面向听障人群的双向手语互译平台分层架构与关键技术研究

李澍惠 郑雅雯 郭丽 黄雨欣
嘉兴大学商学院，中国·浙江 嘉兴 314001

摘要：就国内听障者所遇手语翻译资源缺乏、目前智能手语工具只作单向使用、弱网不能运行、复杂情况识别不准的实际困难而言，应建立端边云三级协同的分层手语互译平台。将视觉、惯性及肌电信号三者作为统一输入，按一定规则完成预处理流水线设计，利用 VGG16-MLP 并行结构对特征加以提取，结合信息熵自动调节的加权法做融合处理，配合 CNN-LSTM 网络对连续手语进行时序解码；按 INT8 标准压缩模型、用通道剪枝方法降低复杂度，把移动端离线和云端精确两种方式分开运用，同时组织多级公益实施计划。仿真试验给出多模态融合类方案的准确率 99.4%，其离线性能基本保持在线结果的 90% 以上，端到端延迟少于 0.2s。所提分层互译思路既重实时、又利离线、且具普惠意义，是无障碍手语智能产品及长期公益的可行技术路线。

关键词：手语互译；端边云协同；多模态融合；轻量化模型；无障碍交互

Research on the Layered Architecture and Key Technologies of a Two-Way Sign Language Translation Platform for the Hearing-Impaired

Li Shuhui, Zheng Yawen, Guo Li, Huang Yuxin
School of Business, Jiaxing University, China Zhejiang Jiaxing 314001

Abstract: Considering the practical difficulties faced by the hearing-impaired in China—such as the lack of sign language translation resources, current intelligent sign language tools being only one-way, weak network conditions preventing operation, and inaccurate recognition in complex situations—it is necessary to establish a three-tiered sign language translation platform with end-edge-cloud collaboration. By using visual, inertial, and electromyographic signals as unified input and designing a preprocessing pipeline according to certain rules, features are extracted using a VGG16-MLP parallel structure. Fusion is handled via a weighted approach automatically adjusted by information entropy, paired with a CNN-LSTM network for temporal decoding of continuous sign language. Models are compressed to INT8 standards and complexity is reduced with channel pruning, applying offline mobile and precise cloud processing separately, while also organizing multi-level public welfare implementation plans. Simulation experiments show that the multimodal fusion approach achieves an accuracy of 99.4%, with offline performance maintaining over 90% of the online results and end-to-end latency under 0.2 seconds. The proposed layered translation approach emphasizes real-time performance, supports offline usage, and has social inclusivity significance, offering a feasible technical path for barrier-free intelligent sign language products and long-term public welfare.

Keywords: Sign language translation; End-edge-cloud collaboration; Multimodal fusion; Lightweight model; Barrier-free interaction

1 相关基础技术

本文介绍面向听障者端边云双向手语译平台核心技术和创新点。平台包含边缘机器并行计算、多模态传感采集、深度特征学习、时序处理以及移动端推理轻量化技术，提供完备的手语双向翻译技术体系，配合多个模块适配于手语应用场景，提供快捷、可靠的手语无障碍翻译。

平台为端、边、云协同分层架构，分层均按算力排序进行任务拆分、层层下放运算。终端只执行视频、肌电、惯性感官原始数据和原始数据简单预处理，无大型网络部署，避免不必要的负荷开销在终端设备算力和功耗上。边缘机完成手部特征提取、短时间手语序列推理的中等算力工作，解决移动端算力低、实时性差的问题。云端完成处理长句子手语的解析工作、手语高阶映射和 3D 虚拟手语的渲染等算力大工作，云端通过统一的 API 接口实现对不同终端的同步更新。实现机器单终端算力有限、单纯云上传重传输性的两大不足，适应弱网和无网环境，能够进行

