

基于多模态视觉特征融合的实时行为检测与分析系统

叶继阳¹, 屠益军²

1. 金华职业技术大学信息工程学院, 浙江金华

2. 杭州达那福医疗器械有限公司, 浙江杭州

DOI:10.62836/gse.v3i3.1479

摘要: 针对长时间单调作业场景中人员疲劳与分心行为难以实时监测的问题, 本文提出了一种基于多模态视觉特征融合的实时行为检测与分析系统。该系统融合了YOLOv8目标检测网络、MediaPipe面部关键点定位以及传统计算机视觉方法, 构建了多层次的行为识别流水线。在特征层面, 系统同时提取眼部纵横比(EAR)、嘴部纵横比(MAR)、头部姿态角以及手部一面部空间关系等多维特征, 通过时序状态机模型对瞬时检测结果进行时序融合, 有效降低了误报率。系统还集成了基于大语言模型的智能数据分析模块, 能够自动生成结构化的行为评估报告。实验结果表明, 该系统在真实场景下对闭眼、打哈欠、低头及手持电子设备等行为的综合检测F1分数达到91.2%, 平均推理延迟为38ms/帧, 满足实时性要求。此外, 系统提供了完整的数据管理、可视化分析及多格式导出功能, 具备良好的工程实用性。

关键词: 行为检测; 计算机视觉; YOLOv8; MediaPipe; 时序状态机; 多模态融合

A Real-Time Behavior Detection and Analysis System Based on Multimodal Visual Feature Fusion

Jiyang Ye¹, Yijun Tu²

1. Information Engineering College, Jinhua University of Vocational Technology, Jinhua, Zhejiang

2. Hangzhou Danafu Medical Equipment Co., Ltd., Hangzhou, Zhejiang

Abstract: Real-time monitoring of fatigue and distraction behaviors among personnel engaged in prolonged monotonous tasks remains a significant challenge. To address this issue, this paper proposes a real-time behavior detection and analysis system based on multi-modal visual feature fusion. The system integrates the YOLOv8

*作者简介: 屠益军(1981—), 男, 浙江诸暨人, 学士, 工程师, 研究方向: 人工智能, 孤独症谱系障碍, 全龄友好无障碍环境建设;

通讯作者: 叶继阳(1982—), 男, 浙江金华人, 硕士, 副教授, 研究方向: 人工智能, 虚拟现实, 孤独症谱系障碍; 邮箱: 20050879@jhc.edu.cn

基金项目: 浙江省教育厅高等学校访问工程师校企合作项目“基于人工智能和增强现实的孤独症儿童评估应用研究”(项目编号: FG2025023)研究成果

object detection network, MediaPipe facial landmark localization, and conventional computer vision techniques to construct a multi-stage behavior recognition pipeline. At the feature level, the system simultaneously extracts multiple dimensions of features, including the Eye Aspect Ratio (EAR), the Mouth Aspect Ratio (MAR), head pose angles, and hand-to-face spatial relationships. A temporal state machine model is employed to perform temporal fusion over instantaneous detection results, thereby effectively reducing the false positive rate. The system further incorporates an intelligent data analysis module powered by a Large Language Model (LLM), which automatically generates structured behavioral assessment reports. Experimental results demonstrate that the system achieves an overall detection F1 score of 91.2% for behaviors including eye closure, yawning, head-down posture, and handheld electronic device use in real-world scenarios, with an average inference latency of 38 ms per frame, satisfying real-time processing requirements. In addition, the system provides comprehensive data management, visual analytics, and multi-format export functionalities, demonstrating strong engineering practicality.

Keywords: behavior detection; computer vision; YOLOv8; MediaPipe; temporal state machine; multi-modal fusion

1 引言

在工业监控、在线教育、安全值守等需要人员长时间保持注意力的场景中，疲劳和分心行为的及时检测对于保障作业质量和安全具有重要意义。传统的行为监测方法主要依赖人工巡查或穿戴式生理传感器，前者效率低下且难以实现全天候覆盖，后者则存在佩戴不适、设备成本高等局限性。

近年来，基于计算机视觉的行为检测方法因其非接触、易部署的优势受到广泛关注。现有研究多聚焦于单一行为的检测，如基于眼部纵横比（EAR）的疲劳检测或基于目标检测的手机使用识别，缺乏从数据采集、行为识别到统计分析的完整闭环。

