基于视觉方法的手语翻译系统

项目背景

通过机器将手语翻译成自然语言可以使聋哑人更容易与非听障人群交流。目前手语识别主要有数据传感手套和视觉识别两种方式。本项目将通过视觉方法实现手语识别、翻译，并以文字或语音的形式输出，辅助聋哑人士与普通用户交流。手语翻译系统可应用于智能家具、机器人交互、游戏控制、XR 耳机智能设备交互等领域。

随着计算机视觉和机器学习技术的发展，目前已有大量关于手语识别的研究。深度学习技术，尤其是 CNN 和 RNN 模型，已经被广泛应用于手语识别。目前的挑战主要在于：背景环境及灯光的变化、不同人的手型差异、动态手语识别、识别准确性、鲁棒性、响应速度。

项目结构

1. 采集图像数据、图像数据集增强、数据预处理，得到训练数据。

2. 图像裁切算法，检测手部边界框。目前成熟的网络架构优化有：使用卷积神经网络检测、RGBD 多通道融合。本项目目前采用 Mediapipe 库的手部坐标模型约束扫描以获取手部边界框，后续考虑提取手部边界以增加背景鲁棒性。

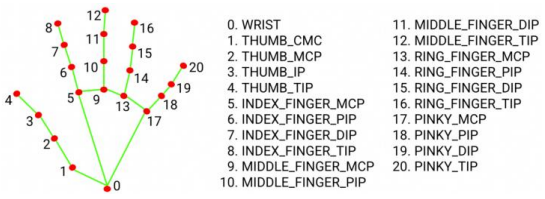
3. 图像识别算法，通过机器学习训练模型，匹配特征。目前成熟的算法有：DTW（动态时间规划调整，比较序列相似度）、HMM（根据概率分布建模）、2DCNN（提取图像空间特征，静态手语识别）、3DCNN（提取图像时空特征，动态手语识别）。本项目采用 2DCNN 匹配数据集和被检测图像的灰度值分布以识别手势。

