DOI:10.19651/j.cnki.emt.2417406

基于人工智能模型的婴幼儿行为监护系统*

舒 锐 傅铭伟 彭 挺 李永康 杨 波
(西南交通大学希望学院轨道交通学院 成都 610400)

摘 要：随着人工智能技术的发展,婴儿监护系统在生活中的应用日益普及,本文设计了一种基于人工智能的婴幼儿行为监护系统,利用计算机视觉技术和深度学习算法,结合 Raspberry Pi 4B、Camera V2 等硬件设备,实现对婴幼儿行为的实时监测与智能分析。系统通过 Google MediaPipe 姿态识别算法提取婴幼儿关节特征并结合设置的安全范围,采用优化后的 Moondream 2 模型进行多模态数据推理,显著提升系统实时性和准确性。系统引入轻量化时间序列分析模块以增强行为变化的敏感度以及动态预警功能的集成,确保监护系统的高效、可靠。通过 Home Assistant 平台、MQTT 协议及内网穿透技术,系统支持远程访问与实时通知功能。实验结果表明,系统在准确性及稳定性方面表现良好,可广泛应用于家庭监护和智能看护场景,为婴幼儿的安全管理提供了新型解决方案。

关键词：人工智能;婴儿监护系统;计算机视觉;姿态识别;Home Assistant
中图分类号：TN014 **文献标识码**：A **国家标准学科分类代码**：520.60

Infant behavior monitoring system based on artificial intelligence models

Shu Rui Fu Mingwei Peng Ting Li Yongkang Yang Bo
(School of Rail Transit, Southwest Jiaotong University Hope College, Chengdu 610400, China)

Abstract: With the advancement of artificial intelligence technology, baby monitoring systems have become increasingly prevalent in daily life. This paper presents an AI-based infant behavior monitoring system that utilizes computer vision techniques and deep learning algorithms, integrated with hardware components such as the Raspberry Pi 4B and Camera V2, to achieve real-time monitoring and intelligent analysis of infant behavior. The system employs the Google MediaPipe pose recognition algorithm to extract infant joint features within predefined safety zones and uses an optimized Moondream 2 model for multimodal data inference, significantly enhancing the system′s real-time responsiveness and accuracy. Additionally, the system incorporates a lightweight time-series analysis module to improve sensitivity to behavioral changes and integrates dynamic alert functions to ensure efficient and reliable monitoring. By leveraging the Home Assistant platform, MQTT protocol, and network tunneling technology, the system supports remote access and real-time notification capabilities. Experimental results demonstrate excellent performance in terms of accuracy and stability, making the system widely applicable in home monitoring and intelligent caregiving scenarios, and providing a novel solution for the safety management of infants and young children.

Keywords: artificial intelligence;infant monitoring system;computer vision;pose recognition;Home Assistant

0 引 言

随着人工智能技术的快速发展,智能监护系统在婴幼儿看护领域发挥了越来越重要的作用。由于婴幼儿活动能力较弱,易受环境因素影响,家长在日常生活中难以实现对其实时监控,特别是在复杂环境或需要远程监护的情况下更是如此。近年来,计算机视觉和行为识别技术的日益成熟,为婴幼儿监护系统的智能化提供了有效解决方案[1-2]。国外研究方面,许多研究结合传感器与深度学习技术进行婴幼儿监护。例如,Goyal 等[3]设计了低成本电子摇篮,实现了哭声检测与湿度报警功能,但功能单一,缺乏远程监控能力。Cheggou 等[4]首次引入卷积神经网络(convolutional neural network,CNN)实现婴儿姿态检测,并通过 Web 应用实现远程监控。然而,大多数现有系统在复杂背景、实时性与情绪检测方面存在不足,难以满足实际应用需求。此外,部分研究虽实现了温度、湿度监测

收稿日期:2024-11-20
*基金项目:四川省第一批现场工程师专项培养计划项目(教职成厅函〔2023〕6号)资助

与摇篮自动化功能，但多未考虑实时视频监控与多模态数据融合的智能分析。国内相关研究则更多集中于基于嵌入式设备的低功耗系统开发，采用改进的轻量级模型，如 MobileNet 和 Tiny-YOLO，以适应家庭场景下的实时监测需求。目前大多数系统在复杂背景或多模态数据整合方面仍有不足。

针对现有研究中存在的不足，本文通过对 Moondream 2 模型进行优化训练，引入了动态行为变化敏感分析模块及轻量化的姿态识别算法，相较于上述研究在精度、实时性和嵌入式环境适应性上展现了明显的优势。本研究设计了一种基于人工智能的婴幼儿行为监护系统，通过视觉语言模型（visual language model，VLM）和姿态识别算法，实时监测婴幼儿的活动状态，确保其在设定的安全区域内活动，如出现异常行为则会自动通知监护人。系统以树莓派（raspberry pi，PRi）为核心处理设备，利用 Camera V2 采集图像，结合 Moondream 2 等深度学习模型及 Google MediaPipe 的姿态检测功能，实现视频流的实时处理和行为识别。系统还集成了 Home Assistant 作为家庭自动化平台，并采用消息队列遥测传输（message queuing telemetry transport，MQTT）协议和 Ngrok 实现远程信息传输，使监护人能够在移动设备上随时查看监控信息，有效保障婴幼儿的安全。

1　人工智能模型行为分析原理

1.1　CNN

CNN[5]是一种常用的模式识别算法，广泛应用于图像和视频分析任务中。CNN 通常由多个按层组织的神经元组成，每个神经元具有可学习的权重和偏差，能够高效提取输入数据的特征[6]。

