基于图像识别的问诊手语辅助系统设计与开发

张茹
陕西国际商贸学院，中国 ·陕西 咸阳 712046

摘要：针对听障人士就医沟通障碍问题，本文设计并实现了一种基于 YOLOv12 的手语识别辅助系统。该系统结合 MediaPipe 手部关键点提取与 YOLOv12 模型，实现对医疗场景下手语手势的实时识别。基于 CSL-500 数据集构建医疗手语数据集，实验表明系统识别准确率达 94.2%，精确率 93.8%，召回率 94.1%。系统采用前后端分离架构，支持实时视频采集、文字显示与语音播报，为听障人士与医护人员的沟通提供了有效的智能辅助工具。
关键词：图像识别；手语识别；YOLOv12；智慧医疗

Design and Development of Interrogation Sign Language Assistant System Based on Image Recognition

Zhang Ru
Shaanxi International Business College, China Shaanxi Xianyang 712046
Abstract: Aiming at the problem of medical communication barriers for hearing-impaired people, this paper designs and implements a sign language recognition assistant system based on YOLOv12. The system combines MediaPipe hand key point extraction and YOLOv12 model to realize real-time recognition of sign language gestures in medical scenes. Based on the CSL-500 data set, the medical sign language data set is constructed. Experiments show that the system recognition accuracy rate is 94.2%, the accuracy rate is 93.8%, and the recall rate is 94.1%. The system adopts the front-end and back-end separation architecture, supports real-time video acquisition, text display and voice broadcast, and provides an effective intelligent auxiliary tool for the communication between the hearing-impaired and the medical staff.
Keywords: Image recognition; Sign language recognition; YOLOv12; Intelligent medical

0 引言

随着数字化技术与人工智能在医疗领域的持续渗透，智慧医疗已成为当代医疗发展的重要趋势。《中国数字经济发展报告 2023》显示，我国数字医疗市场规模已达 7180 亿元，年复合增长率超过 15%。然而，在智慧医疗服务日益普及的背景下，听障人士在接受医疗照护时仍面临显著的沟通障碍。

在国外，基于计算机视觉的手语识别研究起步较早，已形成较为成熟的技术体系，当前研究重点集中于深度学习模型优化、实时检测加速及多模态融合等方向。WU Yaqin 等[1] 提出基于 YOLO 的轻量级目标检测算法，在保持检测精度的同时显著降低计算量，并通过对模型剪枝与网络结构优化实现手机端部署，对医疗场景下的便携式手语识别系统具有启发意义。Guo Mu 等[2] 针对自动驾驶中的交通灯检测问题，提出了一种适应不同光照与运动条件的视觉识别方法，其多尺度特征提取与时序信息融合思想对手语手势分割与动作序列分析具有借鉴价值。传感器技术的发展为手语识别提供了新路径。Chong Wang 等[3] 基于纳米流体离子整流传感器的研究，展示了新型传感器在精准识别与信号处理方面的优势，其模式识别思路对多模态手语识别及生理信号融合应用具有参考意义。在文本检测领域，Wondimu DIKUBAB 等[4] 构建了阿姆哈拉语场景文本检测与识别基准数据集，其数据集构建方法、评价指标设定及多语言处理策略对医疗手语数据集的建立及手语方言差异研究具有指导作用。此外，雷达信号与多载波信号检测技术的发展也为手语识别带来启发。WAN Tao 等[5] 利用可见图技术实现低截获概率雷达信号的检测与识别，其信号处理、特征提取及时序模式分类方法对手语动作序列分析具有借鉴意义，尤其对医疗环境中存在干扰的手语识别场景具有重要参考价值。Zhang Xing 等[6] 在多载波信号盲干扰检测方面的工作，以及 Fei Gao 等基于核函数判别分析在 SAR 图像目标识别中的研究，在特征提取、多通道信息融合及干扰抑制等方面为手语手势检测提供了理论支撑。

