计算机视觉技术在开放教育中面向特殊人群的应用 *

文◆延安开放大学　党　英

引言

信息技术的快速发展推动了开放教育的普及和创新。虽然开放教育在推广教育公平和普及知识方面取得了显著进展，但特殊人群（如听障、视障人士等）的参与度仍然有限。随着计算机视觉技术的不断发展，手语识别技术逐渐成为国内外研究的热点，目的是将手语转换为文本信息，从而消除沟通障碍[1]。国内外专家普遍认为，从传统技术到智能识别方法的演变，不仅为手语识别的技术融合提供了新思路，还拓展了特别是面向听障人士的手语识别的应用场景[2]。同时，文本检测与识别技术主要结合卷积神经网络（CNN）、循环神经网络（RNN）等深度学习模型为场景文本检测提供技术支持，但在模型复杂性与场景适应性方面仍存在改进空间。丁慧洁等[3]发现现有的读屏软件是基于文本的读取方式，而无法识别图片等类型的非文本信息，这一发现揭示了技术的局限性，为视障人士远程学习问题提出了改进方向。

1 计算机视觉技术的应用优势

计算机视觉技术在开放教育中的应用（特别是针对特殊人群手语和图像转语音系统的实现），提升了开放教育的无障碍性和包容性。手语是听障人士和健听人士之间沟通的桥梁，但复杂多样的手语动作增加了手语的学习成本，给听障人士和健听人士交流带来了阻碍。而计算机视觉技术的手语识别和翻译技术消除了沟通障碍。手语识别是将手语所表达的一个或多个手语词逐个识别，并映射为对应的标注文本；手语翻译则是在其基础上进一步将手语标注序列转换为符合自然语言语序的句子。通过视觉特征提取和上下文信息捕捉，手语识别与翻译能够将输入的手语视频转化为文本输出。

图像文本转语音为视障人士的学习提供了诸多便利。许多学习材料和参考书籍通常以图片和文字的形式呈现，视障人士难以直接获取这些信息。而通过图像文本转语音模块，能够将学习材料能够迅速转换为语音，视障人士可以随时随地获取所需知识。这种方式增强了视障人士的学习主动性，使他们能够自主参与学习。

2 平台系统构建

2.1 系统架构设计

系统架构主要由手语识别、图像转文本和文本转语音 3 个核心模块组成。输入类型包括手语视频和图像两种输入形式。第一，手语识别模块。首先，手语识别需要采集大量的手语视频素材，包括不同性别、年龄以及手势习惯的手语使用者，并对素材进行去噪和增强对比度等预处理。其次，采用 OpenPose 算法进行关键点检测，提取出手语的关键点特征。再次，结合 CNN 和长短期记忆网络（LSTM）来捕捉手语的空间与时间特征。最后，输出为文本。第二，图像转文本模块。首先，对输入的图像进行去噪、增强对比度和裁剪等预处理，以提高后续处理的效率和准确性。其次，基于目标检测的受控场景文字定位方法，从预处理后的图像中精确提取文字区域的关键特征。最后，利用序列学习框架进行文字行的识别，该框架使用 CNN 提

*【基金项目】陕西开放大学 2024 年度科学研究项目“计算机视觉技术在开放教育中对特殊人群的应用”（2024KY-B22）
【作者简介】党英（1997—），女，陕西渭南人，硕士，研究方向：计算机应用、开放教育。

取文字区域的空间特征，并生成相应的文本输出。第三，文本转语音模块。首先，通过文本前端将获得的原始文本转换为字符或音素序列。其次，利用声学模型将这些字符或音素转换为声学特征。最后，利用声码器将这些声学特征合成为波形输出生成的语音。

2.2 手语识别

传统手语识别方法根据获取手语信息的方式，可分为基于可穿戴设备的手语识别方法和基于视觉的手语识别方法。基于可穿戴设备的手语识别方法相比基于视觉的手语识别方法，只需普通摄像设备即可完成数据采集和处理。但传统手语识别方法主要依赖人工特征提取，提取的特征泛化性较差，仅能应用于特定场景且算法复杂。而基于深度学习的手语识别逐渐取代传统方法成为当前研究的主要方向。该方法能显著提升手语识别的准确性和拓宽应用范围。