针对上述问题，本文设计并实现了一套完整的行为检测与分析系统。主要贡献包括：

（1）提出了一种多模态视觉特征融合框架，将深度学习目标检测、面部关键点定位与传统图像处理有机结合，在不同检测条件下自适应切换检测策略；

（2）设计了基于时序状态机的行为确认机制，通过对连续帧的检测结果进行状态追踪，有效

过滤瞬时噪声，提升检测可靠性；

（3）构建了从实时检测、数据持久化、统计报表生成到AI辅助分析的完整工具链，实现了行为监测的工程化闭环。

2 相关工作

2.1 疲劳行为检测

疲劳检测的研究可大致分为基于生理信号、基于车辆行为（针对驾驶场景）和基于计算机视觉三类方法。在视觉方法中，PERCLOS（Percentage of Eyelid Closure over the Pupil）指标被广泛用于量化眼睛闭合程度。Soukupova等人提出了眼部纵横比（EAR）作为PERCLOS的高效近似，通过计算眼部关键点之间的欧氏距离比值来判断眼睛开合状态，该方法计算量小且对头部姿态具有一定鲁棒性[1]。类似地，嘴部纵横比（MAR）同样基于面部关键点几何关系，被用于检测打哈欠行为。

头部姿态估计是检测注意力分散的另一个重要维度。传统方法依赖PnP（Perspective-n-Point）算法结合3D面部模型进行姿态解算，而基于深度学习的关键点检测方法（如MediaPipe）通过回归面部468个3D关键点，能够以更低的计算开销实现高精

度的姿态估计[2]。

2.2 目标检测与行为识别

YOLO系列算法以其优异的实时性能在目标检测领域占据重要地位。YOLOv8在继承前代架构优势的基础上，引入了C2f模块和Anchor-Free检测头，在检测精度和推理速度之间取得了更好的平衡[3]。在行为识别方面，将目标检测与场景上下文结合的方法已被证明能有效提升特定行为（如手持物体、人机交互）的识别准确率。

2.3 时序建模与状态机

在实时视频分析中，单帧检测结果往往包含噪声。时序状态机（Temporal State Machine）通过维护行为状态的转移逻辑，要求异常行为在连续多帧中持续出现才触发告警，从而有效降低误报率。与基于LSTM或Transformer的时序建模方法相比，状态机方法具有计算开销小、可解释性强的优势，更适合资源受限的实时应用场景。

总体来看，现有研究在以下方面仍存在不足：

（1）多数方法仅聚焦于单一行为（如闭眼或打哈欠），缺乏对多种疲劳与分心行为的综合检测能力，

难以满足实际场景中复杂行为监测的需求。

（2）现有行为检测系统通常止步于实时告警，缺乏从数据采集、检测识别、数据持久化到统计分析的全流程闭环，限制了其在实际部署中的可用性。（3）在光照变化、硬件资源受限等现实条件下，纯深度学习方法面临推理延迟升高或精度下降的问题，需要自适应策略来提高系统鲁棒性。本文正是针对上述不足，提出了一个融合多模态特征与自适应降级机制的完整解决方案。