综上所述，尽管国内外手语识别技术已取得长足进步，深度学习与多模态融合显著提升了识别精度，但面向

医疗场景的专门化应用仍显不足，现有系统普遍缺乏针对医疗环境特点的定制化改进。因此，开发面向医疗场景的手语识别系统具有明确的研究价值与现实意义。本文围绕面向医疗场景的手语识别检测系统展开研究，主要包括三方面工作。第一，构建医疗手语数据集：以 CSL-500 为基础，采集并筛选医疗常用词汇手语视频，完成数据清洗、标注与增强。第二，设计轻量化识别方法：利用 MediaPipe 提取手部关键点，构建基于轻量级 YOLOv12 的时序神经网络，对连续手语动作序列进行建模与分类，实现实时推理。第三，开发前后端分离系统：后端采用 Python Flask 处理视频流与模型推理，前端构建支持文字显示与语音播报的用户界面，划分为管理员与用户两大模块，满足实际医疗场景需求。

1 研究基础

1.1 YOLOv12 模型

YOLOv12 于 2025 年初提出，是首个以注意力机制为核心的 YOLO 系列模型[7-8]，突破了 CNN 在 YOLO 框架中的长期主导地位。该模型引入三大关键创新：（1）区域注意力模块，通过将特征图划分为等分区域显著降低自注意力计算成本；（2）残差高效层聚合网络（R-ELAN），引入残差连接与缩放因子以优化梯度流动和特征聚合；（3）FlashAttention 及架构精简策略，移除位置编码并调整 MLP 比例以提升内存访问效率。YOLOv12[9] 作为最新迭代版本，在精度与速度上实现了更优平衡，在手语识别任务中，其注意力机制能够有效捕捉手势的全局空间构型与局部细微动作特征，其时序推理能力为动态手语序列建模提供了基础架构支持，适合部署于医疗等实时性要求较高的场景。

1.2 Flask Web 开发框架

Flask[10] 是一个轻量级的 Python web 开发框架，基于 wsgi 标准构建，采用 Jinja2 模板引擎和 werkzeug wsgi 工具集，具有简洁高效、灵活易用的特点，非常适合快速构建 web 应用原型与中小型 web 系统。Flask 的核心优势体现在：一是核心功能简洁精炼，仅保留了 Web 开发必需的核心组件，没有多余的冗余功能，使得框架本身轻量高效；二是路由机制灵活强大，支持基于装饰器的路由定义方式，能够方便地将 URL 路径与视图函数关联起来，实现请求的分发处理。三是模版引擎功能完善，jinja2 模板引擎支持模板继承，变量替换，条件判断、循环等功能，便于构建动态交互的 web 网页；四是扩展性强，Flask 提供了丰富的扩展插件，能够根据项目需求灵活扩展功能；五是易于与其他 Python 库集成，尤其适合与深度学习框架结合，可在同一 Python 进程内完成模型加载与推理，减少跨进程通信的开销。

2 模型构建

2.1 数据集

本研究所用数据集源自中国科学技术大学发布的 CSL-500 中文手语数据集[11]，该数据集包含 500 个常用手语词汇的 1250 个 RGB 视频样本，以 25 帧 / 秒采样率在受控环境下录制，保证了视觉特征的清晰度与一致性。数据集动作边界划分明确、标注规范，可有效支撑动态手势建模研究。据统计，该数据集在国内手语识别研究中的应用率达 78%，具有广泛的学术认可度。采用 CSL-500 作为实验数据来源，有助于构建具有对比性与可复现性的实验基准，为本研究提供可靠的数据支撑。部分数据切片如图 1 所示。

图1 数据采集图

2.2 数据预处理

本文采用多层级的数据预处理策略：

（1）视频归一化：将所有样本统一为 30 帧 / 秒的帧率及 640×480 像素分辨率，保证输入数据一致性。

（2）手部关键点提取：采用 MediaPipe 算法[12] 检测手部区域，输出 21 个三维关键点坐标，作为后续时空建模的特征输入。

（3）数据增强：对原始样本执行随机旋转（±15°）、缩放（0.8-1.2 倍）及亮度调整（±20%），将数据集扩充至 2250 个样本作为训练数据集，其中 80% 用于训练，10% 用于验证，10% 用于测试。

