基于 YOLO 与大语言模型的智能手语翻译

范子杨 闫屹伟 李昕明 朱志刚(通讯作者)

大连理工大学城市学院，辽宁省大连市，116600；

摘要：当前，听障群体对智能手语翻译提出了更多的要求，需要一种能够安装在移动电子设备上的智能手语翻译软件。依托 YOLO 技术及大语言模型搭建智能手语翻译系统有利于满足人们的相关需求。本文说明基于 YOLO 与大预言模型搭建智能手语翻译系统的流程与方法，探讨听障群体如何借助该智能手语翻译软件助力人们高效学习与生活。

关键词：YOLO 技术；大语言模型；智能手语翻译系统

DOI：10.64216/3080-1494.25.10.053

前言

手语是听障人士与他人进行沟通交流的主要方式，但对于不了解手语的交流对象来说，可能会出现理解手语和回应交流方面的困难，这就产生了手语翻译需求。传统手语翻译中存在较多的局限性，无法满足沟通需求。随着现代人工智能技术的成熟与发展，智能手语翻译智能化程度持续提高，为了进一步满足听障人士交流需求，项目组拟使用 YOLO 算法与大预言模型搭建一款轻量智能手语翻译软件，实时助力听障人士顺畅交流。该软件预计能够为手语翻译提供精准的视觉输入，能够准确且快速的识别手语动作中的位置信息及关键姿态，从而将其转化成自然流畅的文字、语音或动画，以便不了解手语的人们迅速理解听障人士使用手语表达的寓意，提升双方交流效率及交流有效性。

1 项目研究背景与意义

1.1 项目研究背景

随着科学技术的发展，YOLO 算法及大语言模型的为智能手语翻译软件的搭建提供了技术支撑，如，YOLO 算法能够精准却快速的识别物体的类别与位置，能高效捕捉手语动作中的关键信息；使用 YOLO 算法连接大语言模型能够将 YOLO 算法识别与捕捉到的手语关键信息生成文字语言或语音，从而快速实现准确的手语翻译，帮助听障人士与他人顺畅沟通。

1.2 项目研究意义

从社会层面来看，使用 YOLO 与大语言模型搭建智能手语翻译系统有利于解决听力残障人群沟通中出现的真实问题，帮助听障人群打开与人沟通的大门。

从技术层面讲，此项目研究中创新性的将自然语言处理技术与计算机视觉技术进行融合，充分发挥出两种技术的优势，精准提炼出两种技术的连接点，为在特殊领域对二者进行融合应用提供了新思路参考。

2 拟解决的关键问题

（1）解决手语翻译软件跨平台兼容性问题：不同平台（如移动端、桌面端、智能硬件配置等）的硬件和操作系统差异加大，导致手语识别技术的跨平台兼容性差异。例如，视频捕获的分辨率、处理能力、输入设备等不同可能影响手语识别的准确性和效率。

（2）解决实时性和计算资源的限制问题：实时手语翻译要求系统能够快速处理视频流中的语音信息并生成文字或语音。对于深度学习模型而言，实时处理大量视频数据需要强大的计算资源，尤其是跨平台应用时，不同平台的计算能力差异会影响性能。

（3）优化硬件和环境因素：手语识别技术通常依赖于摄像头等硬件设备进行视频捕获，但不同设备的质量差异（如分辨率、帧率、颜色深度等）可能会影响识别的效果。此外，背景噪声、光线变化、手部暴露等因素也可能影响识别性能。

（4）解决文化差异下的手语翻译挑战：手语翻译不仅是语言转换，还涉及文化背景的传递。不同地区的手语中可能包含文化的特定表情、言语或符号，且不同的手语具有不同的语法结构和表达方式。跨文化的手语翻译内容很大，尤其是在自动化翻译中，很难处理文化方面的复杂性。

3 使用 YOLO 算法与大语言模型搭建智能手语翻译系统的可行性

3.1 技术基础成熟度保障可行性

当前阶段，YOLO 算法与大语言模型都是较为成熟的技术，技术的成熟度为两者之间的融合提供了诸多保障。YOLOv5 算法技术具备高精度检测能力，能够实现每秒十帧的检测。在医疗领域，我们可以使用该算法迅速定位患者病灶区域，为医生进行医疗决策提供真实数据；可以在交通领域利用该算法进行行人识别和车辆检测。诸多实际应用案例证明，YOLO 算法具备处理动态手语动作

所需的准确性和及时性。

大语言模型也是当下较为成熟的一种智能技术，能够为智能手语系统的搭建提供有力支持。如，LLaMA 等语言模型已经经过万亿级参数的训练，具备了跨领域知识推理能力，这一功能有利于为手语语义理解提供技术参照，能准确转换手语语义，应用该技术搭建智能手语翻译系统有利于解决文化差异下手语翻译不够精准的问题。

