizia [2009] taxonomy, ModiFly is a multimodal system that uses gestures (Graphics-Light-Vision) for continuous drone movements, and voice commands (Acoustics-Sound-Hearing) for discrete setup and secondary commands, as its primary modalities.
This interaction model is based on:

- **Specialization:** Each modality has an exclusive, assigned function (e.g., gestures for movement, voice for setup);
- **Concurrency:** Both modalities operate in parallel, enabling simultaneous actions without a required merge point.

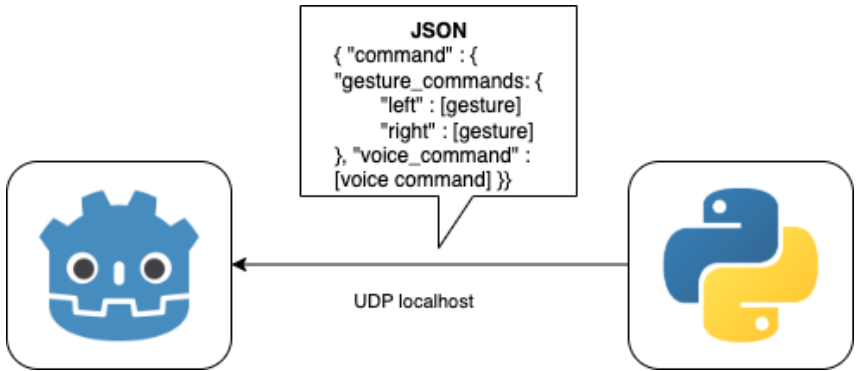
Gesture control



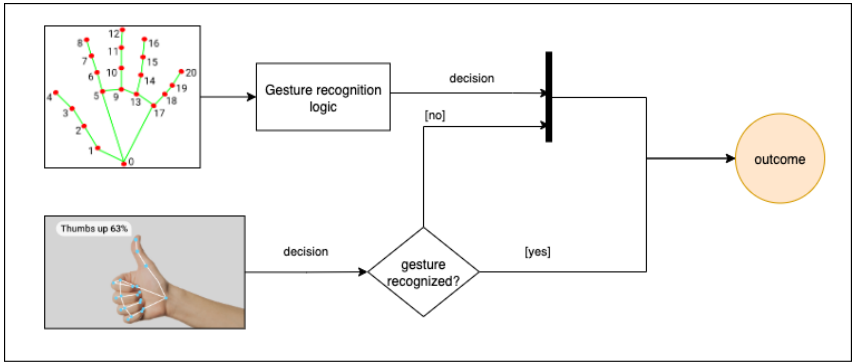
Voice control



System architecture



Fusion strategy for gesture recognition



Gesture Recognition implementation

These gestures are recognized using the MediaPipe library, which is employed in two distinct ways:

- **Hand landmark recognition:** The hand landmark recognition is based on the following formulas:
  - **Thumb left:** $lm[4].x < lm[3].x < lm[2].x \ \wedge \ \forall k \in \{5, 9, 13, 17\}\ lm[k+3].y > lm[k].y$;
  - **Thumb right:** $lm[4].x > lm[3].x > lm[2].x \wedge \forall k \in \{5, 9, 13, 17\}\ lm[k+3].y > lm[k].y$.
- **Gesture classifier**: The MediaPipe library also includes a built-in classifier capable of recognizing a set of predefined static gestures, such as ”thumb up”, ”thumb down”, ”open palm” and etc.
  - In our case all the gestures apart from 1 (Thumb left) and 2 (Thumb right) use the classifier to recognize the gestures.

Voice Recognition implementation

- **API Utilization:** Uses Google's Web Speech API for processing to offload the computational burden.
  - Requires an internet connection.

- **Continuous Listening:** A dedicated thread runs in the background, continuously listening for voice commands without blocking the main application;
- **Command Recognition:** The transcribed text is matched against a predefined list of commands, such as "open," "shut down," "zoom in," and "zoom out."

Conclusion & Future work

- **Novel interface design:** it has been successfully developed and validated a groundbreaking HDI, replacing traditional joystick controls with an intuitive interaction based on gesture and voice;
- **Cost effective:** It eliminates the need for expensive, dedicated hardware. The only required components are standard, widely available devices like a computer or mobile phone;
- **Highly Optimized:** The remarkable efficiency of the system was proven by its successful development on a 2019 MacBook Air, showcasing its minimal computational footprint;
- **Real-World Validation:** The next critical step is to transition from simulation to reality by testing the system with a physical drone in a real operational context.

Modifly

Ready to see it in action?
Let's dive into the demo!