Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070115
ISSN 2516-5003 Vol. 7 Num. 1

Design of an End-Cloud Collaborative Urban Security Robot System Based on Large Model

Saibo Li, Jiatong Lv, Wenxuan Zhang, Jianbin Feng, Xinning Zhang, Xinghai Wu, Maoxiang Chu*

University of Science and Technology Liaoning, Anshan, 114051, Liaoning, China
754136793@qq.com

Keywords: Smart city; Security robot; Large model; End-cloud collaboration; RBF interpolation algorithm; Zero-shot detection

Abstract: Aiming at the pain points of slow response, poor generalization ability of target recognition and high labor cost in traditional urban security systems in complex environments, this paper designs and implements an end-cloud collaborative security robot system driven by multimodal large model. The system adopts a three-terminal collaborative architecture of cloud decision-making, edge scheduling and end execution. The cloud deploys DINO-X Pro multimodal large model to realize accurate recognition of arbitrary semantic targets by its zero-shot learning ability. The edge side takes RDK-X5 as the main control core to build a three-level cascade framework of large detection, small detection and tracking, which solves the contradiction between large model inference delay and end-side computing power limitation. The end realizes high-precision intervention through hexapod chassis and GD32 shooting control board. The experimental data show that the system tracking frame rate is stable at 29.9 FPS, and the shooting hit rate reaches 75%, which is 55% higher than the traditional incremental PID algorithm. It has significant application value in the field of smart city dynamic security.

1. Introduction

1.1. Research Background and Significance

With the deepening of smart city construction and the rapid development of Internet of Things technology, the urban public security system is undergoing a profound transformation from passive monitoring to active defense, which puts forward stricter requirements for the intelligent, motorized and practical level of security equipment. Traditional urban security methods mostly rely on fixed monitoring cameras or manual patrols, which have obvious limitations. Fixed monitoring has a large number of blind spots and can only verify after events, so it is difficult to give early warning and block sudden violent events in real time. Most mobile security robots use traditional target detection algorithms, which rely heavily on labeled data sets and have poor generalization ability in open urban environments.

The rapid development of multimodal large models brings subversive breakthroughs to machine vision cognition. Large models have strong zero-shot and open-set detection capabilities based on

prior knowledge accumulated by massive data pre-training. However, the large number of parameters leads to high computing power consumption and high network delay, which cannot meet the real-time requirements of security robots[1]. In order to solve the problem that high-level cognitive ability and low-delay execution ability are difficult to obtain at the same time, this paper designs an end-cloud collaborative urban security robot system driven by multimodal large model. The system constructs a closed loop from macro environment perception to micro physical intervention, fills the technical gap between intelligent patrol and real-time blocking in complex environments, and provides a practical engineering implementation paradigm for a new generation of embodied intelligent security terminals empowered by large models[2].

1.2. Research Status at Home and Abroad

Foreign research on security robots started early, and the products have high integration of motion control and environment perception, but the cost is high and the algorithm is closed, which is difficult to adapt to the complex urban environment in China. Domestic mobile security robots have made great progress in chassis motion and basic recognition, but most of them use traditional detection models with limited recognition categories and poor generalization for unknown targets. In terms of end-cloud collaboration, most schemes only complete simple data upload and instruction release, lack the cascade design of large and small models, and cannot balance high-precision recognition and low-delay tracking. The application of multimodal large model in the field of security robots is still in the exploratory stage, and there is a lack of mature engineering solutions for real-time physical intervention[4].

1.3. Main Research Content and Structure

This paper takes the end-cloud collaborative architecture as the core, and carries out the design and research of urban security robot system around multimodal perception, all-terrain movement, precise intervention and remote scheduling. The overall structure is as follows. The second part introduces the relevant technical foundations including multimodal large model, end-cloud collaboration, target tracking and motion control. The third part completes the overall system design including architecture, hardware and software. The fourth part carries out the hardware circuit design of the system. The fifth part designs the software function modules. The sixth part carries out system test and data analysis. The seventh part summarizes the research work and prospects the future optimization direction.

2. Related Technical Foundations

2.1. Multimodal Large Model and Zero-Shot Detection Technology

Multimodal large model integrates image and text features through Transformer architecture, and has open-set recognition ability after large-scale data pre-training. DINO-X Pro model used in this paper supports natural language prompt word detection, can locate any semantic target without retraining, and solves the problem of low recognition rate of long-tail targets in traditional algorithms. Zero-shot detection technology does not rely on labeled data sets of specific categories, and can directly identify unknown targets in complex scenes, which greatly improves the adaptability of the system to open urban environments[5].