3.2 算法协同机制增强可行性

YOLO 算法与大语言模型能够通过多模态融合架构实现协同工作。如，我们可以在前端采用 YOLO 算法对手语动作进行检测，生成包含动作类型、运动轨迹的特征向量；可以在后端采用大语言模型接收该向量，将其转化为文字或语音，实现手语翻译功能。

从数据流处理角度来讲，我们可以将 YOLO 算法与大语言模型边缘计算与云端协同的方式进行融合架构。如，使用 YOLO 算法将特征数据上传至云端。在云端搭载大语言模型对特征数据进行深度解析与转换，生成文字与语音。

4 使用 YOLO 算法与大语言模型搭建智能手语翻译系统的流程与方法

4.1 系统架构设计

4.1.1 分层架构规划

搭建智能手语翻译系统，首先需要做好架构设计，本次设计中采取了四层架构设计，架构设计逻辑为自下而上搭建数据采集层、特征提取层、语义理解层和交互输出层。数据采集层能够实时获取原始手语视频，为上层处理提供数据资源支撑；特征提取层需打造 YOLO 算法，使用 YOLO 算法对手部动作关键点进行检测，对运动轨迹进行分析，实时、准确的识别和获取特征向量；语义理解层需要搭载大语言模型，连接 YOLO 算法获得特征向量，对向量进行分析和推理，结合上下文和不同区域语言文化，完成词汇映射与语法组织；交互输出层将大语言模型翻译结果转化为语音、文字或三维动画，实现智能手语翻译功能。

4.2 模块间通信机制

搭建好智能手语翻译系统分层架构之后，需要通过标准化结构实现各层间的数据传输。在特征提取层与语义理解层使用 protobuf 协议实现序列化通信，将手部运动速度、关键点坐标等内容编码为二进制流，降低网络传输开销；在语义理解层与交互输出层应用 WebSocket 协议，将语义理解结果实时推送到交互输出层，从而实现翻译结果与原始手势同步显示的目的。

4.3 数据采集与预处理

4.3.1 多模态数据采集

（1）多模态数据采集方案

在智能手语翻译系统中同时搭载双目摄像头与 IMU 传感器，使用摄像头获取手语视频等资料，使用传感器获取相关运动数据。其中，双目摄像头能够以较高的帧率和分辨率捕捉手语视频资料，有利于获得更清晰的资料；IMU 传感器能够以 100 赫兹采样率对手腕旋转的角度进行捕捉，从而获取精准运动数据。该配置下，能够实现多源数据融合，从而解决单一传感器配置下数据丢失的问题。

（2）数据标注与增强

智能手语翻译系统标注工作中，我们可以使用半自动流程，先使用预训练模式生成初始标注，然后由人工对标注情况进行审核与修正，做好关键帧修正。标注过程中应当严格遵循 WS-3.0 标准，对位置、运动方向、手型等进行规范标注。数据增强模块实施几何变换、色彩调整、运动模糊、噪声注入、遮挡模拟和时序扰动，从而强化模型泛化能力。如，数据增强模块下单一样本能够生成 200 个变体。

4.4 YOLO 算法实现与优化

4.4.1 模型选型与适配

YOLO 算法是该手语智能翻译系统的核心技术之一，选择使用 YOLOv8 为基础框架。为了提升手部小目标检测精准度，需要对锚框生成策略进行修改，将最小锚框尺寸像素调整至 4*4。此外，还需增加浅层特征图融合，实现小目标检测 Recall 率的提升。

4.4.2 实时优化策略

智能手语翻译系统使用 TensorRT 加速引擎量化压缩模型，实现精度从 FP32 向 INT8 的转变，从而提升 3 倍推理速度。同时开发轻量版本适配 Jetson Nano 设备，这有利于剪除冗余通道，在保证准确率的前提下减少内存占用。

4.5 大语言模型集成与应用

4.5.1 模型选择与微调

大语言模型是智能手语翻译系统的另一核心组成模块，因此，应当选择一种准确率较高的语言模型，如，我们可以使用 LLaMA-2-7B 大语言模型，该模型能够实现高准确率识别。之后，还需使用 LoRA 微调技术对参数进行适当调整，从而实现领域适配。训练数据包含 20 万条手语-文本平行语料，覆盖教育、交通、医疗等八大场景。微调后的模型，有利于提升长句翻译质量。

