International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 10 Issue 4, Jul-Aug 2026 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Human AI Collaboration: System Design for Real-time Recognition and Response to Traffic Police Gestures for Autonomous Vehicles

Qi Yuxuan, Zhou Hanyu, Feng Linshu, Liu Tonghan, Wu Yi

Beijing Wuzi University, Beijing, China

ABSTRACT

With the rapid advancement of artificial intelligence technologies, autonomous vehicles are increasingly being deployed in complex open road scenarios. However, existing autonomous driving systems generally lack effective recognition and response capabilities when confronted with onsite traffic police gesture commands, which constitutes a significant safety hazard for the deployment of autonomous driving technology. Focusing on the core pipeline of "visual perception-instruction parsing-vehicle control," this paper proposes a real-time traffic police gesture recognition and response system based on lightweight deep learning models and edge computing architecture. The scheme adopts a dual algorithm parallel technical route combining Media Pipe Hands and a lightweight YOLO, and employs a master–slave hardware architecture consisting of a Raspberry Pi and an STM32 microcontroller. A gesture control set covering core instructions such as "stop," "go straight," "turn left," and "accelerate" is designed, incorporating state machine logic and multilevel latency optimization strategies. The proposed system balances recognition accuracy and real-time performance, featuring low cost, transferability, and ease of deployment, and may serve as a reference for human–machine collaborative interaction in autonomous driving within complex traffic scenarios.

KEYWORDS: *gesture recognition; autonomous driving; Media Pipe; YOLO; edge computing; human–computer interaction.*

How to cite this paper: Qi Yuxuan | Zhou Hanyu | Feng Linshu | Liu Tonghan | Wu Yi "Human AI Collaboration: System Design for Real-time Recognition and Response to Traffic Police Gestures for Autonomous Vehicles" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-10 | Issue-4, August 2026, pp.1322-1326, URL: www.ijtsrd.com/papers/ijtsrd142181.pdf

IJTSRD142181

I. INTRODUCTION

1. Research Background and Significance

The introduction of the Transformer architecture in 2017 greatly accelerated the development of generative artificial intelligence. Subsequently, large language models such as GPT, BERT, LLaMA, as well as domestic models like DeepSeek, Doubao, and Qwen, have emerged one after another, with AI technologies permeating all aspects of social life at an unprecedented pace. In 2025, "AI+" was included in the government work report, and the State Council subsequently issued the Opinions on Deeply Implementing the "AI+" Action, systematically deploying six major action areas including science and technology, industrial development, and more. Meanwhile, AI has been incorporated into the national medium and long term planning; the Recommendations of the Central Committee of the Communist Party of China on Formulating the 15th Five Year Plan for National Economic and Social Development explicitly propose "fully implementing the 'AI+' action."

Driven strongly by policies, the global user base of generative AI has exceeded 1.2 billion, with Chinese users reaching 602 million by December 2025, representing a penetration rate of 42.8%. Notably, the highest AI usage rate, 91%, occurs among the 18–24 age group; in the workplace, AI tool usage is as high as 93%. AI is gradually evolving from a mere technological tool into a "collaborator" that works deeply and symbiotically with humans.

In the field of autonomous driving, despite rapid technological progress, existing systems are almost incapable of recognizing and responding to onsite traffic police gesture commands. This is a practical and urgent problem-when an autonomous vehicle encounters a human traffic controller in a complex open road environment, how to understand and comply with non standardized traffic instructions directly affects driving safety and traffic efficiency. Current mainstream autonomous driving solutions rely heavily on standardized infrastructure such as traffic lights and roadside units, lacking adequate technical preparation for flexible, dynamic onsite human command scenarios.

Against this background, this paper proposes a system design for real-time recognition and response to traffic police gestures specifically for autonomous driving scenarios. The scheme is oriented toward low cost and lightweight implementation, comprehensively employing Media Pipe Hands for hand key point detection and a lightweight YOLO for object detection. Using a master–slave architecture of Raspberry Pi and STM32, the system achieves a complete closed loop from image acquisition, gesture recognition, to vehicle control. The research aims to explore how to achieve high real-time performance and reasonably high accuracy in gesture recognition under the computational constraints of embedded devices, offering a feasible technical approach for autonomous driving in complex open road environments.

2. Current Research Status at Home and Abroad

Gesture recognition for human–computer interaction has been a hotspot in computer vision research in recent years. Current mainstream gesture recognition methods fall into two categories: traditional image processing based methods and deep learning based methods.

In traditional methods, early research mainly relied on skin color detection, contour extraction, template matching, and similar techniques for gesture segmentation and recognition. These methods are computationally light and easy to deploy, but are highly sensitive to illumination changes, background complexity, and hand pose variations, making them insufficiently robust in open scenes.