2.2. End-Cloud Collaborative Computing Technology

End-cloud collaborative computing distributes computing tasks according to the characteristics of cloud and end equipment. Cloud servers undertake high computing power consumption tasks such as large model inference, and end-side embedded platforms complete lightweight algorithms and real-time control. This architecture gives full play to the cognitive advantages of large models and the real-time advantages of end-side execution, reduces the dependence on network bandwidth, and effectively reduces the delay caused by pure cloud processing. It provides a technical basis for the robot to realize high-frame-rate target tracking and fast response[2].

2.3. Target Tracking and Motion Control Technology

ByteTrack multi-target tracking algorithm associates detection frames to optimize target trajectories, with high tracking stability and low computing power consumption. RBF interpolation algorithm establishes a non-linear mapping model from pixel coordinates to pan-tilt angles, which can compensate for mechanical assembly errors and camera optical distortion, and improve control accuracy. Hexapod bionic motion control adapts to unstructured roads such as stairs and ruins through multi-servo cooperative drive, and realizes stable walking and automatic attitude adjustment[3].

3. Overall System Design

3.1. Overall System Architecture

The system adopts a three-terminal collaborative architecture of cloud decision-making, edge scheduling and end execution, as shown in Figure 1. The cloud is responsible for deploying multimodal large models to complete zero-shot target recognition and feature extraction. The edge side takes RDK-X5 as the core to run end-side detection and tracking algorithms, and schedules motion and shooting tasks. The end includes hexapod chassis, pan-tilt and shooting mechanism to perform physical intervention. The web console realizes remote visualization scheduling and human-computer interaction, forming a closed-loop system of perception, decision-making, execution and management[1].

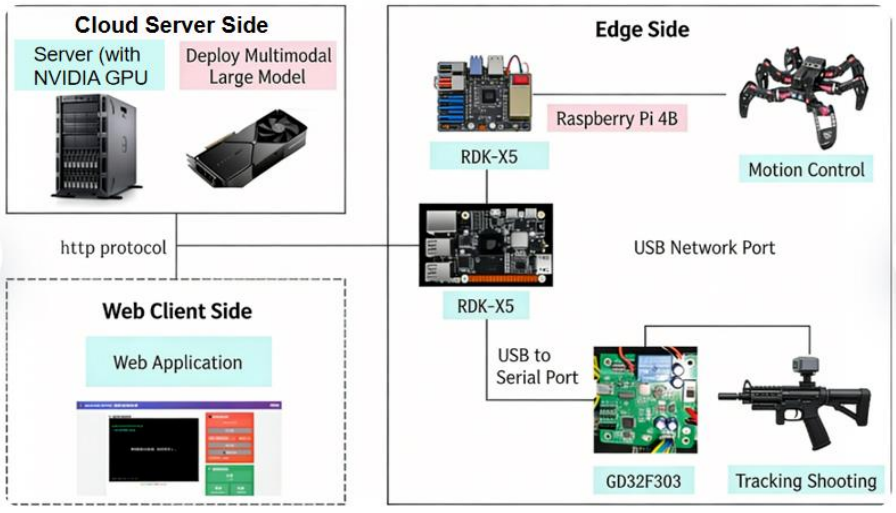

Figure 1 Three-terminal collaborative architecture diagram

3.2. Hardware Structure of Robot System

The physical form of the system adopts a four-layer bearing structure, as shown in Figure 2. The bottom layer is a hexapod bionic chassis with 18 serial bus steering gears, which has strong obstacle surmounting and omnidirectional movement capabilities. The middle layer is a control cabin integrating main control, power supply and communication modules. The top layer is a two-dimensional pan-tilt shooting component, which can realize yaw and pitch adjustment. The hardware adopts CPU-GPU-MCU three-core heterogeneous collaboration to meet the needs of high computing power AI reasoning and high real-time motion control[4].