2.2.1 数据采集与标注

实现数据采集需要构建多种手语手势的数据集。该数据集不仅要包含手语的多样性，还要能够反映出现实交流的复杂性，如手势的形状、手部的运动轨迹等。数据采集主要通过摄像头录制视频或拍摄图像的方式获取，在录制视频时需考虑不同的光线条件、背景环境以及摄像设备的角度和分辨率，以保证数据的适用性。同时，为了应对多样化的实际应用场景，数据集的类别需要包含不同性别、不同年龄和不同手势习惯的人的手语手势信息。

完成数据采集后要对数据进行准确标注。标注过程要将每个手势与其对应的手语含义进行分类和标识，同时需要记录手势开始和结束的时间、手势的位置变化等关键特征。此外，引入关键点检测算法来辅助标注过程，以提高标注的效率和精度。

2.2.2 图像预处理及特征提取

（1）图像预处理。首先，通过高斯滤波等方法，有效去除图像中的随机噪声和冗余信息。其次，通过调整图像的对比度、亮度和锐化细节等，改善图像质量使手部轮廓和手势特征更加清晰。其中，手部检测是图像预处理的核心，通过 Faster R-CNN 目标检测算法对视频中的手部区域进行实时检测和定位。

（2）特征提取。在经过预处理后的图像上，进行关键点检测与特征提取，以捕捉手部的关节和手指位置，从而提取能够描述手势的重要特征信息。OpenPose 关键点检测算法采用自上而下的方式，通过特征图上的热力图和局部亲和场生成关键点的候选位置，并解析这些候选点之间的几何关系和约束，从而能够准确定位出手部关键点的坐标，标记出手部关节的坐标并准确追踪手指的运动。为了提高特征提取的性能，通过多模态优化并利用深度摄像头获取手部三维关节坐标，进一步提升特征提取的精准度。同时，在光照条件复杂或存在遮挡的场景下，通过融合 RGB 图像与深度图像，提高关键点检测的稳定性，从而增强模型在复杂环境下的适应能力。

2.2.3 深度学习模型的设计

传统的 CNN 在处理长期依赖关系和时间序列数据方面存在局限性，而 LSTM 作为递归神经网络的一种变体，能够捕捉序列数据中的长期依赖关系，显著提升模型的预测能力。因此将 CNN 和 LSTM 相结合，能有效提高模型对图像序列的处理能力。

在手语识别中，经关键点提取后，识别模块采用多层次结构模型，能够同时捕捉手语的空间特征和时间特征。模型设计基于 CNN-LSTM 的混合架构。首先，通过 CNN 对预处理后的手势图像序列进行特征提取，捕捉手指的弯曲角度、相对位置、手掌形状等空间特性。其次，将 CNN 提取的高维空间特征输入 LSTM，对手势动作的时间序列特征进行建模，捕捉手语动作在时间维度上的动态变化。再次，在模型中引入注意力机制，提升对手部动作的捕捉能力。最后，将模型最底层设计为全连接层，并将其作为时间序列特征映射的具体手语类别，结合 Softmax 函数实现多分类任务。

2.2.4 模型训练与优化

模型的训练需要使用大量标注好的手语数据集，并通过调整学习率和批量大小等参数来优化性能。在验证阶段，使用与训练数据分离的验证集，通过评估损失函数、准确率和召回率等来评估模型的性能。在测试阶段，需要使用没有训练的数据集，评估模型的泛化能力，特别是在手语识别的实际应用中，实现手语识别的实时性和精准性。因此在测试过程中，不仅要关注模型的识别准确率，还需测试实时响应性能。

2.3 图像转文本

随着计算机视觉技术的发展，图像文本提取转语音（ITTS）技术已被广泛应用于电子书语音阅读、辅助阅读工具等场景。该技术适合视力障碍者和阅读困难人群，能够从图像中提取文本信息，将其转换为语

音，从而提高他们获取信息的便捷性。

光学字符识别（OCR）和文本转语音（TTS）技术的出现为 ITTS 技术的广泛应用提供了一定的技术支持[4]。OCR 通过 CNN 和 RNN 实现了对复杂场景文本的高精度识别，即使在低分辨率、光照不均的情况下也能够准确提取出图像中的字符信息。同时，TTS 技术借助生成对抗网络等深度学习框架，形成了自然的语音，使生成的语音在语调、情感和流畅性上与真人声音更加接近。