In deep learning methods, with the development of convolutional neural networks, deep learning based gesture recognition has achieved remarkable progress. The YOLO series, with its end-to-end, single stage detection characteristics, has been widely applied in object detection. Researchers have further proposed lightweight variants such as YOLOv3-tiny, YOLOv5s, and YOLOv8n to accommodate deployment on edge devices. For gesture recognition tasks, improvements based on YOLOv8 have been studied, including GF-YOLOv8 and YOLOv8-GR. Meanwhile, some studies have explored two stage schemes combining YOLO with Media Pipe.

In lightweight hand key point detection, Google's Media Pipe Hands framework provides real-time, high precision 3D detection of 21 hand key points. The framework uses the Blaze Palm lightweight machine learning architecture, enabling real-time inference solely on CPU, which is highly suitable for embedded devices such as the Raspberry Pi. Media Pipe Hands adopts a two stage detection plus regression architecture: first, a lightweight SSD based network performs palm detection to output a palm bounding box; then, 21 3D key point coordinates are regressed within the detected region.

In hardware platforms, the combination of Raspberry Pi and STM32 is a common architecture for embedded AI projects among student teams and researchers. The Raspberry Pi serves as the upper computer for image acquisition and algorithmic inference, while the STM32 acts as the lower computer for motor control; the two communicate via serial port or Wi-Fi for data exchange. This architecture has clear division of labor and controllable costs, making it a suitable base platform for gesture recognition intelligent vehicles.

However, existing studies mostly focus on optimizing individual algorithms, while comprehensive system level design that applies gesture recognition to the specific scenario of traffic police gesture recognition for autonomous driving remains relatively scarce. In particular, integrated considerations in algorithm selection, hardware adaptation, instruction set design, and latency optimization are still insufficient. This paper addresses this gap by proposing a complete system design from a system integration perspective.

II. System Design

1. Overall Architecture

The system builds a complete closed loop around three stages: "perception–decision–execution." In the perception layer, a wide angle camera mounted at the front of the vehicle captures forward road video streams, covering an effective recognition area of 0.5 to 3 meters. The captured image data is transmitted to the Raspberry Pi (upper computer), where the gesture recognition algorithms run for processing. In the decision layer, the Raspberry Pi runs either Media Pipe Hands or a lightweight YOLO model to detect hands and classify gestures in video frames. The recognition result is converted into corresponding vehicle control commands via instruction mapping logic and sent to the lower computer via Wi-Fi or Bluetooth modules. In the execution layer, the STM32 (lower computer) receives the commands and drives the motors to perform the corresponding vehicle actions. The system also incorporates state machine logic requiring that a gesture be sustained for more than 0.5 seconds before triggering a command, thereby preventing false triggers due to hand jitter.

These three stages are sequentially connected: after image acquisition, the camera transmits frames to the Raspberry Pi for gesture recognition; the recognition results are sent wirelessly to the STM32, which finally drives the motors to execute vehicle actions, forming a complete closed loop from visual perception to vehicle execution.

2. Lightweight Gesture Recognition Algorithm Selection

This scheme adopts a dual algorithm parallel design philosophy, balancing recognition accuracy and real-time performance.

First, the Media Pipe Hands local solution. Media Pipe Hands is an opensource, cross platform hand tracking framework from Google. Its core advantages lie in three aspects. First, it is lightweight and efficient-the Blaze Palm architecture enables high precision detection and real-time inference solely on CPU, without requiring GPU acceleration, making it highly suitable for embedded devices like the Raspberry Pi. Second, it provides fine grained key points, extracting 21 3D key points per hand, including fingertips, knuckles, and palm center. Third, it offers temporal stability through temporal filtering techniques that stabilize key point detection across frames and effectively suppress jitter noise. In this scheme, the Raspberry Pi uses Media Pipe Hands to extract 21 hand key points and then performs geometric logic to determine the gesture type-for example, by calculating the relative positions of fingertips and wrist to determine finger extension states, thereby recognizing "five fingers spread" as a full speed ahead command or "fist" as an emergency brake command. This solution runs entirely locally, without network dependency, and with controllable latency.

Second, the lightweight YOLO cloud assisted solution. For high precision recognition in complex scenarios, this scheme also considers deploying a lightweight YOLO model on the cloud as a supplement. Specifically, using cloud GPU instances from platforms such as CSDN Star Chart, a lightweight YOLO model is deployed. The Raspberry Pi uploads captured video frames via REST API to the cloud, where the model performs inference and returns the results. This solution is suitable for scenes with severe illumination changes, complex backgrounds, or hand occlusion where Media Pipe may perform poorly. It should be noted that the cloud solution relies on network communication and has inherent latency uncertainty, so it serves primarily as a supplement to the local solution, invoked when local recognition confidence is low.