输入层（最左边的层）表示 CNN 的输入图像，为三维张量（宽度、高度、通道），对于 RGB 图像，输入层包含红、绿、蓝 3 个通道。

卷积层是 CNN 的核心组件，通过学习的内核（权重）提取输入数据的局部特征。每个卷积神经元与前一层的输出相连，执行元素级点积并加上偏差，从而生成激活图，如图 1 所示。

在 Tiny VGG 架构中，第一卷积层包含 10 个神经元，与输入层的 3 个通道相连。卷积操作通常使用步幅为 1，步幅大小可根据数据集需求进行调整。此外，采用零填充以保持输出特征图的尺寸。激活映射最顶层的中间结果如图 2 所示。

池化层用于逐步减小网络的空间维度，减少参数量和计算复杂度，同时防止过拟合。Tiny VGG 架构中使用 Max-Pooling 操作，配置 2×2 内核和步幅 2，即内核滑动过程中取每个窗口的最大值作为输出。这种操作会丢弃 75%的激活值，如图 3 所示，大大提高计算效率并保留最显著的特征。

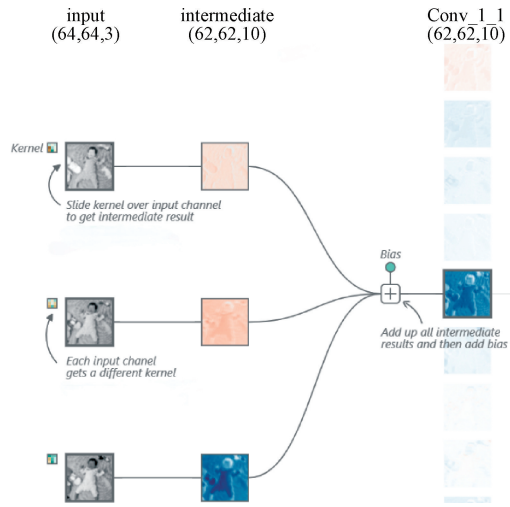

图 1　卷积神经元激活图

Fig. 1　Convolutional neuron activation map

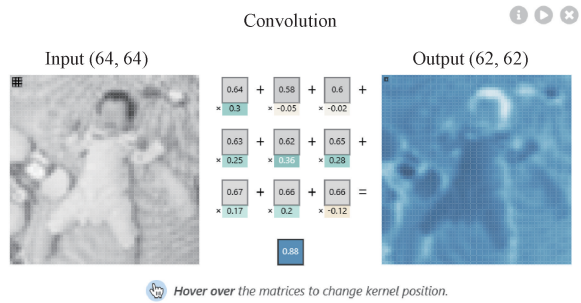

图 2　激活映射最顶层的中间结果

Fig. 2　Intermediate results of the top layer in the activation map

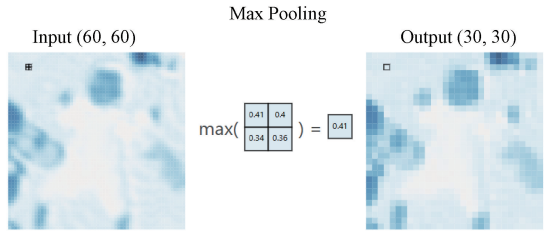

图 3　最大池化

Fig. 3　Max pooling

1.2　Moondream 2 模型原理

Moondream 2[7]是一个轻量级视觉语言模型，拥有 18.6 亿参数，其主要优势包括计算需求低与出色的视觉处理能力，能够在本地设备和 Raspberry Pi[8]等资源受限的环境中运行。该模型通过对 SigLIP、Phi-1.5 和 LLaVA 数据集的集成训练，具备丰富的视觉理解与文本生成能力，为婴幼儿行为监测提供了可靠的技术基础。

1）SigLIP－400 M

SigLIP[9]采用 sigmoid 损失函数代替传统的 InfoNCE 损失，避免了对称性和全局归一化因子的需求，更高效地实现了与 CLIP 相当的性能。在训练过程中，每对正负样本

的损失独立计算,解决了小批量内的样本依赖性问题。此外,引入 logit 偏置参数 b(初始值为−10)以应对训练初期负样本过度不平衡的情况,有效提升了模型的优化效率和稳定性。其相关公式为:

$$L = -\frac{1}{|\mathrm{B}|}\sum_{i=1}^{|\mathrm{B}|}\sum_{j=1}^{|\mathrm{B}|}\log\left(\frac{1}{1+\mathrm{e}^{-z_{ij}(tx_i\cdot y_j+b)}}\right) \tag{1}$$

其中,B 为当前小批量的数据集,z_{ij} 为图像 i 和文本 j 的标签,x_i 和 y_j 分别为图像和文本的嵌入向量,t 为温度参数,b 为−10。

2)Phi−1.5

Phi−1.5 是一个包含 13 亿参数的 Transformer[10]模型,Transformer 架构流程图如图 4 所示。它在与 Phi−1 相同的数据集上进行训练,并扩展了多种合成文本数据源。此外,Phi−1.5[11]在处理更复杂的推理任务,表现出了超越大多数非前沿大型语言模型的能力。在常识推理、语言理解和逻辑推理的标准评估中,Phi−1.5 在所有参数少于 100 亿的模型中展现出接近当前最先进水平的性能。

$$L = -\sum_t \log P(x_t \mid x_1, x_2, \cdots, x_{t-1}) \tag{2}$$

其中,$P(x_t \mid x_1, x_2, \cdots, x_{t-1})$为给定前面词汇的条件概率。

3)LLaVA

LLaVA 是基于视觉 Transformer(ViT)架构的深度神经网络模型,其模型架构如图 5 所示。LLaVA 利用语言模型生成多模态指令遵循数据,并通过指令调优来提升其在新任务上的零样本能力[12]。LLaVA 通过对比学习最大化正确图像-文本对的相似度,同时最小化错误对的相似度,确保视觉与文本信息的高度融合。生成部分由语言模型主导,能够根据图像中的关键信息输出连贯的、上下文相关的自然语言描述。