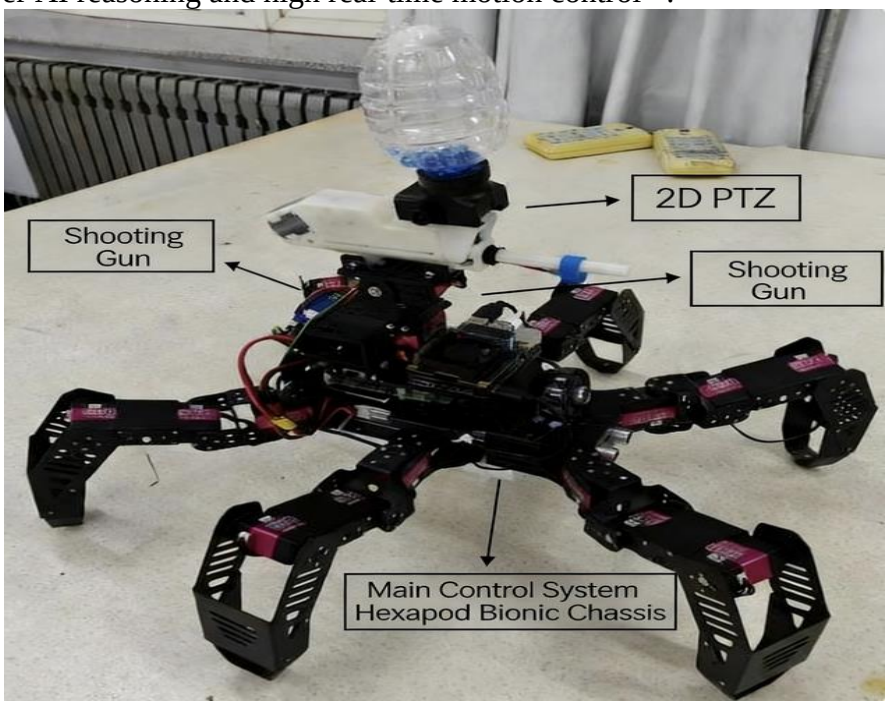

Figure 2 Overall view of the work

3.3. Core Functional Modules

The system is divided into four core functional modules: semantic perception module, motion control module, precise intervention module and remote scheduling module. The semantic perception module completes arbitrary target extraction through cloud large model and end-side template matching. The motion control module drives the hexapod chassis to realize all-terrain patrol and adaptive level adjustment. The precise intervention module calculates the pan-tilt angle through RBF algorithm and completes non-lethal shooting. The remote scheduling module realizes real-time video return and instruction release through the web console[5].

4. Hardware Design of the System

4.1. Overall Hardware Framework

The hardware design adopts a modular structure with three-core heterogeneous collaboration, which is divided into main control computing module, motion control module and shooting control module. The main control computing module uses RDK-X5 with 10 TOPS equivalent computing power to process visual data and network communication. The motion control module uses Raspberry Pi 4B to drive the hexapod chassis. The shooting control module takes GD32F303 as the

core to drive the pan-tilt and shooting mechanism. The power supply adopts double-stage voltage stabilization design to isolate the power supply of logic components and high-power components, ensuring system stability.

4.2. Main Control and Communication Module

RDK-X5 is the core of edge computing, which receives the target features returned by the cloud, runs template matching and ByteTrack tracking algorithm, and maintains the tracking frame rate at 29.9 FPS. The communication module uses lightweight network protocol and internal network penetration technology to realize low-delay data interaction between cloud, end and web console. The module supports real-time return of video stream and equipment status, and receives remote instructions such as target selection and shooting authorization.

4.3. Motion and Shooting Drive Module

The motion drive module controls 18 steering gears through Raspberry Pi 4B, and realizes actions such as forward, backward, turning and obstacle surmounting by loading action group files. The chassis automatically adjusts the body height according to the terrain to keep the body level and reduce the impact of vibration on visual tracking. The shooting drive module uses GD32F303 to receive coordinate deviation, outputs PWM signals to drive the pan-tilt steering gear, and controls the shooting motor through a relay to complete accurate shooting. The circuit design of the shooting control substrate is shown in Figure 3, which integrates main control core circuit, power management circuit and relay drive circuit to ensure safe and stable execution of physical intervention.

