State of the Art in Gesture-Based, Vision-Based, and Learning-Based Drone Control

Mohammed-salih Diyari*, Chaabeni Ilyes*
* Master 2 E3A – Smart Aerospace & Autonomous Systems
Université d'Évry-Val-d'Essonne
†Project Supervisor: Naima AitOufroukh
Email: naima.aitoufroukh@univ-evry.fr

Contents

I. Introduction and Motivation

Unmanned Aerial Vehicles (UAVs), commonly referred to as drones, are increasingly used in inspection, surveillance, education, and research. Despite advances in autonomy, most consumer drones are still controlled using handheld remote controllers, which require training and limit accessibility.

Human–robot interaction research has long explored alternative interaction modalities, including vision-based hand gestures, to enable more intuitive control. Wachs et al. [1] highlight that gestures provide a natural and expressive communication channel for controlling machines, particularly in robotic systems.

Recent progress in computer vision and machine learning has further enabled gesture-based and learning-based drone control. This report reviews the current state of the art and positions the present project within this research landscape.

II. Current Project Overview and Objectives

The current project implements a gesture-controlled drone using a monocular camera. Static hand gestures are detected frame by frame and mapped directly to drone commands such as takeoff, landing, and directional movement.

While functional, the system has several limitations:

- Gesture recognition is static and frame-based.
- Users must hold poses, reducing natural interaction.
- No temporal understanding of gesture motion exists.
- No autonomous follow-me or target tracking mode is implemented.
- Control and visualization are limited to local OpenCV windows.

The objectives of this project are therefore to:

- Introduce dynamic gesture recognition using LSTM models.
- Add a vision-based follow-me mode based on user detection.
- Develop a web-based dashboard for control and monitoring.

These objectives are motivated by limitations identified in the state of the art.

III. Gesture-Based Drone Control

Gesture-based drone control replaces traditional remote controllers with vision-based hand gestures. Taylor et al. [2] present a recent and representative system using Google MediaPipe Hands to extract 21 hand landmarks from RGB images. Static gestures are mapped to drone commands in real time.

This work demonstrates that gesture-based piloting is feasible, accessible, and suitable for novice users.

What this work provides:

- Lightweight hand landmark extraction.
- Safe gesture-to-command mapping.

Limitations:

- Gestures are static and frame-dependent.
- Performance is sensitive to lighting and hand orientation.

These limitations are consistent with those identified in earlier surveys on gesture-based interaction [1]. This work represents the current state of the art for static gesture-based drone control and serves as the baseline for the current repository.

IV. Dynamic Gestures Using LSTM

Static gestures restrict expressiveness and robustness. Dynamic gesture recognition addresses this by modeling motion across time.

Ikram and Liu [3] propose a skeleton-based dynamic hand gesture recognition system using Long Short-Term Memory (LSTM) networks. Their system processes sequences of joint positions, enabling recognition of motion-based gestures such as swipes and rotations.

Similarly, Sundar et al. [4] demonstrate that pose estimation combined with LSTM networks enables real-time recognition of human actions with low computational cost.

Relevance to this project:

- Enables natural motion-based gestures.
- Improves robustness compared to static gestures.

As emphasized by Wachs et al. [1], static gesture systems suffer from ambiguity and user fatigue, directly motivating the transition toward temporal gesture modeling in this project.

V. Vision-Based Follow-Me Systems

Follow-me systems allow drones to autonomously track a user using visual feedback. The drone adjusts its yaw and pitch to keep the target centered in the camera frame.

Deep-learning-based approaches often rely on object detectors such as YOLO. The Drone-YOLO repository [5] demonstrates real-time object detection for UAV applications using YOLO-based architectures.

Insights:

- Vision-based tracking enables autonomous behaviors.
- Detection pipelines are effective but computationally heavy.

Due to hardware constraints, this project adopts lighter alternatives such as face detection combined with classical trackers for follow-me functionality.

VI. Web-Based Control and Monitoring

Most gesture-controlled drone systems rely on local OpenCV windows. Web-based dashboards improve accessibility by enabling control and monitoring from any device on a local network.

A web interface allows:

- Live video streaming.
- Mode switching between gesture and follow-me control.
- Visualization of system state and confidence.

Integrating such an interface with gesture-based and autonomous control remains largely unexplored, motivating this extension.

VII. System Evolution by the State of the Art

Based on the reviewed literature, the proposed system integrates gesture recognition, temporal modeling, autonomy, and web control.

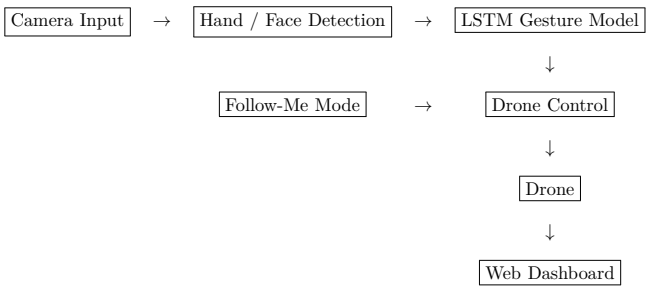

Fig. 1. High-level system evolution based on the state of the art

Figure 1 presents the high-level evolution of the proposed drone control system inspired by the state of the art. Visual input from a camera is processed for hand and face detection, enabling both gesture recognition and user tracking. An LSTM-based model analyzes gesture sequences to support dynamic, motion-based commands, while a follow-me mode provides autonomous user tracking. A central control module integrates these inputs to generate safe flight commands for the drone and to communicate system status to a web-based dashboard for monitoring and interaction.

VIII. Discussion and Research Gap

The literature shows that gesture control, temporal modeling, vision-based tracking, and web interfaces are well studied individually. However, existing systems rarely integrate all these components into a single lightweight framework.

This project addresses this gap by combining:

- Dynamic gesture recognition using LSTM,
- Vision-based follow-me autonomy,
- Web-based control and monitoring.

IX. Conclusion

This state-of-the-art review analyzed recent research relevant to gesture-based, vision-based, and learning-based drone control. By building upon current methods and addressing their limitations, the proposed system aims to deliver a more natural, robust, and accessible drone control platform.

References

[1] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan, "Vision-Based Hand-Gesture Applications," *Communications of the ACM*, vol. 54, no. 2, pp. 60–71, 2011.

[2] B. Taylor et al., "Enhancing Drone Navigation and Control: Gesture-Based Piloting, Obstacle Avoidance, and 3D Trajectory Mapping," *Applied Sciences*, vol. 15, no. 13, 2025.

[3] M. Ikram and Y. Liu, "Skeleton-Based Dynamic Hand Gesture Recognition Using LSTM and CNN," in *Proc. IPMV*, 2020.

[4] K. C. M. Sundar et al., "Real-Time Human Action Recognition Using Pose Estimation and LSTM," *Sensors*, 2024.

[5] TKN TU Berlin, "Tello Drone Control with Flask and YOLOv8," GitHub repository. Available: https://github.com/tkn-tub/drone_Yolo