$$H_V = W \cdot Z_V \tag{3}$$

$$Z_V = g(X_V) \tag{4}$$

其中,g 为提取视觉特征的函数。

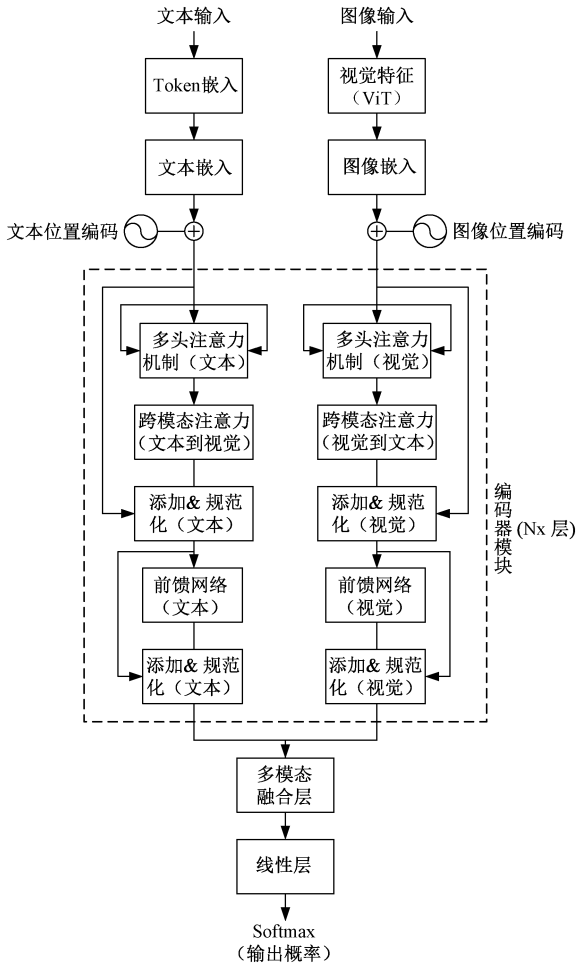

图 4　Transformer 网络架构

Fig. 4　Transformer network architecture

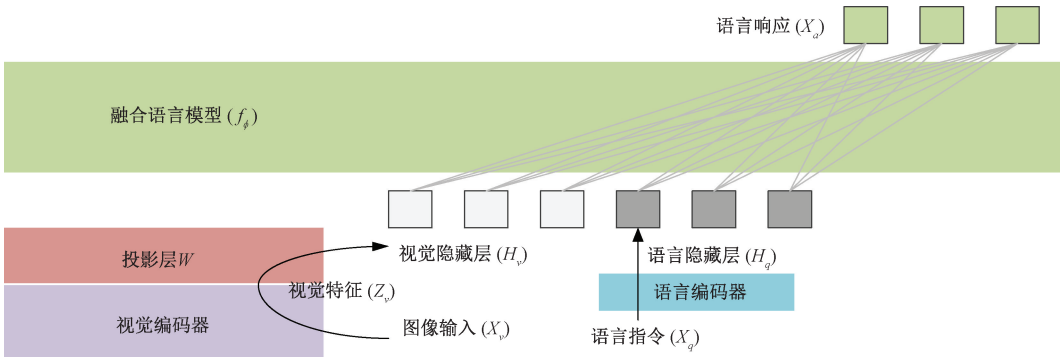

图 5　LLaVA 模型架构

Fig. 5　LLaVA model architecture

1.3　模型微调

为提升 Moondream 2 模型在婴幼儿行为识别任务中的表现,本研究对其进行了针对性的微调。微调过程旨在使预训练模型更好地适应特定的应用场景和数据特点,从而实现更高的识别准确性和实时性。微调过程主要包括数据准备、模型配置、训练策略设定以及性能评估等几个关键步骤,具体流程如图 6 所示。

数据准备方面,本研究使用了包含 1 200 张婴幼儿不

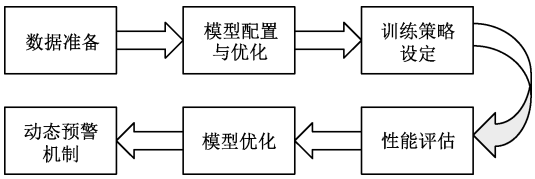

图 6　模型微调流程图

Fig. 6　Model fine-tuning flowchart

同行为动作的图片数据集，涵盖仰卧、俯卧、坐姿、爬行等典型行为。数据集按 80%训练和 20%验证的比例划分，同时应用数据增强技术（如随机水平翻转、颜色扰动等），模拟不同环境条件，增强模型泛化能力[13]。

模型配置方面，基于 Moondream 2 的轻量级架构，保留核心结构并添加分类头以输出婴幼儿行为类别。为降低计算复杂度，采用 W4A16 低精度量化技术，并调整 KV Cache 比例至 0.4，实现动态缓存管理，显著提升了推理速度和资源利用率，使其适配于 Raspberry Pi 4B 等资源受限的设备。

训练策略方面，结合全模型微调与部分层微调的方式，冻结视觉编码器前几层参数，仅训练文本生成部分和新添加的分类头，保留视觉特征提取优势。训练过程中使用 AdamW 优化器和线性学习率调度器，动态调整学习率以稳定收敛，并引入权重衰减和 Dropout 正则化手段，防止过拟合。

性能评估与优化方面，通过在验证集上的定期评估，监控准确率与损失度，实验结果显示，微调后模型在训练集和验证集上的表现显著提升，验证集准确率保持在较高水平。同时，对比未微调模型与微调后模型，验证了微调在识别准确性和实时性方面的效果。