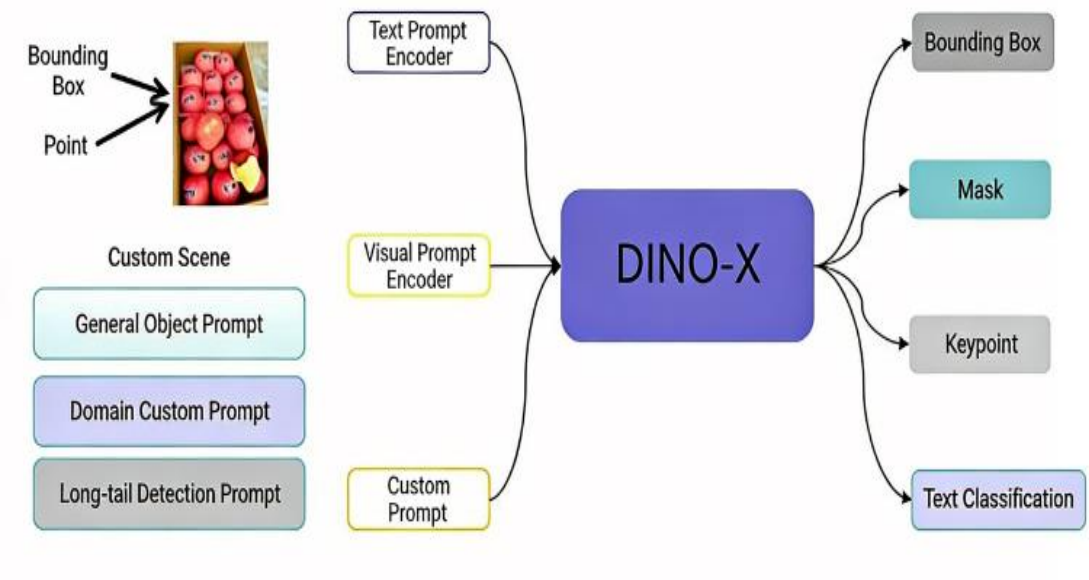

Figure 3 Schematic diagram of the shooting control base plate circuit

5. System Test and Data Analysis

5.1. Test Scheme

The test simulates real campus and indoor scenes. The target moves at about 5 km per hour. The test indexes include mean average precision, tracking success rate, tracking frame rate and shooting hit rate.

5.2. Image Processing Test

Table 1 Image processing contrast test data

Processing scheme	Mean average precision	Tracking success rate	Tracking frame rate
Pure end-side detection	72.4%	79.5%	29.8 FPS
Pure cloud detection	90.4%	23.2%	4.5 FPS
End-cloud collaborative detection	89.1%	75.4%	29.9 FPS

As shown in Table 1, the pure end-side scheme achieves high frame rate but delivers low recognition accuracy. The pure cloud scheme obtains high accuracy yet suffers from serious transmission delay. The end-cloud collaborative scheme maintains high detection accuracy and stable high frame rate, which effectively balances target recognition performance and real-time tracking capability.

5.3. Shooting Accuracy Test

Table 2 Shooting hit rate contrast data

Control algorithm	Shooting hit rate
Incremental PID algorithm	20%
RBF interpolation algorithm	75%

As shown in Table 2, the traditional incremental PID algorithm is severely affected by nonlinear errors such as camera distortion and mechanical clearance, resulting in an extremely low shooting hit rate. In contrast, the RBF interpolation algorithm effectively compensates for these errors, increasing the hit rate by 55% and achieving high practical reliability.

6. Conclusion

This paper designs an end-cloud collaborative urban security robot system based on multimodal large model. The system constructs three-terminal collaborative architecture, uses end-cloud cascade mechanism to improve tracking frame rate, and applies RBF interpolation algorithm to improve shooting accuracy. The test shows that the system has high recognition accuracy, stable tracking and excellent intervention effect, which can be widely used in dynamic security scenarios.

In the future, the system can add lidar and other sensors to enhance autonomous navigation ability, optimize model deployment to reduce network dependence, and expand voice interaction and other functions to improve scene adaptability.

Acknowledgement

Liaoning Province College Students Innovation and Entrepreneurship Training Program.

References

[1] Zijian L. Integration of Artificial Intelligence and Internet of Things to Construct Urban Security Monitoring System[J].Journal of Safety Science and Engineering,2024,1(4):DOI:10.62517/JSSE.202408405.

[2] ZhengJ ,JeonS ,YangX .GRDATFusion: A gradient residual dense and attention transformer infrared and visible image fusion network for smart city security systems in cloud and fog computing[J].Expert Systems,2024,42(2):e13685.DOI:10.1111/EXSY.13685.
[3] GuoS ,WuK ,JeonS , et al.IETAFusion: An illumination enhancement and target-aware infrared and visible image fusion network for security system of smart city[J].Expert Systems,2024,42(1):e13538.DOI:10.1111/EXSY.13538.
[4] Sun T ,Xia Y ,Gan Y .Discussion on Integration of Urban Video Surveillance System[J].Procedia Engineering,2011,15(C):3255-3259.DOI:10.1016/j.proeng.2011.08.611.
[5] Yihan K ,Xuanang R ,Renjie Z .Deep learning for image super-resolution and its application in urban security[C]//The Storm King School (United States);Chengdu Foreign Language School (China);Chengdu Tanghu Foreign Language School (China),2022:DOI:10.1117/12.2639314.