离线轻量化、在线云端高精度双端翻译模式，有效运行。

采集与预处理多源数据是实现高精度手语识别的基础。平台利用手机RGB相机采集手部视频，对采集视频图象进行灰度转换、高斯滤波、直方图均衡等图像增强方法实现阈值分割，消除背景信息，采用MediaPipe轻量化框架的帧手姿全局定位和帧手姿局部追踪算法，能够实时地输出42个关节的完整手部的二维关键点坐标，避免了后续计算的冗余。对于肌电、惯性一维时序数据采用后处理方法对测量数据做异常值去除、滑动窗口分段和Z-score标准化来减弱误差偏差的影响，基于视频时间戳对各模态进行统一成多模态采样的时间，以消除多模态数据的采样时间不准的现象，从而实现多源数据规整同步。

平台搭建VGG16、MLP两个分支的手语识别特征提取结构分别对应手语识别的多模态。VGG16每一层卷积、池化结构都对轮廓特征、纹理特征、全局手势空间等提取比较稳定的特征特征适合静态手势的识别。MLP分支的肌电信号、惯性运动信号中提取较小的发力、运动特征结果空间维度一致，所得结果特征向量，这适合多模态得到的特征成统一空间规格进行特征向量融合统一输入。

平台使用特征层加权融合与CNN-LSTM时序解码提高连续手语识别精度；使用基于信息熵算法的模态加权权重动态分配模态的有效特征，在平时抑制视觉模态特征，在暗光、遮挡等各种条件情况下增强肌电和运动特征，来解决单一视觉识别抗干扰性不强、手势混淆等问题。CNN-LSTM结合网络空间特征提取与时序记忆机制来弥补传统RNN网络中存在的梯度消失问题，可以提取一段连续手语文本里有效手语的动作与过渡帧的区别，来进行长周期手语语义准确解码。

平台用INT8量化进行模型轻量化和移动端离线时序推理，将FP32浮点参数量化为整数8位，经过校准微调，小精度损失，大幅降低模型的大小和计算代价，不必改动原有网络结构，在断网下仍保证足够翻译效果。系统最终实现完美地双向翻译：手语翻译成文本，对接时序解码和常用的通用手语语法规整语义；文本翻译成手语，分词映射，拼接动作时序得到标准手语，借助数字孪生渲染3D虚拟手语人，实现手语与文字、语音的双向可视化互译，拉近听障者与健听者间的“隐性鸿沟”。

2 端边云协同整体架构总体思路与分层设计

2.1 架构总体思路

面向听障人群自然手语双向翻译的实时交互、低延迟响应、离线使用与模型轻量化落地需求，本文手语互译平台搭建端侧、边缘、云端三级协同分层架构，解决传统纯云端方案网络依赖强、传输时延高、隐私风险突出等问题。架构主要承载移动端离线手语互译、线下场景低时延实时翻译、云端数据集训练与模型迭代三大核心能力，依托手语采集、多模态解析、语义转换、可视化输出与数据回流的完整链路实现层级解耦。各层功能边界清晰、接口标准统一，可适配手机、智能穿戴、公共无障碍终端等设备，覆盖日常沟通、政务办事、医疗问诊、线上通话等多元化无障碍服务场景[4]。

2.2 分层层级详细设计

端侧为用户交互核心入口，部署经剪枝、INT8量化压缩的轻量化模型，支撑断网状态下基础双向手语翻译。硬件涵盖智能手机、无障碍平板等设备，集成数据采集、本地推理、交互输出与离线缓存四大模块，可采集手部骨骼、肢体姿态与语音信息，完成本地降噪、手势推理及3D虚拟手语渲染，依托本地两千余条高频词汇句式缓存保障离线稳定性。同时端侧完成原始生物数据脱敏，仅上传特征向量，从源头保护用户隐私。

边缘节点部署于政务大厅、医院、特教学校等公共场景，承担端侧与云端的中转调度工作，主要负责多终端并发处理、长时序手语与复杂长句翻译任务。节点仅接收脱敏特征向量，无需传输原始视频流，依托场景专属手语词典优化语义解析效果，弥补端侧长句识别精度不足的问题。通过并发调度机制严控交互时延，短期缓存交互样本并脱敏后批量上传云端，不留存用户原始数据。