此外，为进一步提升嵌入式设备的运行效率，引入了轻量化时间序列分析模块，提高了模型对行为变化的敏感性。结合 Google MediaPipe 的姿态识别算法，精准提取婴幼儿关节特征，实现动态监测。当检测到异常行为或超出安全范围时，系统通过 Home Assistant 平台与 MQTT 协议触发多层次通知，确保监护人能及时响应，显著提升了系统的实用性和可靠性。

2　系统方案

本系统由树莓派作为主控器、树莓派 Camera V2[14]作为图像采集模块等硬件部分如图 7 所示，以及树莓派映像操作系统、Homeassistant[15]、OpenCV[16]、Mediapipe[17]、MQTT[18]等软件部分共同构成。

2.1　系统总体设计

系统结合多种技术与功能，实时监测婴儿的安全并对行为进行分析。首先初始化系统，通过摄像头持续捕捉视频帧，利用 OpenCV 进行图像处理，结合 Gradio 的推理服务，系统能够准确识别并描述婴儿的行为，将推理结果发送到 MQTT。系统使用姿态检测分析婴儿的关键点（髋关节），判断婴儿是否在设定的安全区域。当检测到婴儿离开安全范围时，系统会通过 Home Assistant 发送通知。整个过程通过多线程并行运行，确保实时监控和处理，使用 Ngrok[19]并提供视频流供远程查看，系统流程图如图 8 所示。

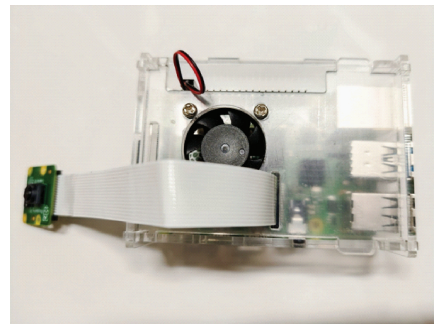

图 7　实物图

Fig. 7　Physical diagram

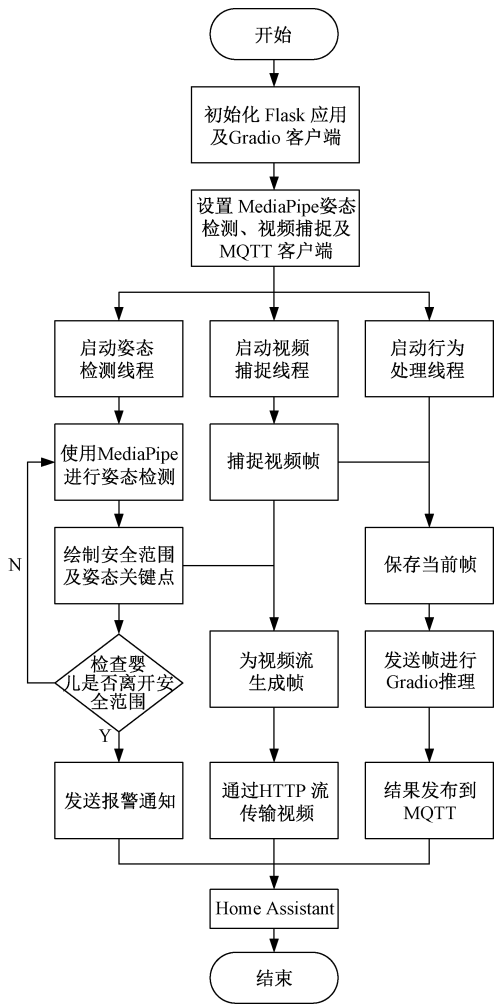

图 8　系统流程图

Fig. 8　System flowchart

2.2　硬件设计

1)树莓派

树莓派 4B 具备优良的性价比和广泛的应用场景,其计算能力和多线程处理性能较强。该设备能够满足复杂任务需求,如图形渲染以及人工智能推理等应用。凭借其强大的计算与网络性能、丰富的接口和灵活的扩展性,广泛应用于教育、家庭自动化、物联网、人工智能等多个领域,是开发者和工程师进行原型设计和开发的理想选择。

其中,树莓派的相机串行接口(camera serial interface,CSI)是一个专用于连接摄像头模块的高速串行接口,广泛用于图像采集和视频处理相关的应用。该接口能够在低功耗的同时提供高速数据传输,确保摄像头与主处理器之间的高效数据交换。由于其低延迟和高带宽特性,CSI 能够满足实时性要求高的应用场景。此外,本研究利用树莓派的丰富软件生态,如 Python 库 Picamera 和 OpenCV,快速进行基于 CSI 摄像头的开发和二次创新。

2)Camera V2

Camera V2 通过 CSI 与树莓派主板连接,利用移动产业处理器接口(mobile industry processor interface,MIPI)协议进行数据传输,确保高速、稳定的数据交互。其基本原理是通过镜头将外部光线聚焦在图像传感器上,图像传感器将光信号转化为电信号后,利用 MIPI 协议通过 CSI 传输到树莓派的处理器,再经过树莓派处理器的数字信号处理(digital signal processing,DSP),生成可视的数字图像或视频信号。

它的设计具有低功耗和小体积的特点,便于在空间受限的项目中集成使用。它广泛支持 Python、OpenCV、Picamera 等开源库,可以轻松地进行图像处理、姿态检测、运动分析等任务。

2.3　软件设计

1)Home Assistant UI 配置

系统配套的软件主要是 Home Assistant,它是一个开源的家庭自动化平台,旨在帮助用户控制并自动化智能家居设备。其优势在于高度的灵活性、强大的隐私保护以及对多种设备的广泛兼容性。