2.3.1 图像预处理

实现图像文本提取转语音的首要步骤是对图像进行预处理。首先，对输入图像的大小、分辨率和颜色等进行优化。其次，通过高斯滤波方法去除图像中的噪声，降低环境对识别效果的干扰影响，同时利用边缘增强技术突出图像中的边缘区域。再次，采用直方图拉伸技术，增强文本区域与背景之间的对比度，使文字在复杂背景下更加清晰。最后，通过裁剪图像去除无关的背景区域，不仅降低了计算复杂度，还减少了背景信息对文字检测过程的干扰。

2.3.2 光学字符识别

OCR 技术是一种基于计算机视觉的文本提取技术，通过检测暗、亮的模式确定其形状，用字符识别方法将形状翻译成计算机文字[5]。传统的 OCR 技术采取自上而下的切分式，适用于版面规则、背景简单的场景。而基于深度学习的 OCR 技术能够处理复杂的背景、多样的字体和多角度排布的文本。该过程首先对输入图像进行预处理，其次进行文字检测。选择基于 Faster R-CNN 的受控场景文字定位方法，该方法适用于结构化文本场景，能够在规则版面中精确识别文字区域；选择基于 FCN 的非受控场景文字定位方法，能够处理复杂背景、多种字体和不规则排布的文字区域。此外，通过基于序列学习的识别框架进行文字行的识别，该框架利用 CNN 提取文字区域的空间特征，再通过 RNN 建模字符的顺序关系，实现对文字行的高效识别，最后输出文字识别结果。因此，基于深度学习的 OCR 技术能显著提升复杂场景文本的识别能力。

2.3.3 文本处理

文本处理是确保转语音效果流畅且准确的关键环节。首先，去除提取的文本内容中多余的字符、符号以及可能干扰语音输出的冗余信息。其次，调整文本的排版格式，使其符合语言的逻辑顺序和语法规则。最后，对文本内容进行拼写错误的检查和修正，从而提升语音输出的准确性，避免因拼写问题导致生成的语音出现错误或不连贯。

2.4 文本转语音

文本转语音（TTS）技术是一种从文本到语音端到端的语音合成技术。首先，输入原始文本，通过文本前端模块将其转换为字符或音素。在这一过程中，需要对输入文本进行断句、词性分析和归一化处理。其次，利用声学模型将字符或音素转换为声学特征，如线性频谱图和线性预测编码特征等，这些特征能够精确描述语音的频率分布和音色特性，为语音生成提供基础数据。最后，声码器将生成的声学特征转换为波形输出。该过程依赖深度学习模型的支持，使生成的语音自然流畅，接近人类语音的真实效果。

结语

本文主要研究了计算机视觉技术在开放教育特殊人群中的应用，为提升开放教育的包容性提供了技术支持。首先，分析了计算机视觉技术的优势。其次，针对开放教育的教学场景，研究设计了手语视频与图像转语音系统。系统结合手语识别与文本转语音技术，实现手语视频到音频或文本的双向转换，并引入多模态融合算法，提高复杂环境下的识别准确性，为听障群体搭建起高效的“沟通桥梁”。未来，随着算法迭代与硬件升级，该技术体系有望进一步拓展应用场景，如结合智能教学终端实现个性化学习支持或融入远程协作系统促进特殊人群的社交参与，从而推动开放教育真正迈向“全纳、公平、高质量”的发展目标。

引用

[1] 李宇楠,耿熙,苗启广.基于计算机视觉的手语识别与翻译研究综述[J].微电子学与计算机,2025,42(6):15-36.

[2] 孟巾凯,彭健钧,肖智东,等.模块化连续手语识别算法及技术综述[J].小型微型计算机系统,2024,45(10):2428-2441.

[3] 丁慧洁,张晓霞,李芷筠.以英国开放大学视障人士学习为视角分析我国视障人士远程学习的障碍及对策[J].中国多媒体与网络教学学报(上旬刊),2021(2):54-56.

[4] 陈曦,陆利坤,王彤,等.基于视觉语言的文字识别方法综述[J].北京印刷学院学报,2024,32(6):35-43.

[5] 张亚停.多语种TTS文本语料库构建关键技术研究[D].北京:北京印刷学院,2023.