

基于 DeepSeek 的行为监测系统的设计与实现

李晨, 黄勇萍*

广西民族师范学院数学与计算机科学学院, 广西 崇左 532200

DOI: 10.61369/TACS.2026060003

摘 要 : 随着社会智能化和数字化的快速发展, 公共安全、家庭安全和交通管理等领域对行为监测和异常情况检测的需求日益迫切。基于此, 设计一个基于 DeepSeek 大模型与 YOLOv8n 融合的行为检测系统, 实现视觉信号到自然语言描述的转换。系统采用分层架构设计, 前端层基于微信小程序实现用户交互与视频展示, 业务逻辑层通过云函数处理请求转发与用户认证, AI 服务层集成 YOLOv8n 模型进行实时目标检测, 并通过规则引擎触发 DeepSeek API 调用, 生成异常行为的自然语言描述。由于大模型 API 调用频率过高导致成本高的问题, 设计了异常防抖与 LLM 冷却两级节流策略。

关 键 词 : 行为监测; Deepseek; YOLOv8n; 异常检测

Design and Implementation of a Behavior Monitoring System Based on DeepSeek

Li Chen, Huang Yongping*

School of Mathematics and Computer Science, Guangxi Normal University for Nationalities, Chongzuo, Guangxi 532200

Abstract : With the rapid advancement of societal intelligence and digitalization, the demand for behavior monitoring and anomaly detection in public safety, home security, and traffic management has become increasingly urgent. To address this, a behavior detection system integrating the DeepSeek large model with YOLOv8n is designed to enable the conversion of visual signals into natural language descriptions. The system employs a hierarchical architecture: the front-end layer facilitates user interaction and video display via a WeChat mini-program; the business logic layer handles request forwarding and user authentication through cloud functions; and the AI service layer integrates the YOLOv8n model for real-time object detection, triggering DeepSeek API calls via a rule engine to generate natural language descriptions of anomalous behaviors. To mitigate high costs associated with excessive API calls to large models, a two-stage throttling strategy combining anomaly debouncing and LLM cooling is implemented.

Keywords : behavior monitoring; DeepSeek; YOLOv8n; anomaly detection

引言

传统监控系统大多采用录像回放的方式, 缺乏智能异常检测、主动预警等能力, 需要人工检测异常行为; 而人工异常行为检测效率低且往往存在歧义理解^[1]。近年来, 大语言模型 (Large Language Model, LLM) 的兴起为智能感知与人机交互带来了新机遇和发展, 将视觉检测与大模型语义理解相结合, 成为提升系统可解释性的重要方向。国内外学者在行为检测领域开展了大量研究。文献^[2]提出了一种结合 YOLOv8n 与 MediaPipe 两种检测模型结合的跌倒检测系统, 主要是利用目标框宽高比的变化作为跌倒判据, 当人体从站立时的“高瘦”状态变为“矮胖”状态时触发告警。2025 年, Cao 等^[3]在针对 YOLOv8n 进行多项改进的情况下, 引入 CARAFE 上采样算子和 LSKNet 注意力机制, 有效提升了老年人异常行为检测精度; Shi 等^[4]提出了自上而下的跌倒检测算法, 在 YOLOv8n 检测基础上结合人体关键点信息, 如: 质心速度、躯干角度等作为判断依据; Wang 等^[1]则将改进的 DeepSeek 与 YOLOv8n 相结合, 提出了 SAL-YOLO-DeepSeek 框架并应用于自动扶梯安全监控。当系统检测到异常时立即调用 DeepSeek 动态生成响应策略, 实现了检测到决策的端到端智能化。

项目信息:

广西民族师范学院自治区级大学生创新创业训练计划项目 (项目编号: S202510604093);

广西民族师范学院 2025 年度校级科研项目 (项目编号: 2025ZD015)。

作者简介: 李晨 (2004—), 男, 湖北利川人, 本科生; 研究方向: 计算机网络、计算机视觉。

* 通讯作者: 黄勇萍 (1985—), 女, 广西藤县人, 高级实验师, 硕士; 研究方向: 计算机网络、人工智能。

然而，目前的行为检测方案多关注算法层面，轻量级端云协同部署及 API 调用成本控制方面仍有优化空间。基于此，本文设计并实现一个基于 DeepSeek 与 YOLOv8n 融合的端云协同行为监测系统，采用“本地 AI 服务 + 微信云函数 + 小程序前端”的架构，在成本控制方面设置了异常防抖与 LLM 冷却两级节流策略以降低 API 调用成本，实现了从视觉信号到自然语言描述的实时转换，并且人机交互性能良好，为独居老人看护、校园安全等场景提供了有效的异常行为检测解决方案。

一、系统总体设计

本系统采用分层架构设计，采用端-边-云协同范式^[5]，具体划分为前端层、业务逻辑层、AI 服务层和数据层，如图1所示。其中，前端层基于微信小程序实现，负责用户交互、实时视频流展示、告警弹窗及历史记录查询。业务逻辑层部署在微信云环境中的云函数，承担用户认证、请求转发、与本地 AI 服务通信等任务。AI 服务层运行于本地 PC 的 Flask 应用中，集成 YOLOv8n 目标检测模型和 DeepSeek API，负责视频采集、实时推理、异常判断、语义生成及告警推送。数据层使用云开发数据库存储用户信息、及历史告警记录等信息。