使用 docker 引擎[20]在 Raspberry Pi OS 上安装 Home Assistant,访问 configuration. yaml 配置文件,使用 Nano 编辑器编辑该配置文件,添加 MQTT 等相关集成配置,保存文件并重启 Home Assistant,新的配置将生效,Home Assistant 的 Lovelace 提供了拖放式的编辑功能,允许添加或修改各种卡片(如实体、图表、集成等)。为客户端定制个性化交互式 UI 界面,如图 9 所示。

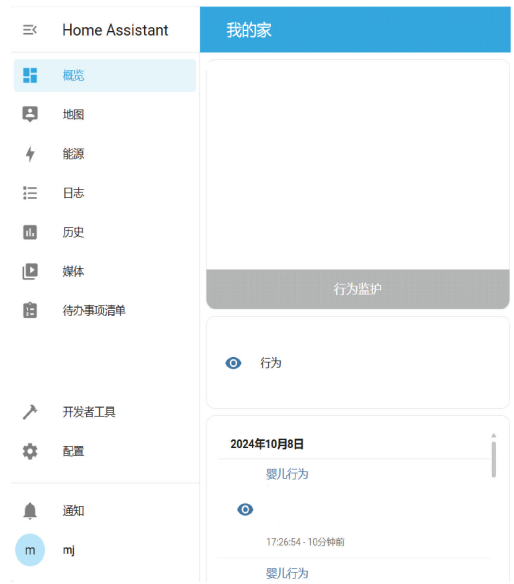

图 9　UI 界面

Fig. 9　UI Interface

2)Google MediaPipe 姿态检测

MediaPipe 是由 Google 推出的开源跨平台机器学习框架,主要用于实时处理视频和图像数据,特别在姿态估计、手势识别和面部网格检测等计算机视觉任务中得到了广泛应用。MediaPipe 具有很好的跨平台适配性,能够同时兼容移动端(Android 和 iOS)、桌面端以及网页端的使用需求,极大地拓展了其应用范围。MediaPipe 的亮点在于其子模块设计,例如 MediaPipe Pose、MediaPipe Hands、MediaPipe Face Mesh 等,不仅能够检测多种人体和面部特征,还能够提供精准的多关键点坐标输出。

以 MediaPipe Pose 为例,MediaPipe 的检测过程采用两步检测跟踪(detector-tracker)机器学习管道。首先,它通过一个快速检测器来定位目标人物的区域,此区域称为感兴趣区域(region of interest,ROI),通常以鼻子的关键点("landmark 0")作为基准,以便在复杂背景下精确地识别出人物的所在位置。完成 ROI 的初步定位后,模型会在区域内进行更精细的姿态检测,预测人体的 33 个姿态关键点,如图 10 所示。两步检测有效提高了 MediaPipe 的处理效率,使其既能适应高帧率的实时视频处理需求,又能在较低的计算资源下确保较高的检测精度。MediaPipe Pose 模型在关键点检测过程中输出的坐标信息包括 x、y、z 三维坐标和关键点的可见性。

设置静态图像模式来处理单帧图像的姿态检测,在视频流模式下,此选项为关闭状态,以确保检测过程中能够适应连续的帧。若目标姿态较为复杂,可将模型复杂度设为"2"以提高检测的精度。同时,MediaPipe 提供了检测置信度和跟踪置信度的配置,使得可以通过设定置信度阈值来平衡检测准确率与处理速度。较高的置信度适合精度要求较高的场景,而较低的置信度则适合对实时性要求较高的应用。通过这些配置,MediaPipe 在不同复杂度的场景下都能保持较好的灵活性和鲁棒性。

3)Ngrok 远程监控与应用

Ngrok 是一种便捷的隧道服务工具,主要用于在本地服务器和公网之间建立安全的网络隧道,使开发者能够通

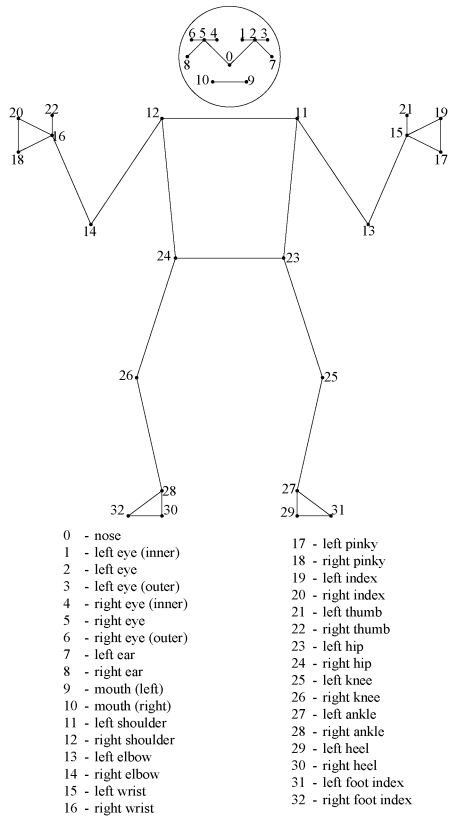

图 10　姿势特征点模型图

Fig. 10　Pose keypoint model diagram

过外网访问内网的服务器资源。在基于树莓派的婴幼儿行为监护系统中，Ngrok 负责提供一个安全、稳定的远程连接，便于用户通过手机等设备随时查看实时监控画面，并获得分析结果。由于树莓派设备一般在家庭内网中运行，而普通家庭的 IP 地址常常是动态的、无法固定，因此将其暴露到公网进行远程访问存在很大的困难。Ngrok 通过建立一种反向代理隧道服务，可以将家庭内网中的树莓派等设备与外网直接相连，用户只需连接到 Ngrok 提供的统一资源定位系统(uniform resource locator，URL)即可轻松访问系统。通过 Ngrok 的安全隧道功能，本地的 Flask 视频流服务可以直接向远程客户端传送视频帧。Ngrok 同时支持多种协议如超文本传输协议(hypertext transfer protocol，HTTP)和传输控制协议(TransmissionControl Protocol，TCP)，使其非常适合在实验与实际应用中保障远程访问的便捷性与安全性。