（4）环境鲁棒性增强：针对医疗场景特点，添加高斯噪声及模拟光照、背景变化的样本，提升模型真实环境适应性。

（5）序列及特征归一化：通过填充或截断将动作序列

统一至固定帧数，并采用 Z-score 标准化对特征向量进行归一化处理，确保模型稳定收敛。

2.3 模型训练效果

2.3.1 实验环境

为实现手语动作有效识别，本研究采用 CSL-500 数据集进行仿真验证，具体的仿真环境配置如下表 1 所示。

表1 实验环境配置表

组件/软件名称	版本/规格
CPU	Intel Core i7-12700K
内存	32GB DDR4
GPU	NVIDIA GeForce RTX 4060 Laptop GPU (8 GB)
开发语言	Python 3.8
深度学习框架	TensorFlow 2.9
手部关键点检测库	MediaPipe 0.8.10
视频处理库	OpenCV 4.6

2.3.2 评估指标

本模型性能评价指标包括混淆矩阵、准确率、精确率、召回率和 F1-Score[13]。混淆矩阵表示如表 2 所示，其中 TP 表示真阳率，FN 表示假阴率，FP 表示假阳率，TN 表示真阴率。准确率（Accuracy）是分类任务最常用的评价指标，指模型预测正确的样本数占总样本数的比例，直观反映模型整体分类效果。数值越接近 1，代表模型对手语图形的识别能力越强，其计算方法见式（1）。精确率指标（Precision）用来衡量“预测得准不准”，数值越接近 1，代表模型对手语图形的识别能力越强，其计算方法见式（2）。召回率指标（Recall）用来衡量模型“找得全不全”，数值越高，性能越好，其计算方法见式（3）。F1-Score 是精确率和召回率的调和平均数，综合二者表现，取值[0,1]，越接近 1 效果越好，其计算方法见式（4）。

表2 混淆矩阵

预测结果 \ 真实标签	正常	异常
正常	TP	FN
异常	FP	TN

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1}-\text{score}=\frac{2\times precison\times recall}{precision+recall} \quad (4)$$

2.3.3 识别效果

训练过程中选用的相关参数如下表 3 所示，同时使用早停法来观察验证集上的损失情况，具体训练迭代过程见图 2 所示，如果连续 15 次验证集损失没有明显降低就停止训练。经多次实验，模型混淆矩阵如下图 3 所示。经计算，模型整体识别准确率为 94.2%，精确率为 93.8%，召回率为 94.1%，F1-score 为 93.9%。