云端作为训练与服务中枢，不参与实时翻译，核心负责数据集管理、模型迭代、专业语义解析与全域运维。云端搭建细分领域多模态手语样本仓库，依托海量数据优化算法模型，通过知识蒸馏、量化剪枝生成轻量化模型并推送至全终端。同时内置国标手语语义知识库，补充医疗、法律等专业术语，完成高精度语义校验，统一实现设备运维、版本更新与数据统计。

2.3 三层协同业务流程与架构优势

平台核心手语翻译链路实现层级协同联动：简单短句手语可通过端侧本地模型直接完成识别翻译，即时输出文字与语音；复杂长时序手语由端侧上传脱敏特征至边缘节点，完成特征融合与时序解码；含专业术语的内容则转发至云端完成语义补全与精度校验，优化结果逐层回传终端实现无障碍交互。该分层架构优势显著，可实现算力分级均衡调度，兼顾离线使用与高精度翻译需求；本地与边缘优先的推理模式大幅降低传输时延，满足实时对话要求；

分级脱敏传输机制有效规避用户生物隐私泄露风险。此外，各模块相互解耦，可独立迭代升级，适配各类新增场景，且硬件适配门槛低，便于平台规模化公益落地推广。

3 多模态手语数据处理与识别算法

3.1 多模态手语数据采集与预处理

采用对多模态的深入整合手段，我们不仅仅将传统的视觉数据的采集推到了极致，还将动作的、肌电的等其他的核心模态的数据都有效的融入了其中，形成了从外在的视觉到内在的动作、肌电的三大核心的数据的多模态的采集平台。视觉模态依托移动端 720P/1080P、30fps 摄像头采集手语视频，记录手部姿态、掌心朝向与面部微表情。采用对多模态的深入整合手段，我们不仅仅将传统的视觉数据的采集推到了极致，还将动作的、肌电的等其他的核心模态的数据都有效的融入了其中，形成了从外在的视觉到内在的动作、肌电的三大核心的数据的多模态的采集平台。

原始多模态数据受光照、噪声、遮挡等因素干扰，需对视觉、动作、肌电数据进行分层预处理。视觉数据预处理首先采用高斯滤波去除传感器噪声，将帧图像统一缩放至 224×224，转换为灰度图并进行直方图均衡化，通过公式 $H_{eq}(x,y)=T[H(x,y)]$ 重新映射像素灰度值分布，有效消除光照不均与强逆光带来的对比度差异；最后基于 MediaPipe 完成手部 21 个关键点坐标的空间校准与背景干扰剔除。动作数据预处理采用滑动平均滤波对关键点坐标序列进行平滑处理，滤除高频抖动噪声，设定阈值剔除异常抖动点后提取速度、加速度等运动学特征。肌电数据预处理经带通滤波器（20~450Hz）滤除工频干扰与运动伪迹，信号经滤波后表示为 $EMG_{filtered}=EMD_{raw}*h(t)$，经基线校正后提取峰值、均值、方差等统计特征。三类数据以视觉数据的时间戳为主参考轴，通过插值重采样完成时间轴同步，时间轴同步的校验条件为 $T_{sync}=\max(T_{visual},T_{motion},T_{emg})$ 确保同一手语动作的多模态数据在时间上严格匹配，输出标准化同步化特征数据集。

3.2 基于 YOLOv8+MediaPipe 的双引擎识别架构

单一目标检测算法难以兼顾手语识别的精度与效率，YOLOv8 手部定位准确率高但计算量大，MediaPipe 轻量化高效但复杂环境鲁棒性不足[1-2]。对此，本平台采用二者融合的双引擎级联架构。依托 YOLOv8 多尺度检测能力精准裁剪手部 ROI 区域，坐标误差控制在 2 像素内；再通过 MediaPipe 两级定位模型输出 21 个手部 2.5D 关键点与三维坐标，精准提取手部几何特征以区分手语状态。系统实现自适应协同调度，简单场景依托 MediaPipe 低延迟推理，复杂场景激活 YOLOv8 辅助校正，有效弥补单一视觉算法的性能短板。