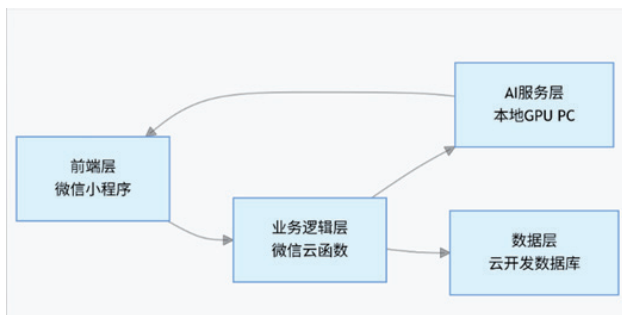


图1 系统分层架构

这种分层设计既发挥了本地 GPU 的算力优势保障实时检测，又通过云函数隔离内外网提升系统安全性，同时基于微信生态环境降低用户使用门槛^[5]。

二、关键技术实现

（一）视觉感知层：YOLOv8n 实时检测

1. 模型部署与优化

本地 AI 服务使用 ultralytics 库加载预训练的 YOLOv8n 模型，并通过 model.to(cuda) 将模型移至 GPU。在模型预热阶段，进行自检测试以确保 GPU 及其加速依赖正常工作。在工作阶段：检测主循环从摄像头读取帧，经 BGR 转 RGB 后输入模型推理，得到目标边界框、类别和置信度。由于本系统在模型部署中参考了基于改进 YOLOv8n 的融合算法^[3]的优化思路，在使用 CUDA 加速后对单帧的处理时间稳定在 30~50ms，与同类系统^[4]的 18.1fps 相比，本系统在 GPU 加速下可达 60FPS，满足实时监控需求。

2. 异常判定规则

本系统的异常判定规则：利用目标框宽高比的变化检测摔倒行为。当系统检测到人体目标时，先计算其边界框的宽高比，如

若宽高比小于 0.8，且置信度高于 0.7，则判定为疑似摔倒^[2]。此外，鉴于生活场景的复杂需求本系统针对小物体（如瓶子、杯子）快速移动及大型物体（如椅子）倾倒也可进行相应拓展。未来可借鉴引入人体关键点信息以进一步提高检测准确率的方法^[4]。

（二）语义理解层：DeepSeek 自然语言生成

1. Prompt 设计与调用策略

Prompt 设计：包括角色设定（安全监控分析员）、场景上下文（办公室监控）、检测结果（目标类别、置信度以及位置信息）和要求输出内容（简短自然语言形式）^[1]。如下所示：

你是专业安防监控助手，请基于以下检测结果以简练、专业语言说明当前场景存在何种异常行为：

（场景上下文）环境：办公室监控

（检测到的异常行为）1. 类别：person | 置信度：0.95 | 位置：[100,200,300,400]

（输出要求）1. 有人疑似摔倒。2. 语气客观冷静。

系统实现的过程中首先将视觉检测结果传给 DeepSeek 大模型进行语义分析，实现由视觉信息到文字的跨模态转换，在检测出异常行为后才会发起对 DeepSeek API 的请求，而在正常情况下不会产生不必要的开销。

2. 资源节流策略

为了解决大模型 API 调用成本较高问题，本系统提出了一种异常防抖以及 LLM 冷却两种限流方式。

（1）异常防抖：两次告警最小间隔 5 秒，防止短时间内重复告警。

（2）LLM 冷却：10 秒内只请求一次 DeepSeek API。

在连续异常情况的测试下，API 调用次数由理论上最大值 360 次/小时降低到 72 次/小时，下降的比例达到 80%，节省了费用。

3. 告警队列与推送

本地服务维护一个 recent_alerts 队列来保存最近 10 条告警。云函数定时从本地服务的 /get_alerts 接口获取最新的告警并显示到微信小程序中。目前的研究显示 DeepSeek 生成的文章具有很高的自然度，因此本系统告警的信息也是可以阅读的^[5]。

（三）交互应用层：微信小程序

1. 实时视频流实现

小程序通过云函数 getFrame 轮询本地服务的 /current_frame 接口，获取 Base64 编码的视频帧，并更新页面中的 <image> 组件实现类实时显示。关于轮询间隔需要用户选择的画质作为动态调整：流畅 100ms、标清 150ms、高清 250ms。这种轮询方式参考云端终端协作系统^[7]中关于端云协同视频传输的讨

论，在5G网络下标清模式可稳定运行，丢帧率低于5%。

2. 告警展示与用户交互

每当云函数检测到新的异常告警时，小程序按流程主动弹出组件显示 DeepSeek 生成的描述，并且触发振动和可选的声音提示。历史记录告警列表查询预留有接口后续拓展可实现。用户可以通过小程序开关声音、切换画质、查看系统状态。