4.5.2 上下文理解增强+文化识别

在智能手语翻译系统中引入记忆增强机制，如搭建滑动窗口，对最近 5 个手势语义向量进行缓存，以便系统在检测到重复手势时参考历史上下文消除歧义，进行

准确翻译；此外，需在系统中输入系统应用地本地语言特色文化，以便系统基于文化差异背景更准确的进行手语翻译。

4.6 多模态融合与输出

4.6.1 特征对齐与融合

在智能手语翻译系统中建立时空对齐模块，实现 YOLO 输出的关键坐标与大语言模型生成的语义标签之间的关联。使用图神经网络捕捉手指间相对运动特征，形成包含运动特征、空间位置和运动特征的多模态向量，为手语翻译提供有力支撑。

4.6.2 输出形式定制化

智能手语翻译系统支持三种输出模式，即文字输出、语音输出和动画输出。用户可以在不同场景选择不同输出模式，如，特殊学校教师可使用文字或动画模式面向听障学生教学；如，公共服务场景中，可选择同时激活文字+语音模式，满足沟通需求。

4.7 系统测试与迭代

4.7.1 性能测试指标

在智能语言翻译系统中建立思维评估体系，评价内容包括准确率、鲁棒性、实时性即用户体验。词汇识别准确率达到 96.3%，性能波动在 300lux-10000lux 光照范围内小于 2%，端到端延迟小于 280ms，则为性能测试合格。

4.7.2 持续优化机制

智能手语翻译系统需持续为残障人群提供服务，因此需要持续对系统机制进行优化。我们需在系统中部署在线学习模块，根据用户反馈进行持续优化。如，某用户连续 5 次对翻译结果进行修正，则触发系统自动微调功能，这有利于系统翻译结果更加贴合用户需求。

上述搭建流程模块化设计，能有效实现技术解耦，使各组件间既可以协同工作又可以独立优化。随着 5G+边缘计算的普及，智能手语翻译系统将进一步向低功耗和轻量化方向发展，为听障人群实现普惠型无障碍沟通提供技术保障。

5 用户如何使用智能手语翻译系统进行手语翻译

该款智能手语翻译系统使用流程较为简单方便。用户开启系统后，将电子设备放置在恰当的位置，面向摄像头进行手语表达。此时，系统内的 YOLO 算法迅速启动，检测及提炼关键信息，大语言模型将这些信息转化成文字或语音，用户将文字或语音呈现给沟通对象。

如，在医院场景中，听障患者使用该系统可以迅速将手语转化为语音、文字，进行导诊和就诊咨询。患者进入医院后，可面向医院提供的智能手语翻译系统进行手语表达，询问就诊科室的位置。系统准确理解患者意图后，即向患者显示科室位置和相关路线。如果患者需要导诊人员帮助，即可点击系统中的语音模式，导诊人员听到语音播报后即前来为听障患者提供帮助，这为听障患者高效就医提供了较大助力。

6 结语

基于 YOLO 与大语言模型的智能手语翻译系统，能够通过 YOLO 算法实现对手部关键动作的精细与完整捕捉，生成语言向量，大语言模型接收到向量后即可将其转化为文字、语音或动画，供用户按需选择，具备较强的准确性、及时性便捷性。该款智能手语翻译软件搭建流程包括系统架构设计、模块间通信机制、数据采集与预处理 、YOLO 算法实现与优化、大语言模型集成与应用、多模态融合与输出及系统测试与迭代。将 YOLO 算法与大语言模型融合应用于智能手语翻译系统是一个创新性与实用性均较强的模式，我们应在日后加强对相关技术的研究及应用，以便进一步提升智能手语翻译软件功能。

参考文献

[1]朱向阳,张定国,李成璋.新型的智能手语翻译与人机交互系统及其使用方法:CN 201310101297[P][2025-09-05].DOI:CN103279734 A.

[2]叶显锋,李涛,王超.一种基于大语言模型的汉语手语翻译方法及系统:202411212390[P][2025-09-05].

[3]冯文静,王岩,张天宇,等.基于 YOLO V5 和 Transformer 的实时手语识别和翻译方法研究[J].移动信息,2024,46(10):316-319.

[4]刘继兴,周昕,张帅峰,等.基于人工智能的手语翻译系统实现[J].科技创新与应用,2022,12(23):4.DOI:10.19981/j.CN23-1581/G3.2022.23.010.

[5]邓增辉,程胜月,王建彬,等.基于Web和深度学习的手语翻译系统的设计与实现[J].移动信息,2024,46(8):358-360.

基金项目：辽宁省大学生创新创业项目(项目编号:X202513198161，项目名称：聆听无声之语-基于 YOLO 与深度学习的手语翻译)。