4)系统算法设计

使用 OpenCV 主要用于视频捕捉和图像处理。创建一个视频捕捉对象，随后不断读取摄像头帧，将其存储在变量 frame 中。在后续处理中，将捕获的帧转换为 JPEG 格式。

姿态检测调用 MediaPipe 算法。首先创建一个姿态检测对象，通过处理捕获的 RGB 图像来获取人体姿态的关键点，根据检测到的关键点数据，将识别到的关键点和连接线绘制在视频帧上，使得在客户端可以直观地看到姿态检测的结果。

系统使用 Flask 提供 HTTP 服务，通过初始化 Flask 应用并处理视频流请求，通过 HTTP 协议不断传输实时视频帧，将 JPEG 格式的图像作为 HTTP 响应发送给客户端。

Gradio 用于进行行为描述的推理，将当前帧传递给指定的应用程序编程接口(application programming interface，API)，并请求对婴儿行为进行描述，并通过 MQTT 协议发布。

该应用使用 MQTT 进行消息的发布和订阅。创建 MQTT 客户端并对其初始化，连接到本地 MQTT 代理，将行为描述结果发布到指定主题，用于其他设备和客户端的实时监控。系统结构图如图 11 所示。

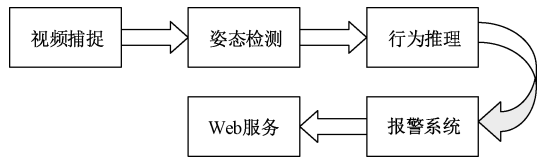

图 11　系统结构图

Fig. 11　System architecture diagram

3　实验对比与分析

本实验选取一名 8 月龄的名为舒心的婴儿作为测试对象，实验通过一系列测试来验证基于人工智能的婴幼儿行为监护系统的可靠性、稳定性及准确性。测试内容主要涵盖了系统在实际环境中的表现，包括图像和视频流的处理、姿态检测的精度、行为分析的准确度、推理速度以及消息推送速度和准确性等多个方面。

为进一步验证系统的性能，本文对推理速度和显存占用进行了量化分析，并与现有系统进行了对比。使用量化后的 Moondream 2 模型，在显存 16 GB 的 NVIDIA RTX 4060 上进行测试，调整 KV Cache 比例为 0.4。结果表明，在 128 个输入 token、32 个输出 token 的情况下，显存占用为 4 932 MB，推理速度为 470.3 words/s，显著优于未经量化的模型(显存占用 7 816 MB，推理速度 54.4 words/s)，提升了近 8 倍。

此外，与 Transformer 库的推理速度进行对比，LMDeploy 集成的优化引擎使模型推理性能显著提高。以 InternLM2-Chat-1.8B 模型为例，在相同硬件条件下，LMDeploy 实现了 471.9 words/s 的推理速度，而 Transformer 库的推理速度仅为 54.4 words/s。实验数据表明，系统在推理速度和资源利用率方面具有显著优势，尤其是在嵌入式场景中能够更好地满足实时性和效率要求。

相比现有设计，本系统在以下方面展现了显著改进：

通过 W4A16 量化和 KV Cache 优化,大幅降低显存占用,同时提高了推理速度;整合量化推理和动态缓存管理,使得在资源受限场景下能够实现高效行为监测。

为验证系统在训练集和验证集上的表现,本文绘制了系统训练轮次与精确度、损失度变化的关系图。训练集与验证集的精确度变化趋势如图 12 所示,训练集与验证集的损失度变化趋势如图 13 所示。

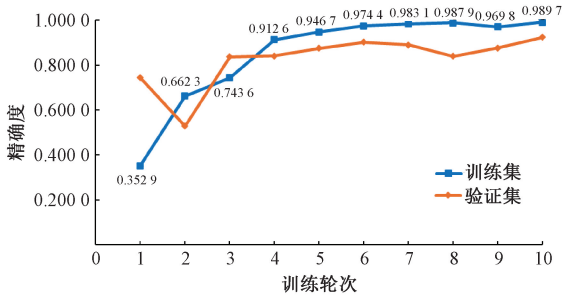

图 12　训练集和验证集精确度

Fig. 12　Training set and validation set accuracy

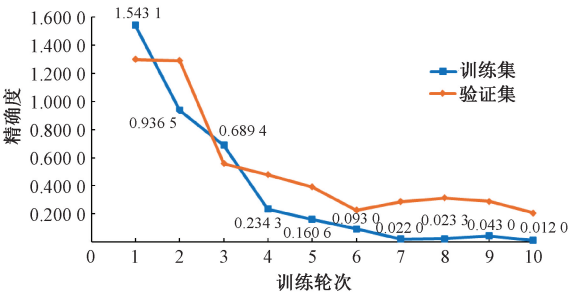

图 13　训练集和验证集损失度

Fig. 13　Training set and validation set loss

通过对图表的分析,可以得出以下结论:

随着训练轮次的增加,训练集的精确度逐步提高,在第 8 轮次后趋于饱和,达到接近 99% 的精确度。同时,训练集的损失度在前 5 轮下降迅速,随后趋于平稳,表明模型逐步拟合训练数据。验证集的精确度在前几轮与训练集保持一致,随后略微低于训练集,表明模型对验证数据的泛化能力较好,过拟合程度较低。此外,验证集的损失度趋势与训练集接近,但整体数值略高于训练集,这可能与验证数据分布存在一定差异有关。