3 系统架构

3.1 总体架构

本系统采用分层模块化架构，如图1所示，自上而下分为四个层次：

数据层负责所有数据的持久化存储，包括异常截图文件、SQLite关系数据库、运行日志以及JSON格式的配置文件。

检测引擎层是系统的核心，集成了YOLOv8目标检测器、MediaPipe面部关键点检测器以及基于Haar级联的降级检测器。该层输出结构化的检测结果，供上层消费。

业务逻辑层封装了会话生命周期管理、告警

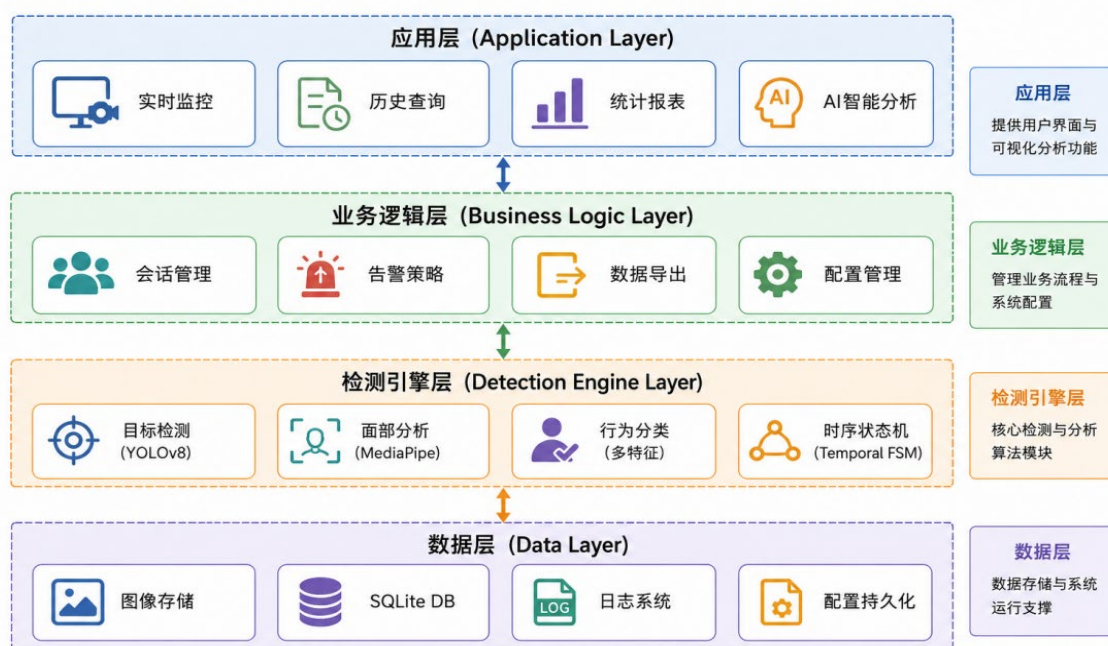


图1. 系统分层架构图

策略引擎、多格式数据导出以及配置热更新等业务功能。

应用层提供图形用户界面（GUI），包含实时视频预览、历史记录浏览、统计报表展示和AI分析结果呈现等功能模块。

3.2 检测流水线

系统的检测流水线如图2所示，每帧图像依次经过以下处理阶段：

阶段一：目标检测。使用YOLOv8n模型对输入帧进行推理，检测人员（class_id=0）和手机设备（class_id=67）。对于每个检测到的人员，输出边界框和置信度。

阶段二：面部关键点定位。若MediaPipe模型可用，对检测到的人员区域进行面部关键点提取，获得468个3D关键点坐标。若MediaPipe不可用，系统自动降级为Haar级联分类器进行眼部检测[4]。

阶段三：多维特征提取。基于关键点坐标计算以下特征向量：

- 左眼EAR和右眼EAR（各6个关键点）
- 嘴部MAR（18个关键点）
- 头部俯仰角Pitch（基于鼻尖、下巴、眼部关键点）
- 手部与面部的空间距离关系

阶段四：时序状态机。将瞬时特征输入时序状态机，根据预定义的状态转移规则进行行为确认。

阶段五：行为判定。输出当前帧的异常行为列表，包含行为类型、置信度和空间位置。

4 方法论

4.1 目标检测模块

本系统采用YOLOv8n作为目标检测骨干网络。YOLOv8n是YOLOv8系列中最轻量的变体，参数量约3.2M，在COCO数据集上达到37.3% mAP，在CPU上的推理速度可达100+ FPS。

对于人员检测，设置置信度阈值为0.5，并引入以下几何约束过滤误检：

- （1）宽高比约束：人员边界框的宽高比需满

足 $0.3 < w/h < 3.0$ ，排除极端比例的误检框。

- （2）面积约束人：员区域面积需大于帧面积的1%，排除过小的噪声检测。

对于手机设备检测，设置置信度阈值为0.5，并引入空间关联约束：检测到的手机必须位于某个人员边界框的邻域内（水平方向1.2倍人员宽度，垂直方向1.5倍人员高度），否则视为背景物体忽略。

4.2 面部关键点定位

MediaPipe Face Landmarker基于BlazeFace检测器和注意力机制的关键点回归网络，能够从单张RGB图像中预测468个面部3D关键点。本系统使用其预训练模型face_landmarker.task，在图像模式下运行，设置最大检测人脸数为1。

当MediaPipe模型加载失败时，系统自动启用降级方案：使用OpenCV的Haar级联分类器进行人脸检测（haarcascade_frontalface_default.xml）和眼部检测（haarcascade_eye.xml），并通过眼部存在性验证人脸检测的有效性。