3.3 多模态特征提取与融合算法

单一视觉传感器易受光照、遮挡与背景干扰，仅依靠单模态信息难以完整表征手语语义。为此，本平台融合视觉、动作、肌电三类模态，通过多维度特征联合提取与加权融合提升手语识别精度与环境鲁棒性[3-5]。视觉特征依托 VGG16 网络提取，该网络由 13 个卷积层、4 个最大池化层与 3 个全连接层构成，采用 3×3 小卷积核堆叠结构，参数量约 1.47 亿，以 MediaPipe 裁剪的 224×224 手部 ROI 区域为输入，输出视觉特征向量 S1。动作特征通过 MLP 多级全连接拓扑完成非线性映射，得到动作特征向量 S2。肌电信号经带通滤波、基线校正去噪后提取统计特征，生成肌电特征向量 S3。平台通过计算三类模态特征的余弦相似度实现动态加权融合，有效发挥多模态互补优势。计算公式为：

$$\cos\theta=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|}$$

其中 A、B 分别为待识别样本与已知词汇库样本的特征向量。

随后，按视觉特征权重 0.5、动作特征权重 0.3、肌电特征权重 0.2 的比例进行加权融合，得到综合相似度：

$$S=0.5S_1+0.3S_2+0.2S_3$$

当 $S\geqslant 90\%$ 时匹配成功。平台实现对 5000+ 词汇的精准识别，融合延迟低于 50ms，综合识别准确率稳定在 99.4% 以上。

3.4 基于 CNN-LSTM 的时序语义识别模型

单帧检测模型无法捕捉时序依赖。本平台构建 CNN-LSTM 模型，通过 CNN 提取帧内空间特征、LSTM 建模帧间时序依赖。

CNN 空间特征提取层：CNN 通过卷积操作在局部感受野上提取图像的纹理、边缘、形状等深层特征。在手语识别任务中，手部 ROI 图像经过裁剪与归一化后输入 CNN 模块[6]。CNN 空间特征提取层沿用 VGG16 卷积结构，2D-CNN 的卷积操作可表示为[6]：

$$y_{m,n}=\sum_{j=0}^{J-1}\sum_{i=0}^{I-1}x_{m+i,n+j}\cdot w_{i,j}+b$$

其中 x 为输入的手部 ROI 图像，w 为卷积核的权重，（m，n）为图像像素坐标，I 和 J 分别为卷积核的高度和宽度，b 为偏置项，$y_{m,n}$ 为卷积输出特征[7]。CNN 将各帧映射

为空间特征向量形成特征序列输入 LSTM。

本平台采用 LSTM 网络对连续帧的空间特征序列进行时序建模。LSTM 通过遗忘门、输入门和输出门的三门控机制，有效解决传统 RNN 的梯度消失与长期依赖问题。

LSTM 三门控的核心计算过程如下[7]：

遗忘门 f_t：

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right)$$

输入门 μ_t

$$\mu_t = \sigma\left(W_{x\mu}x_t + W_{h\mu}h_{t-1} + b_\mu\right)$$

$$\widetilde{C}_t = \tanh\left(W_{xC}x_t + W_{hC}h_{t-1} + b_C\right)$$

细胞更新：

$$C_t = f_t \odot C_{t\text{-}1} + \mu_t \odot \widetilde{C}_t$$

输出门 σ_t：

$$\sigma_t = \sigma\left(W_{x\sigma}x_t + W_{h\sigma}h_{t-1} + b_\sigma\right)$$

$$h_t = \sigma_t \odot \tanh(C_t)$$

其中 x_t 为 t 时刻的输入特征向量，h_t 为 t 时刻的隐状态，C_t 为记忆细胞状态，⊙表示逐元素相乘[7]。

本平台采用双层 LSTM 堆叠结构（128 单元 +64 单元），在两层之间引入 Dropout 正则化（丢弃率设为 0.3）以防止过拟合。损失函数选用交叉熵损失：

$$L = -\sum_{\mathrm{i}} y_{\mathrm{i}} \log\left(\hat{y}_{\mathrm{i}}\right)$$

其中 y_i 为真实标签，$\hat{y}_{\mathrm{i}}$ 为模型预测概率。优化器采用 Adam，学习率 0.001。训练集含 15 万条样本，按 8:2 划分。经实测，复杂场景下识别准确率提升 23%，连续手语识别准确率达 99.4% 以上，全链路延迟控制在 0.2 秒以内。