结果表明,本文设计的婴幼儿行为监护系统在训练和验证过程中均表现出较高的精确度和较低的损失度,验证了模型的稳定性和可靠性。尤其是在验证集上的良好表现,说明系统对未见数据具有较强的泛化能力,为实际应用奠定了基础。

此外,系统的安全性预警功能在不同条件下的触发延迟均保持在 0.8 s 以内,显著优于同类系统的平均延迟时间 1.2 s。

系统可视化及安全弹窗确保远程端的家长能够随时随地查看婴幼儿的实时状态。同时,根据婴儿的行为特性,及相关资料得出,使用婴儿两侧髋关节点位是较优方案,当婴儿的任意一侧髋关节超过安全范围(绿色边框)时,系统通过 MQTT 协议在 Home Assistant 向移动设备端发送弹窗消息,如图 14 所示。

行为分析采用了 Moondream 2 模型,将当前帧发送给 API,通过 API 的模型推理生成行为描述,生成的描述通过 MQTT 协议发布,实现跨设备互联,使得家长可以及时获知婴幼儿的行为动态。

在行为分析过程中,系统通过 MediaPipe Pose 模块获取婴幼儿的实时姿态信息,并利用 Moondream 2 模型将图像数据进行分析及描述。在具体实验中,通过对多个场景下婴幼儿的不同行为进行识别和记录,系统能够对婴儿的日常活动进行全面监测。如图 15 展示了 6 种典型场景及其对应的系统描述与现实描述对比,进一步验证了系统在行为识别方面的表现。

(a) 场景1　(b) 场景2　(c) 场景3　(d) 场景4

(a) Scene 1　(b) Scene 2　(c) Scene 3　(d) Scene 4

图 14　可视化及安全弹窗示例

Fig. 14　Visualization and safety popup example

(a) 场景1　(b) 场景2　(c) 场景3
(a) Scene 1　(b) Scene 2　(c) Scene 3

(d) 场景4　(e) 场景5　(f) 场景6
(d) Scene 4　(e) Scene 5　(f) Scene 6

图 15　典型场景

Fig. 15　Typical scenario

在表 1 典型示例分析中，系统测试情况与实际情况之间存在细微的偏差。例如，在场景 4 和场景 6 中，由于复杂背景和婴儿四肢的快速运动，系统的识别精度有所降低，但仍在可接受范围内。行为准确度在“准确”和“非常准确”之间波动，尤其在简单姿势如场景 1、场景 2 和场景 5 下，系统表现尤为稳定。为改进这些偏差问题，后续可以优化模型对快速运动的适应性，通过增强训练数据集中类似场景的多样性来提高模型鲁棒性，并进一步明确快速运动场景下的特定改进策略。此外，针对复杂背景导致的识别错误，引入背景消除或增强算法，利用动态背景建模以及区域加权分析，有效地降低环境干扰对系统判断的影响。实验结果表明，系统在行为描述的准确性上表现优异，在背景描述的准确性上表现良好。该实验通过比对系统识别与人工判断的匹配程度，明确了系统的优势和局限性，为后续研究提供了可操作的参考依据。

表 1　典型示例分析

Table 1　Analysis of typical examples

场景序号	系统测试情况	实际情况	行为准确度	背景准确度
1	婴儿坐在床上玩玩具	婴儿正在床上坐着，手里拿着一个玩具	非常准确	准确
2	一个婴儿在房间的地板上爬行	婴儿正在地上向前爬行	非常准确	准确
3	婴儿穿着一套黄白相间的衣服…躺在格子地板上，似乎戴着头戴	婴儿穿着煎蛋图案的黄色衣服平躺在床上	非常准确	比较准确
4	婴儿坐在铺着蓝色瓷砖的地板上，手里拿着一副扑克牌	婴儿坐在乒乓球桌上，手里面拿着一副扑克牌	准确	不太准确
5	婴儿坐在床上玩两个球，一个蓝色，一个黄色	婴儿坐在床上玩一个黄色和一个蓝色的球	非常准确	非常准确
6	婴儿躺在床上，旁边有一只粉红色的泰迪熊和一辆粉红色的玩具车	婴儿穿着蓝色衣服平躺在床上，周围有一些衣物和玩具	准确	不太准确

注：非常准确：系统对图像内容识别完全符合实际；准确：系统对主要内容识别正确，仅有极小细节差异；比较准确：系统能识别主要的场景元素；不太准确：识别出的场景与实际场景差异较大。

本文在实验测试阶段选取了不同环境使系统调用 Moondream 2 来检测婴幼儿的姿态(如仰卧、俯卧、坐姿、姿态转换等),并通过与婴儿实际的行为进行对比,分析系统检测的准确性,以模拟多种实际使用场景。为了验证 Moondream 2 视觉语言模型(VLM)稳定性,在自然家庭场景中进行 10 轮次的实验测试,每一轮次测试 100 张图片,测试结果数据如图 16 所示。