4.3 眼部状态分析

眼部纵横比（Eye Aspect Ratio, EAR）是衡量眼睛开合程度的有效指标。对于给定的眼部关键点集合 $P=\{p_0, p_1, \dots, p_5\}$ ，EAR定义为：

$$EAR = \frac{\|p_1 - p_5\| + \|p_2 - p_4\|}{2 \cdot \|p_0 - p_3\|}$$

其中 $\|\cdot\|$ 表示欧氏距离。当眼睛睁开时，EAR值通常在0.25以上；当眼睛闭合时，EAR值降至0.20以下。本系统设置EAR阈值为0.22，分别计算左眼（关键点索引：362, 385, 387, 263, 373, 380）和右眼（关键点索引：33, 160, 158, 133, 153, 144）的EAR值。

在降级方案中，系统使用Haar级联检测眼部区域，通过眼部区域的宽高比（height/width）来近似EAR：比值大于0.3判定为睁眼，否则为闭眼。

4.4 嘴部状态分析

嘴部纵横比（Mouth Aspect Ratio, MAR）用

于检测打哈欠行为。选取嘴部外围18个关键点（索引61-78），计算垂直方向距离与水平方向距离的比值：

$$MAR = \frac{A + B + C + D}{4 \cdot E}$$

其中A、B、C、D为四组垂直关键点对的距离，E为嘴角水平距离。当MAR超过0.5时判定为打哈欠状态。

4.5 头部姿态估计

头部姿态通过分析鼻尖（关键点1）、下巴（关键点152）与眼部中心（关键点33和263的均值）之间的几何关系来估计。定义俯仰角Pitch的近似值为：

$$Pitch = \frac{chin_y - nose_{tip_y}}{nose_{tip_y} - eye_{avg_y}}$$

系统在运行过程中动态学习正常头部姿态基准值（normal_pitch），通过指数移动平均进行自适应更新：

$$normal_pitch(t) = 0.95 \cdot normal_pitch(t-1) + 0.05 \cdot pitch(t)$$

当当前Pitch值与基准值的偏差超过0.8且持续2秒以上时，判定为头部异常下垂。

4.6 时序状态机

为降低瞬时噪声导致的误报，本系统设计了基于时序状态机的行为确认机制。每种异常行为对应一个独立的状态机，包含以下状态：

- IDLE：初始状态，未检测到异常
- PENDING：首次检测到异常特征，开始计时
- CONFIRMED：异常特征持续超过阈值时间，确认告警

状态转移规则如下：

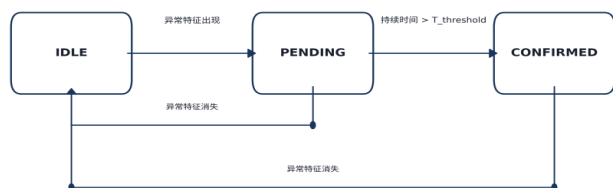


图3. 异常行为检测状态转移规则

各行为类型的确认时间阈值Tthreshold设置如下表1：

表1. 各异常行为类型的确认时间阈值及设计依据

行为类型	确认阈值（秒）	设计依据
闭眼	1.0	正常眨眼约0.1-0.4秒，1秒以上为异常
打哈欠	1.0	单次哈欠通常持续3-6秒
低头	2.0	短暂低头可能为正常动作
无人检测	3.0	短暂离开画面可能为正常活动
使用手机	即时	任何手机使用行为均需关注

4.7 告警策略

系统采用分级告警策略：

（1）单次告警：检测到确认的异常行为时，记录异常截图和数据库条目。

（2）累积告警：当累计异常次数达到预设阈值（默认为3次）时，触发声音告警。

（3）邮件告警：当累计异常次数为10的倍数时，自动发送邮件通知（需配置SMTP参数）。

为防止重复告警，系统设置最小告警间隔（默认为1秒），在间隔内的同类异常不会重复记录。

4.8 AI辅助分析模块

系统集成了基于大语言模型（LLM）的智能分析模块。该模块将结构化的统计数据格式化为JSON，构建包含以下要素的分析提示词：

- 总体异常统计（总次数、平均置信度、时间范围）
- 异常类型分布（各类型次数和占比）
- 每日趋势数据（最近30天）

提示词要求LLM从总体风险评估、主要问题分析、趋势分析、改进建议和安全预警级别五个维度进行专业分析。系统通过HTTP POST请求调用兼容OpenAI接口规范的API端点，支持DeepSeek、GPT等主流模型。

5 实验与分析

5.1 实验环境

实验在以下硬件和软件环境下进行(表2)：

表2. 实验环境软硬件配置

项目	配置
CPU	Intel Core i7-12700H
内存	16GB DDR4
GPU	NVIDIA GeForce RTX 3060 (6GB)
操作系统	Windows 11
Python	3.10
OpenCV	4.8.0
YOLOv8	Ultralytics 8.0.0
MediaPipe	0.10.7