Finally, the dual algorithm collaboration strategy. The core idea is "local first, cloud auxiliary." By default, Media Pipe Hands is used for local real-time recognition to meet low latency requirements in daily scenarios. When the local recognition confidence falls below a threshold or detection fails, the system automatically switches to the cloud YOLO solution for secondary confirmation. The two solutions can be flexibly configured according to actual deployment scenarios, balancing real-time performance and accuracy.

3. Hardware Platform Design

In terms of hardware architecture, this scheme adopts a master–slave architecture with Raspberry Pi and STM32. The Raspberry Pi serves as the upper computer, the core perception and decision unit, responsible for connecting the wide angle camera for image acquisition, running gesture recognition algorithms for inference, parsing recognition results into control commands, and sending them via the communication module. The STM32 serves as the lower computer, the execution control unit, receiving command signals from the Raspberry Pi and driving motors to perform actions such as forward movement, steering, and braking. The advantage of this architecture is clear division of labor: the Raspberry Pi focuses on computationally intensive tasks like image processing and algorithm inference, while the STM32 focuses on real-time critical tasks like motor control, resulting in high overall system stability.

In terms of communication protocols, the upper and lower computers establish a communication link via Wi-Fi or Bluetooth modules. The communication protocol is designed to ensure command transmission latency below 100 ms. This can be optimized by adopting a lightweight protocol to reduce packet header overhead, encoding commands in binary rather than text to compress data size, and setting a reasonable baud rate.

For visual sensing, a wide angle camera is mounted at the front of the vehicle to simulate the driver's forward view. The camera must cover an effective recognition area of 0.5 to 3 meters ahead to ensure that police gestures can be captured at a reasonable distance. Image resolution can be set to 480p to reduce data transfer and processing pressure while maintaining recognition quality.

4. Gesture Command Set and Control Logic

This scheme defines an intuitive gesture control command set. A fist corresponds to emergency braking, which has the highest priority as a safety mechanism. Five fingers spread corresponds to full speed ahead as the basic forward command. Index finger pointing up corresponds to acceleration during motion. Palm waving left/right corresponds to steering to control direction.

To prevent false triggering due to natural hand jitter or accidental movements, the system introduces a "gesture duration" judgment logic. Only when the same gesture remains stable in consecutive frames for more than 0.5 seconds will it be recognized as a valid command and trigger the corresponding vehicle action. This design effectively filters out transient and accidental hand movements, improving system reliability and safety. Additionally, the duration threshold can be adjusted based on actual testing to accommodate different operators' habits.

5. System Latency Optimization Strategies

End-to-end latency-from camera capture to motor response-is a key indicator affecting system real-time performance. This scheme optimizes at three levels.

At the image acquisition and transmission level, video resolution is reduced to 480p, reducing data volume in exchange for higher frame rates while preserving necessary image details for recognition. Camera exposure and white balance parameters are reasonably set to minimize framerate fluctuations caused by automatic adjustments.

At the algorithm inference level, Media Pipe Hands is already optimized for mobile and embedded devices, achieving milli second level inference on Raspberry Pi. For the cloud YOLO solution, techniques such as model quantization and pruning can further compress model size and accelerate inference.

At the software architecture level, unnecessary visualization drawing functions are disabled, outputting only control commands to reduce computational overhead. Multithreaded programming is employed to ensure parallel processing of image acquisition, algorithm inference, and command transmission, avoiding serial blocking. Through these optimizations, the end-to-end latency can be controlled within an acceptable range to meet the basic requirements of real-time control.

III. Innovations and Features

1. Scenario Adaptation Innovation

Existing autonomous driving solutions rely heavily on standardized infrastructure such as traffic lights and roadside units, lacking effective recognition and response capabilities for non standardized, highly dynamic traffic police gestures. This scheme focuses on this core interactive pain point in real traffic scenarios, breaking through the dependency on traditional traffic signals and enabling real-time recognition and response to onsite police gestures, filling the technical gap in autonomous driving under manual traffic command scenarios.

2. Technical Solution Innovation

This scheme adopts a dual algorithm parallel lightweight technical route. The Media Pipe Hands solution preserves the ability to capture fine hand features while requiring only CPU operation, perfectly matching the computational constraints of embedded devices like the Raspberry Pi. The lightweight YOLO solution serves as a cloud supplement, providing higher recognition accuracy in complex scenarios. The collaboration between the two solutions achieves a balance between high accuracy and high real-time performance, avoiding the problem of being unable to deploy on low cost hardware when pursuing accuracy alone.