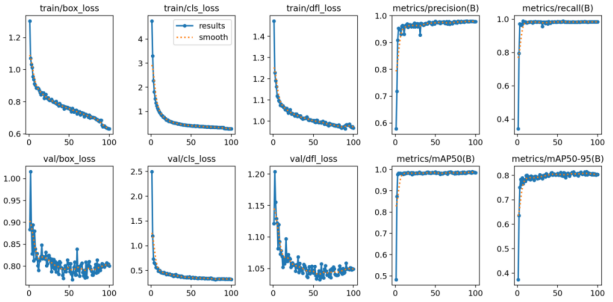

图2 模型训练过程变化趋势图

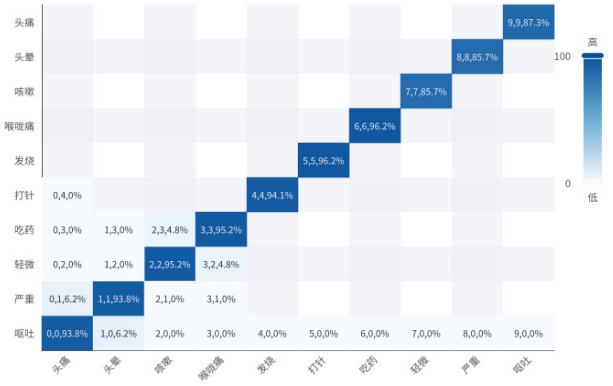

图3 模型测试混淆矩阵

表3 训练参数设置表

组件/软件名称	版本/规格
优化器	Adam
损失函数	交叉熵损失
初始学习率	0.001
批次量大小	32
训练次数	200

3 系统实现

为满足医疗场景下听障患者与医护人员的实时沟通需求，本系统采用前后端分离架构，将计算密集型的模型推理与交互型的前端展示解耦[14-15]。整体架构自底向上分为数据层、服务层与交互层：数据层负责医疗手语数据集、模型文件、运行日志及手语词库的分类存储与管理；服务层基于 Python 与 Flask 框架，集成 OpenCV、MediaPipe 及 TensorFlow 等工具，实现视频流处理、手部 21 个关键点三维坐标提取、基于 YOLOv12 融合模型的推理以及接口提供，并通过 RESTful API 与 WebSocket 分别完成非实时数据交互与实时视频流传输；交互层采用现代 Web 技术栈，支持多终端设备，负责视频采集、结果可视化展示及系统管理。系统运行流程为：前端采集手语视频流并推送

至服务层，经预处理与手部关键点提取后，由 YOLOv12 模型将特征序列实时转换为手语识别结果，再依据手语词库映射为文字并反馈至前端，从而形成手语到自然语言的实时转换，实现高效、简洁的人机协作与沟通。

3.1 后端设计

后端基于 Python 3.8、Flask 框架，部署在 Ubuntu 20.04 上，结合 OpenCV、MediaPipe、TensorFlow 等库，开发四大核心模块与辅助功能，保障系统业务高效运行。

视频流处理模块：接收前端视频流，用 OpenCV 解析、预处理和标准化视频帧，统一分辨率和帧率，去除噪声与光照干扰，还支持本地视频处理，满足非实时需求。

手部特征提取模块：集成 MediaPipe 算法，检测手部区域、提取关键点，输出 42 维特征序列，优化检测逻辑，在复杂场景下确保关键点提取准确率。

模型推理模块：封装 YOLOv12 融合模型为推理函数，接收特征序列进行分类推理，输出识别结果并匹配词库，轻量化优化模型，控制单帧推理时间，保证实时响应。

系统接口模块：基于 Flask 搭建 RESTful API 和 WebSocket 接口，前者支持系统管理，有鉴权机制保障安全；后者实现视频流和识别结果双向传输，满足实时沟通需求。

3.2 前端设计

用户端是听障患者与医护人员进行沟通的核心界面，主要实现实时视频采集、识别结果展示三大核心功能，界面划分为图像采集区、结果展示区、功能控制区三个区域。视频采集区支持电脑摄像头、外接摄像头的实时视频采集，也支持本地手语视频文件的上传，实时显示采集的视频画面；结果展示区以大字体、高对比度的形式清晰显示手语识别对应的医疗文字信息，便于医护人员和患者快速查看；功能控制区包含识别结果保存等功能按钮，满足不同医疗场景的沟通需求，识别结果保存功能可将沟通记录保存为文本文件，便于后续的病历整理。用户端通过 WebSocket 技术与后端实时通信，保证视频流和识别结果的低延迟传输，提升交互体验。

3.3 系统展示

用户端各类功能按键以及检测面板见图 4 所示。用户可以选择视频摄像头捕获数据进行检测，也可以进行上传图片检测、视频检测，对于线上听障用户很便捷。用户通过上传图片进行手语识别，并以中文向用户反馈识别结果，具体见图 5，同时用户可以选择摄像头进行检测，具体见图 6。