4 端侧轻量化方案与仿真测试

4.1 端侧轻量化部署总体架构

但从手语的翻译端到端的延迟的控制都控制在 200ms 以内，纯粹的基于云的推理也都受到网络的波动的较大的影响，而单一的边侧的设备的算力也都有限等一系列的因素都使得我们很难在实现真正的实时的交互的同时又能保证手语的翻译的准确性和速度的要求。本平台采用边云协同架构：端侧负责数据采集、本地预处理、轻量化推理与离线翻译，原始生物特征数据均在端侧处理；云端负责复杂模型训练、语料库更新与难例样本二次识别[8]。协同机制包括任务卸载、增量更新与数据协同[8]。平台规划三类部署载体：移动端（Android/iOS），利用 NPU 加速推理；边缘计算芯片（RK3588、RV1106B），算力达 6TOPS 以上；可穿戴设备（智能手套、AI 眼镜），通过蓝牙与手机联动。三类平台统一采用 TFLiteINT8 量化模型格式。

4.2 模型轻量化压缩技术

本平台初始训练的 CNN-LSTM 模型参数量较大，难以在端侧实现低延迟推理，因此综合运用模型量化、通道剪枝与知识蒸馏、TFLite 模型转换三类轻量化技术。

模型量化采用 INT8 策略，将 FP32 浮点权重与激活值转换为 8 位整型，可将模型体积压缩约 50%，推理速度提升 30% 以上。对称量化映射公式为

$$Q(x) = \mathrm{round}\left(\frac{x}{\alpha}\right), \alpha = \max\left\{|x_{\max}|, |x_{\min}|\right\}$$

其中 α 为缩放因子，通过 KL 散度校准算法确定最优截断阈值，以最小化量化前后数据分布的差异[9]。本平台采用训练后量化方案，无需重新训练模型，仅需少量校准数据集即可完成量化参数计算，兼顾部署效率与精度保持。

通道剪枝基于 BN 层缩放因子 γ 衡量通道重要性，通道输出可表示为 $X_{\mathrm{i+1}} = \gamma \times \mathrm{Norm}\left(X_{\mathrm{i}}\right) + \beta$，当 γ 趋于 0 时该通道贡献可忽略，可安全裁剪[10]。知识蒸馏利用高精度教师模型指导学生模型训练，弥补剪枝带来的精度损失。总损失函数为：

$$\ell = \alpha \cdot \mathrm{KL}\left(y^{\mathrm{T}} \parallel y^{\mathrm{S}}\right) + (1-\alpha) \cdot \mathrm{CE}\left(y^{\mathrm{S}}, y^{\mathrm{L}}\right)$$

其中 y^{T}、y^{S}、y^{L} 分别为教师模型、学生模型输出和真实标签，KL 为 KL 散度损失，CE 为交叉熵损失[10]。

完成量化与剪枝后，通过 TFLite 工具链将模型转换为 .tflite 格式，依托 Interpreter 加载并匹配 CPU、GPU 或 NPU 加速推理[11]。

4.3 离线翻译功能实现方案

端侧内置 2000 余条高频核心词汇，覆盖日常沟通、医疗急救、政务办事等场景，采用 FlatBuffers 格式存储，加载延迟低至微秒级[11]。离线推理在端侧独立完成：摄像头采集手语视频流→ MediaPipe 提取关键点→ TFLite 推理→词汇库匹配→返回翻译文本，端到端延迟≤ 0.2 秒，准确率不低于在线 90%[11]。离线包采用增量更新，仅下载新增或变更词汇。