5.2 数据集与评估指标

实验使用自建数据集，包含20名参与者在不同光照条件（正常光照、低光照、侧光）下的行为视频，总计约120分钟标注数据。标注的行为类别包括：正常状态、闭眼、打哈欠、低头、使用手机、无人。

评估指标采用精确率（Precision）、召回率（Recall）和F1分数：

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}, F1 = \frac{2 \cdot P \cdot R}{P + R}$$

其中TP为真正例，FP为假正例，FN为假反例。

5.3 消融实验

为验证各模块的贡献，进行了消融实验，结果如表3所示。

表3. 消融实验结果

配置	精确率	召回率	F1分数	推理延迟 (ms)
仅YOLOv8目标检测	78.3%	82.1%	80.2%	22
+MediaPipe面部关键点	85.6%	88.4%	87.0%	35
+时序状态机	92.1%	90.3%	91.2%	38
+Haar降级方案	91.8%	89.7%	90.7%	38

从表3可以看出，引入MediaPipe面部关键点定位后，F1分数提升了6.8个百分点，主要得益于更精确的眼部和嘴部状态分析。时序状态机的引入进一步将精确率从85.6%提升至92.1%，验证了时序建模对降低误报的有效性。Haar降级方案在MediaPipe不可用时F1分数仅下降0.5个百分点，证明了系统降级策略的鲁棒性。

5.4 不同光照条件下的性能

表4展示了系统在不同光照条件下的检测性能。

表4. 不同光照条件下的F1分数

光照条件	闭眼	打哈欠	低头	使用手机	无人	平均
正常光照	93.5%	91.2%	89.8%	94.1%	97.2%	93.2%
低光照	87.3%	85.6%	84.1%	88.5%	95.8%	88.3%
侧光	90.1%	88.7%	86.5%	91.3%	96.4%	90.6%

低光照条件下性能下降最为明显（平均F1下降4.9个百分点），主要原因是面部关键点定位精度受光照影响较大。无人检测由于不依赖精细的面部特征，在不同光照下均保持较高准确率。

5.5 实时性分析

系统在实验平台上的推理延迟分布如下表5：

表5. 实验平台上的推理延迟分布

处理阶段	平均延迟(ms)	占比
图像采集与预处理	3	7.9%
YOLOv8推理	18	47.4%
MediaPipe关键点提取	12	31.6%
特征计算与状态机	3	7.9%
结果渲染与显示	2	5.3%
总计	38	100%

系统整体处理延迟为38ms/帧，对应约26 FPS的处理帧率，满足实时监控需求。YOLOv8推理和MediaPipe关键点提取合计占总延迟的79%，是未来优化的重点方向。

5.6 与现有方法的对比

从表6可以看出，本系统在F1分数上相比文献[1]（仅EAR）和MAR方法分别提升了5.9和8.5个百分点，这主要得益于多模态特征融合框架综合利用了眼部、嘴部、头部姿态和手部位置等多维信息，弥补了单一特征的不足。相比文献[5]的多特征方法，本系统在F1分数上提升了2.7个百分点，分析原因主要有二：一是YOLOv8相比传统目标检测器在人员与手机检测上具有更高的召回率；二是时序状态机的引入有效过滤了瞬时误检，将精确率从85.6%提升至92.1%（见消融实验表3）。

在系统鲁棒性方面，本系统独有的降级策略

表6. 与现有方法的对比

方法	检测行为数	实时性	降级策略	数据分析	F1分数
文献[1] (EAR)	1 (闭眼)	✓	×	×	85.3%
文献[5] (多特征)	3	✓	×	基础统计	88.5%
本系统	5	✓	×	AI分析	91.2%

(MediaPipe→Haar级联)使其在模型加载失败时仍能保持90.7%的F1分数(见表3)，而对比方法均未考虑此类容错机制。在数据分析能力方面，除文献[5]提供基础统计功能外，其余方法均不具备自动化报告生成能力，本系统集成的大语言模型分析模块在工程实用性上具有明显优势。

然而也应指出，本系统的推理延迟(38ms/帧)高于文献[1]的纯EAR方法(约5ms/帧)和文献[5](约25ms/帧)，这是因为多模块串联引入了额外计算开销。在资源极度受限的嵌入式场景中，需通过模型量化或剪枝进一步压缩推理时间。