3. Implementation Feasibility Innovation

In hardware selection, this scheme uses the Raspberry Pi + STM32 combination, balancing cost and performance. The Raspberry Pi is moderately priced with a mature ecosystem, while the STM32 offers stable and reliable control. Overall hardware costs are controllable, aligning with student teams' R&D conditions and resource ranges. In system design, a standardized command system and state machine logic are constructed, clearly defining the mapping between gestures and vehicle actions, reducing the risk of false triggers. The entire scheme is suitable for low cost transfer and validation from laboratory simulation to real vehicle scenarios.

4. System Completeness Innovation

Unlike existing studies that mostly focus on single algorithm optimization, this scheme builds a complete closed loop system from data acquisition, algorithm parsing, hardware communication, to vehicle execution. Each module's functions, interfaces, and data flows are clearly defined, and standardized interaction commands and verification processes are designed, offering engineering completeness and can be embedded as an independent perception module into broader autonomous driving platforms.

IV. Future Work Outlook

The project has currently completed system design and validation. Subsequent work will continue in three directions.

In technical depth iteration: On the algorithm side, to address Media Pipe Hands' instability in strong light, backlight, and hand occlusion scenarios, we plan to introduce multimodal feature fusion ideas, using software simulation to add virtual lighting, occlusion, and interference data for training optimization. Meanwhile, we will perform pruning and knowledge distillation on the lightweight YOLO model to further reduce model size and accelerate inference. On the hardware side, we will continue testing compatibility with affordable hardware such as Jetson Nano and

ESP32-S3. In the communication link, we plan to simulate 5G edge transmission protocols in software and optimize data interaction logic to reduce latency redundancy via program compression.

In scenario expansion and validation: We will expand scenarios using virtual datasets and simulation platforms, generating synthetic samples for rain, night, backlight, traffic occlusion, etc. We will also extend the gesture command library to include more complex commands such as "pull over" and "slow down and yield." Additionally, we plan to replicate closed scenarios like campus autonomous vehicles, port container trucks, and airport shuttles in simulation platforms to virtually validate remote gesture scheduling logic, and integrate with vehicle–road collaboration simulation modules to build a unified data flow link among police gestures, onboard terminals, and roadside signals.

In outcome consolidation and dissemination: Since our team consists entirely of undergraduate students without AI related majors, limited by technical development capabilities and engineering practice experience, we primarily focus on theoretical synthesis and summarization. Specifically, we will consolidate the ideas, literature reviews, and system framework accumulated during the project investigation and design phase into written materials for future researchers' reference. At the same time, we plan to compile the theoretical analysis, algorithm comparison logic, and system design into documents, hoping to provide some inspiration for fellow students interested in autonomous driving and human–computer interaction. We are fully aware of our limitations; therefore, this project's emphasis is more on raising questions, analyzing problems, and offering feasible solution proposals rather than actual engineering implementation. We look forward to students with relevant technical backgrounds continuing to deepen this work in the future.

V. Conclusion

This paper addresses the practical question of "how autonomous vehicles can understand human traffic commands" by proposing a real-time traffic police gesture recognition and response system based on lightweight deep learning models and edge computing. The scheme adopts a dual algorithm parallel architecture of Media Pipe Hands and lightweight YOLO, combined with a Raspberry Pi and STM32 master–slave hardware platform, designing a complete closed loop system of perception, decision, and execution. The scheme balances recognition accuracy and real-time performance, with low cost, transferability, and ease of deployment.

The preliminary research results of this project have been compiled into a project proposal and participated in the "Challenge Cup" National College Students' Entrepreneurship Competition, winning a third prize. The team will continue to work under low cost, lightweight conditions, iterating through software development, consolidating outcomes, and refining system performance and virtual scenarios, aiming to provide useful references for campus AI innovation research and human–machine collaborative interaction in autonomous driving.

References

[1] Beijing Wuzi University College Student Innovation and Entrepreneurship Training Program Project Group. Exploratory Research on the State of Human–AI Collaboration in the Next Decade – Project Application Report [R]. Beijing: Beijing Wuzi University, 2026.

[2] Google. Media Pipe Hands: Ondevice Real-time Hand Tracking [EB/OL].

[3] Research and Application of Gesture Recognition Algorithm Based on Media Pipe [J]. Intelligent Computers and Applications, 2024(04).

[4] Research on Improved YOLOX Based Gesture Recognition Algorithm and Its Lightweight Implementation [D].

[5] Static Gesture Recognition Detection Model Based on YOLOv5 [J]. Software Guide, 2024(11).

[6] Implementation of Gesture Recognition Smart Car Control via Raspberry Pi & STM32 Serial Communication [EB/OL].

[7] Implementation of Finger Gesture Based Medicine Delivery Robot Control System with Media Pipe [J].

[8] Real-time Gesture Recognition Method Integrating Multiple Lightweight Strategies and Media Pipe [J]. Computer Engineering and Applications, 2026.