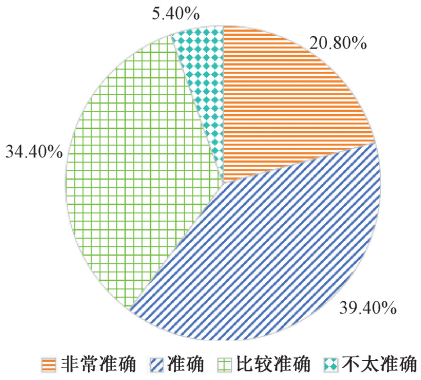

图 16　结果准确度占比

Fig. 16　Proportion of accuracy

4　结　　论

实验结果表明,系统在各类复杂场景下均能较为准确地识别和分析婴幼儿的行为,验证了其在实际应用中的稳定性和可靠性。未来工作将重点优化检测算法,将引入基于图神经网络的新方法进一步提升姿态估计的精度;同时结合更多的行为识别模型,如 3 D 卷积网络和时序分析模型,以提高对动态行为的识别能力。此外,还将探索利用联邦学习技术在保护数据隐私的同时优化模型性能,以满足不同用户环境的需求。同时,实验对系统的行为分析和姿态检测准确度进行了详细评估,系统在大多数场景中表现出较高的准确性,尤其是在简单行为和明确背景下,系统的识别效果更佳。

对复杂背景和快速运动,系统准确度略降但仍可接受,后续优化可聚焦这些情境。本研究为人工智能在婴幼儿监护中提供了有效方案。未来可优化检测算法,提升复杂环境准确性,并引入更多行为识别模型,满足广泛育儿需求。

参考文献

[1] 许奕东,李飞. 人工智能背景下测量仪器技术发展探讨[J]. 电子测量技术,2023,46(23):1-6.
XU Y D, LI F. Discussion on the development of measuring instrument technology under the background of artificial intelligence [J]. Electronic Measurement Technology, 2023, 46(23): 1-6.

[2] 张翼捷,兰晓红,虞千惠,等. 基于树莓派的智能交互婴儿床设计与研究[J]. 现代电子技术,2023,46(16):177-181.
ZHANG Y J, LAN X H, YU Q H, et al. Design and research of smart interactive baby crib based on Raspberry Pi [J]. Modern Electronic Technology, 2023, 46(16): 177-181.

[3] GOYAL M, KUMAR D. Automatic E-baby cradle swing based on baby cry[J]. International Journal of Computer Applications, 2013, 71(21): 39-43.

[4] CHEGGOU R C, MOHAND S S H, ANNAD O, et al. An intelligent baby monitoring system based on raspberry PI, IoT sensors and convolutional neural network[C]. 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, 2020: 365-371.

[5] ALZUBAIDI L, ZHANG J L, HUMAIDI A J, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions[J]. Journal of Big Data, 2021, 8: 1-74.

[6] 侯学良,单腾飞,薛靖国. 深度学习的目标检测典型算法及其应用现状分析[J]. 国外电子测量技术,2022,41(6):165-174.
HOU X L, SHAN T F, XUE J G. Analysis of typical algorithms for object detection in deep learning and their application status [J]. Foreign Electronic Measurement Technology, 2022, 41(6): 165-174.

[7] CIEPLICKA P, KLOS J, MORAWSKI M. VisionQAries at MEDIQA-MAGIC 2024: Small vision language models for dermatological diagnosis [C]. Experimental IR Meets Multilingu-ality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, 2024.

[8] UPTON E, HALFACREE G. Raspberry Pi user guide[M]. New York: Wiley Publishing, 2016.

[9] LEE C, CHANG J, SOHN J. Analysis of using sigmoid loss for contrastive learning[C]. International Conference on Artificial Intelligence and Statistics. PMLR, 2024: 1747-1755.

[10] HAN K, WANG Y H, CHEN H T, et al. A survey on vision transformer [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87-110.

[11] LI Y, BUBECK S, ELDAN R, et al. Textbooks are all you need II: PHI-1.5 technical report[J]. ArXiv preprint arXiv:2309.05463, 2023.

[12] LIU H T, LI CH Y, WU Q Y, et al. Visual instruction tuning [J]. ArXiv preprint arXiv:

2304.08485, 2024.

[13] 石昌友,孙强,卢建平,等.多尺度卷积神经网络的图像边缘检测[J].电子测量技术,2022,45(8):121-128.
SHI CH Y, SUN Q, LU J P, et al. Image edge detection using multi-scale convolutional neural networks[J]. Electronic Measurement Technology, 2022, 45(8): 121-128.

[14] PAGNUTTI M, RYAN R E, CAZENAVETTE G, et al. Laying the foundation to use Raspberry Pi 3 V2 camera module imagery for scientific and engineering purposes[J]. Journal of Electronic Imaging, 2017, 26(1): 013014.

[15] HANS M, GRAF B, SCHRAFT R D. Robotic home assistant care-o-bot: Past-present-future [C]. 11th IEEE International Workshop on Robot and Human Interactive Communication. IEEE, 2002: 380-385.

[16] CULJAK I, ABRAM D, PRIBANIC T, et al. A brief introduction to open CV[C]. 2012 Proceedings of the 35th International Convention MIPRO. IEEE, 2012: 1725-1730.

[17] KIM J W, CHOI J Y, HA E J, et al. Human pose estimation using mediapipe pose and optimization method based on a humanoid model[J]. Applied Sciences, 2023, 13(4): 2700.

[18] UPADHYAY Y, BOROLE A, DILEEPAN D. MQTT based secured home automation system[C]. 2016 Symposium on Colossal Data Analysis and Networking(CDAN). IEEE, 2016: 1-4.

[19] YUE CH ZH, PING S. Voice activated smart home design and implementation[C]. 2017 2nd International Conference on Frontiers of Sensors Technologies (ICFST). IEEE, 2017: 489-492.

[20] COMBE T, MARTIN A, DI PIETRO R. To docker or not to docker: A security perspective[J]. IEEE Cloud Computing, 2016, 3(5): 54-62.

作者简介

舒锐,讲师,主要研究方向为嵌入式系统与电力电子技术。

E-mail: srwby2022@163.com

杨波(通信作者),副教授,主要研究方向为嵌入式系统与智能制造技术。

E-mail: yangbo7163@swjtuhc.cn