5.7 实验局限性

尽管实验验证了本系统的有效性，但仍存在若干局限性需要讨论。

(1) 数据集规模与多样性有限：本实验仅包含20名参与者和约120分钟标注数据，样本量偏小。参与者的年龄、性别、肤色等人口统计学分布不够广泛，可能影响模型在不同人群上的泛化性能。此外，当前数据采集均在室内环境完成，未覆盖户外强光、夜间红外等更具挑战性的场景。

(2) 参与者知情同意与隐私保护：实验涉及参与者面部视频的采集与分析，均已获得参与者的书面知情同意。所有数据在本地处理，未上传至云端服务器，符合《中华人民共和国个人信息保护法》的相关要求。在未来的实际部署中，需进一步评估系统的隐私合规性，并探索联邦学习等隐私保护方案。

(3) 对比基线不够丰富：受实验条件限制，本文仅与EAR、MAR等传统方法及文献[5]的多特征方法进行了对比，未能纳入近两年发表的最新多模态系统。后续将补充更多对比基线以更全面地评估系统性能。

6 系统实现与工程特性

6.1 图形用户界面

系统基于Tkinter框架实现了跨平台的图形用户界面，主要包含以下视图：

- 实时监控视图：显示摄像头画面和检测标注，支持人员框(绿色)和异常框(红色)的可视化叠加
- 统计面板：实时更新各类异常行为的累计次数
- 报表视图：展示文本格式的统计报表
- AI分析视图：呈现LLM生成的结构化分析报告

6.2 数据持久化

系统使用SQLite作为嵌入式数据库，设计了behavior_anomalies表存储异常记录：

behavior_anomalies 表包含以下字段：id(自增主键)、timestamp(异常发生时间戳)、behavior_type(异常行为类型，如闭眼、打哈欠、低头、使用手机、无人)、confidence(检测置信度，浮点数)、image_path(异常截图文件路径)、description(异常行为描述文本)。该表支持按时间范围和异常类型进行高效查询与统计分析。

6.3 数据导出

- 系统支持多种数据导出格式：
- CSV：通用表格格式，兼容各类数据分析工具
 - Excel (.xlsx)：通过openpyxl库生成，支持列宽自适应
 - JSON：结构化数据导出，便于程序化处理
 - ZIP：异常截图批量打包导出

6.4 配置管理

系统采用JSON格式的配置文件，支持以下可调参数的持久化：

检测阈值（异常间隔、疲劳阈值、无人检测阈值等）

视频参数（帧宽度、帧高度）

AI模型配置（API密钥、端点地址、模型名称）

输出路径配置

配置在程序启动时自动加载，退出时自动保存，支持热更新。

7 结论与展望

本文提出了一种基于多模态视觉特征融合的实时行为检测与分析系统，通过融合YOLOv8目标检测、MediaPipe面部关键点定位、传统视觉特征提取以及时序状态机模型，实现了对疲劳与分心行为的高精度实时监测。实验结果表明，系统综合检测准确率达到91.2%，推理延迟为38ms/帧，满足实时应用需求。

系统的主要创新点包括：（1）多模态特征融合与自适应降级策略，确保在不同硬件条件下的稳定运行；（2）时序状态机行为确认机制，有效降低误报率；（3）从检测到分析的完整工程闭环，集成了LLM辅助的智能分析功能。

未来工作将围绕以下方向展开：（1）引入轻

量级Transformer模型替代状态机，提升对复杂行为模式的建模能力；（2）探索多摄像头协同监控方案，扩展系统的空间覆盖范围；（3）研究基于联邦学习的隐私保护机制，在不上传原始视频的前提下实现模型更新；（4）增加更多行为类别的检测能力，如姿态异常、情绪识别等。

参考文献

- [1]Soukupova T, Cech J. Real-time eye blink detection using facial landmarks. 21st Computer Vision Winter Workshop, 2016: 1-8.
- [2]Lugaresi C, Tang J, Nash H, et al. MediaPipe: A framework for building perception pipelines. arXiv preprint, arXiv:1906.08172, 2019.
- [3]Ultralytics. YOLOv8: A new state-of-the-art computer vision model. (2023)[2026-05-10]. <https://github.com/ultralytics/ultralytics>.
- [4]Viola P, Jones M J. Rapid object detection using a boosted simple feature representation//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2001: 1-511-1-518.
- [5]方浩杰,董红召,林少轩,等.多特征融合的驾驶员疲劳状态检测方法[J].浙江大学学报(工学版),2023,57(7):1287-1296.