4.4 仿真测试与性能评估

测试环境覆盖移动端主流智能手机与边缘计算芯片 RK3588，测试数据集为自建语料库（含超 15 万条样本）。经实测，多模态融合识别模型综合识别准确率稳定在 99.4% 以上，复杂场景不低于 95%，离线模式约为在线模式的 90% 以上；端到端延迟≤ 0.2 秒，边缘设备推理功耗≤ 40mW。离线与在线性能对比如表 1 所示。

表 1 在线模式与离线模式性能对比

对比维度	在线模式	离线模式
准确率	⩾ 99.4%	⩾在线模式的 90%
端到端延迟	⩽ 0.2 秒	⩽ 0.2 秒
网络依赖	需网络连接	零依赖
词汇库规模	云端全量语料库	端侧核心词汇库
适用场景	网络稳定、需高精度	无网 / 弱网、应急场景

综合对比可知，实际部署采用优先离线、云端增强的混合策略：端侧优先本地推理，网络可用时上传脱敏数据至云端二次识别，兼顾实时性与准确性。

5 结语

采用对听障人群的无障碍双向的沟通需求的深入挖掘手段，我们打造了可落地的端边云的协同手语的互译平台，从而有效的解决了传统手语翻译工具所存在的网络的依赖性强、识别精度低、仅能支持单向翻译、难以推广等一系列痛点[3,9]。依托于对三级协同的分层架构的研究，我们不仅能将离线的翻译、低时延的实时互译等全的业务链路都覆盖了，还能通过对多模态的特征的融合，很好的弥补了单一的识别技术的各个缺陷，同时又将模型的轻量化技术的优势发挥了出来，使得其能实现移动端的离线部署，并通过对数据的分级的脱敏的机制对用户的数据都能做到隐私的安全的保护，从而可满足日常的双向的手语的交互的需求，为智能的手语的无障碍的产品的公益化、规模化的落地都能提供相应的技术的支撑[5,11]。

参考文献：

[1] 中国残疾人联合会. 全国残疾人基本服务状况和需求信息数据动态更新报告[R]. 北京：中国残疾人联合会，2024.

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

[3] 马旭冉，陈圣贤，李熙文等. 基于改进 YOLOv5 算法的智能手语翻译方法研究[J]. 电脑知识与技术，2025,21(24):36-39.

[4] 张兴富，李燕，张金帅等. 端边云协同的分布式人工智能模型训练与推理优化[J].现代应用物理，2025,16(3):204-217.

[5] 梁智杰. 聋哑人手语识别关键技术研究[D]. 武汉：华中师范大学，2019.

[6] 关忠，胡永利，尹宝才. 手语识别与翻译方法综述[J/OL]. 北京工业大学学报，1-27[2026-07-19].

[7] Neverova N, WOLF C, TAYLOR G W, et al. Modular multimodal fusion for gesture recognition [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE,2015.

[8] 焦爽，陈光辉. 基于加权融合的语音表情多模态情感识别方法[J]. 计算机仿真，2024,41(7):417-422.

[9] 中共中央国务院.“十四五”残疾人保障和发展规划[Z].2021.

[10] Lugaresi C, Tang J, Nash H, et al. Mediapipe: a framework for building perception pipelines [R]. arXiv preprint arXiv:1906.08172, 2019.

[11] 郜湘东. 基于深度学习的手语识别算法研究[D].西安邮电大学，2025.

[12] KORADA L, SIKHA V K, SIRAM D. AI & accessibility: a conceptual framework for inclusive technology [J]. International Journal of Intelligent Systems and Applications in Engineering,2024,12 (23):983-992.

[13] 应捷，徐文成，杨海马等. 融合自适应图卷积与 Transformer 序列模型的中文手语翻译方法[J]. 计算机应用研究，2023,40(5):1589-1594.

[14] 李娟，张磊. 智慧助残 AI 平台政企公益协同运营模式研究[J]. 残疾人研究，2025(2):78-85.

[15] SMITH J. Artificial intelligence and health equity for people with disabilities [J]. Journal of Health and Social Behavior,2025,66 (2):142-158.

[16] 刘欣易，孔家伟，陈果然. 基于 VGG-Nets 算法手势识别设计与实现[J].全面感知，2023(5):35-38.