图 4 用户端各类功能检测板面图

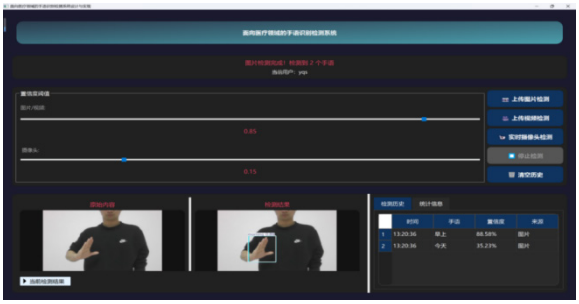

图5 上传图片检测结果展示

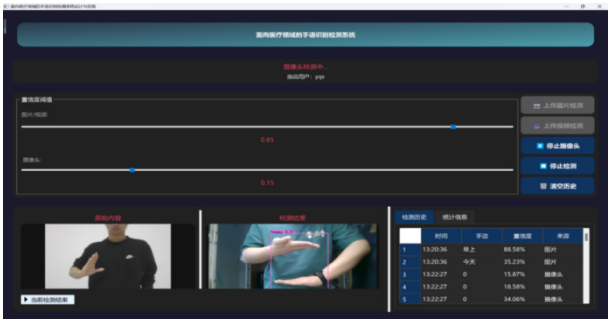

图6 摄像头检测结果展示

4 结语

本文围绕医疗场景中听障人士与医护人员之间的沟通障碍问题，设计并实现了一套基于 YOLOv12 的图像识别手语辅助系统。主要工作包括：构建了面向医疗场景的手语数据集，提出了基于 MediaPipe 与 YOLOv12 融合的轻量化手语识别方法，并开发了前后端分离的系统架构。实验结果表明，该系统在 CSL-500 数据集上的识别准确率达到 94.2%，具备较高的识别精度与实时响应能力。系统前端界面友好，支持多种输入方式（摄像头、图片、视频），识别结果可实时转换为文字与语音，满足实际医疗场景中的沟通需求。未来可进一步扩展手语词库规模，增强模型的泛化能力，并探索多模态融合策略，以提升在复杂医疗环境中的鲁棒性与实用性。

参考文献：

[1] WU Yaqin, ZHANG Tao, NIU Jianjun, et al. YOLO-based lightweight traffic sign detection algorithm and mobile

deployment[J]. Optoelectronics Letters, 2025.

[2] Guo Mu, Zhang Xinyu, Li Deyi, et al. Traffic light detection and recognition for autonomous vehicles[J]. The Journal of China Universities of Posts and Telecommunications, 2015(01): 54-60.

[3] Chong Wang, Hao Xie, Rulan Xia, et al. Nanofluidic ion rectification sensor for enantioselective recognition and detection[J]. Chinese Chemical Letters, 2025.

[4] Wondimu DIKUBAB, Dingkang LIANG, Minghui LIAO, et al. Comprehensive benchmark datasets for Amharic scene text detection and recognition[J]. Science China (Information Sciences), 2022(06): 77-78.

[5] WAN Tao, JIANG Kaili, LIAO Jingyi, et al. Detection and recognition of LPI radar signals using visibility graphs[J]. Journal of Systems Engineering and Electronics, 2020(06): 104-110.

[6] Zhang Xing, Hu Jianhao. Blind interference detection and recognition for the multi-carrier signal[J]. The Journal of China Universities of Posts and Telecommunications, 2017(02): 52-60.

[7] 韩晓冰，胡其胜，赵小飞等. 改进 YOLOv7-tiny 的手语识别算法研究[J]. 现代电子技术，2024(01): 63-69.

[8] 邓玉琼，李维乾，何飞等. 基于改进 YOLOv7 的轻量化手语识别模型研究[J]. 舰船电子工程，2025.

[9] 潘丽. 基于改进的 YOLOv5s 结构的手语识别设计[J]. 西昌学院学报 (自然科学版)，2024(02): 57-63+69.

[10] 殷钦东，刘明良. 基于 Flask 与 Vue 框架的智能垃圾识别系统设计与实现[J]. 电脑编程技巧与维护，2026(5): 132-134+159.

[11] 深度学习模型对 CSL 手语数据集进行中国科学技术大学 CSL 手语数据集训练基于卷积神经网络与循环神经网络结合架构，3D-CNN 或 I3D 特征提取，配合 LSTM 或 GRU 处理时间序列信息. 中科大 csl 手语数据集 -CSDN 博客.

[12] 李红蓉，潘文林，聂腾涛等. 基于 MediaPipe Landmarks 的动态手语孤立词识别[J]. 计算机与数字工程，2026, 54(3): 780-785.

[13] 张恒博. 基于深度学习的手语识别研究与实现[D]. 银川：宁夏大学，2023.

[14] 井钰. 面向医疗领域的舆情分析系统的设计与实现[D]. 哈尔滨：东北林业大学，2022.

[15] 王蕾. 面向医疗健康领域的智能问答系统的设计与实现[D]. 北京：北京邮电大学，2018.

作者简介：张茹（1998-），女，汉族，陕西咸阳人，硕士，助教，研究方向：图像识别与处理、深度学习。