3. 用户认证模块

小程序支持手机号密码登录和微信授权登录两种方式。云函数 login 和 wxLogin 处理认证逻辑，验证用户信息后生成 token 并存入云数据库的 user_sessions 集合。后续请求携带 token 进行身份验证，确保系统安全性。

（四）端云协同通信

1. 本地 AI 服务接口设计

本地 AI 服务提供以下 RESTful 端点：

- （1）/health：用于健康检查，返回服务状态及本地 IP。
- （2）/start：用于启动摄像头与检测线程。
- （3）/stop：用于停止检测并释放资源。
- （4）/detect：用于单帧检测，返回检测结果及 Base64 帧。
- （5）/current_frame：用于获取当前视频帧，用于小程序直播。
- （6）/get_alerts：用于获取最新告警数据。

这种分层接口设计符合云端终端协作系统提出的模块化原则，便于后期的系统扩展和维护。

2. 云函数代理与容错

云函数 behavior-detection 封装了对本地服务的请求，增加用户认证、请求限流、日志记录等功能。针对网络不稳定的问题，设置了重试机制（最多3次，间隔3秒），连接失败时生成测试帧保证 UI 可用性，体现了云端终端协作系统强调的容错能力。

三、实验结果与分析

（一）实验环境

（1）硬件：本地 PC 配备 Intel i7-13700H CPU、NVIDIA GeForce RTX 4060 Laptop GPU（8GB 显存）、16GB 内存；手机端为小米 14。

（2）软件：Python 3.8、PyTorch 2.1.0、ultralytics 8.0.200、Flask 2.3.3、微信开发者工具 Stable 1.06、微信基础库 2.24.0。

参考文献

- [1]Wang Q., Wang D., Lu J. et al.SAL-YOLO-DeepSeek:a lightweight real-time detection and LLM-driven decision framework for intelligent escalator safety monitoring[J].Scientific Reports,2025,15(1):40600.
- [2]Chakurkar P, Kothari V.Towards safer environments: a YOLO and MediaPipe-based human fall detection system with alert automation[J].MethodsX,2025,13:103245.
- [3]Cao X, Li C, Zhai H.YOLO-AB:A Fusion Algorithm for the Elders' Falling and Smoking Behavior Detection Based on Improved YOLOv8[J].Computers, Materials & Continua,2025,83(3).
- [4]Shi H, Wang X., Shi J.Fall detection algorithm using enhanced HRNet combined with YOLO[J].Sensors,2025,25(13):4128.
- [5]Gong L., Yang H, Fang G., et al. A survey on video analytics in cloud-edge-terminal collaborative systems[EB/OL]. arXiv:2502.06581, 2025. <https://doi.org/10.48550/arXiv.2502.06581>.

（3）测试数据：实验室环境下采集大约 24 小时监控视频，有正常行走、摔倒、物体移动等。

（二）功能验证

1. 异常检测验证

在实验模拟过程中，测试人员随机执行 10 次摔倒以及 10 次正常行走，系统正确识别 8 次摔倒，有 2 次误报。说明规则引擎对于常见异常行为是有效的。

2. 语义生成效果验证

选取 50 个异常事件，让 10 位参与者给 DeepSeek 生成描述给出自然程度打分（1-5 分）。结果为平均分为 4.2 分，说明生成告警信息易读性强。比如对于摔倒事件，系统给出“检测到办公室有人疑似摔倒，请及时查看”，语义正确，表达流畅。

3. 端云协同稳定性验证

系统持续工作 24 小时，统计丢帧率以及重试次数，丢帧率为 3.2%，重试次数共 15 次，基本无稳定问题。

（三）对比分析

与 MediaPipe 比较：最初使用 MediaPipe 作为检测模型，但由于存在版本问题以及对复杂场景不够，最后选择使用 YOLOv8n。YOLOv8n 有大量网上资料可供参考，而且有预训练好的模型可以调用，之后还可以加入自己的数据集进行进一步训练，能够识别 80 种不同的物体，更加符合任务的需求。

与无节流策略比较，在无节流策略中，如果遇到连续相同异常情况会引起大量 API 请求造成每天上百元开支难以管理，而使用两级节流之后每天请求量基本保持在较低水平上，节约了 80% 以上开支说明节流是有必要的。

四、结语

本文设计并实现了一个基于 YOLOv8n 与 DeepSeek 融合的端云协同行为监测系统，通过 YOLOv8n 进行异常行为检测，总结数据交给 DeepSeek 分析并生成自然语言描述返回给用户，在成本控制方面通过两级节流机制有效的降低 API 调用成本。就实验与研究结果来看，本系统在检测延迟、帧率、准确率等方面均满足实时监护需求，达到项目目的。未来将在以下两方面进一步研究：（1）边缘部署：随着边缘计算的发展，可将模型压缩后部署于边缘设备，实现更低延迟的本地推理，符合云端终端协作系统指出的发展趋势。（2）多模态融合：引入声音、文本等信息，构建全方位智能感知系统。