4. 以文字、语音等形式输出结果。

注：为便于测试，目前的数据使用数字 0-9 进行录制和测试，后续可扩展到更多数据。

传统方法——二维约束法计算向量夹角

使用 mediapipe 的手部坐标模型，计算出手指指尖与手掌夹角（以拇指为例，（0,2）向量和（3,4）向量的夹角）。

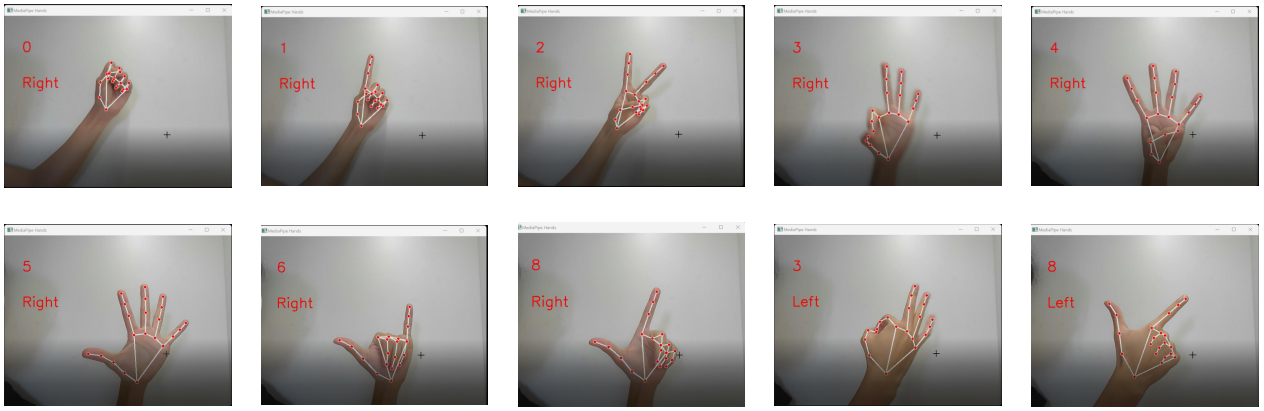


由于角度与试验者手型高度相关，且使用的是传统的计算方法，利用夹角是否落在范围内，判断手指开合。在简单手势时能得

到正确结果，但鲁棒性非常低，且无法拓展至复杂手势（如数字 9 无法判断食指开合，与 0 情形会相同）。

不过，由于该方法可得到唯一结果（或 unknown），目前已加入文本转换为语音功能，使用 python 中 pyttsx3 库实现。

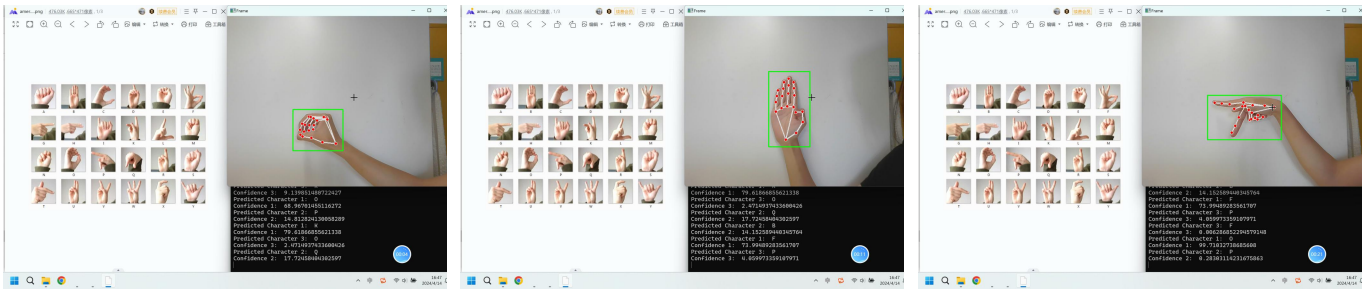
以下是实验结果：



采用 CNN 深度学习模型识别手势-使用开源数据集测试

使用了 tensorflow 中的 CNN 模型，使用了三层卷积层，卷积核大小为（3,3），输出通道数分别为 75,50,25。激活函数使用了 relu。训练了 20 个 epoch。使用开源数据集训练（已预处理），在测试中以同样方式预处理图像数据（图像提取分割，模糊为 28*28 像素值，展开成灰度值大小的一维数组）。结果欠佳，尚能识别出部分手势，准确度低，鲁棒性低，考虑是模型参数与数据集大小不匹配，出现过拟合，在后续自采样的数据集中尝试优化。

以下为实验结果：



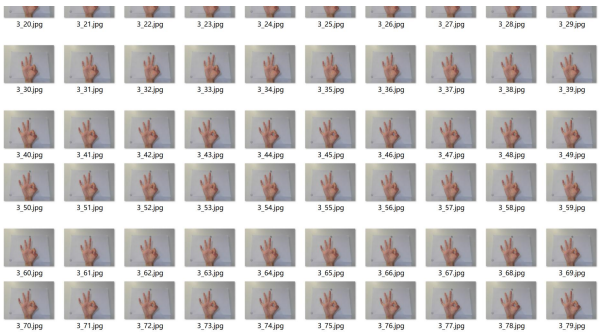
创建数据集

python 调用摄像头拍摄图像数据，并加标签。（采用了白色和黑色背景，最后准确率无明显差异）

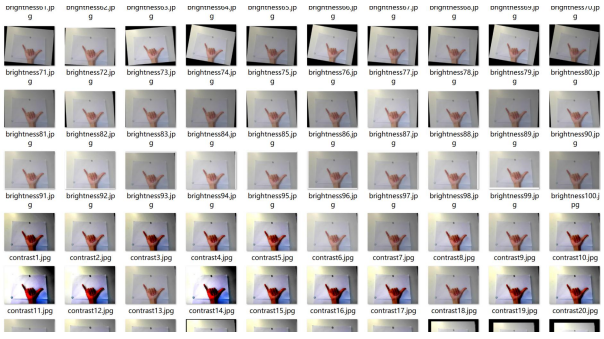
采用了图像数据集增强的方式，对图像做了随机对比度调整、水平翻转、随机亮度调整、随机角度旋转、随机仿射变换、随机错切变换、HSV 数据增强、随机平移、随机放大。迭代运行使其得到二重图像变换。丢弃部分无法识别的图像后，将训练数据扩充到了约 4200 张/标签，训练中的测试数据约 500 张/标签，同时增加训练数据集的多样性，提升系统鲁棒性。

预处理图像数据。扫描 mediapipe 手部坐标模型，获得手部区域估计，对图像裁切，以提高数据效率和压缩后的图像清晰度。将图像转为灰度图，以 56*56 像素处理图像，转换为灰度图像，用图像灰度值运算，降低运算复杂度。展开到一维数组，与标签一同保存到 csv 文件中以便后续调用。使用 panda 库对上述保存的预处理文件进行打乱，防止训练时出现阶段性遗忘。

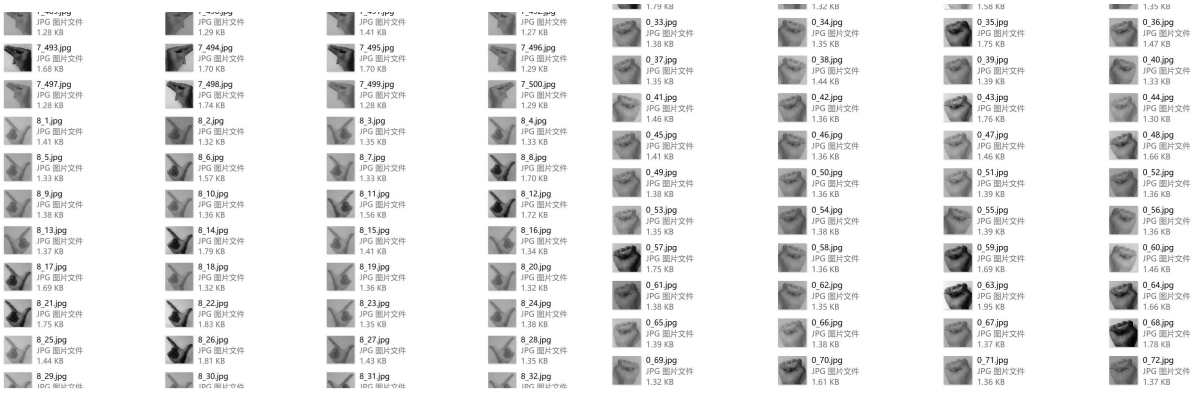
根据从预处理文件还原的图像数据观察，可以认为数据信息量保留程度合适。以下为部分图像数据预览：



拍摄得到的图像数据



图像增强后的图像数据

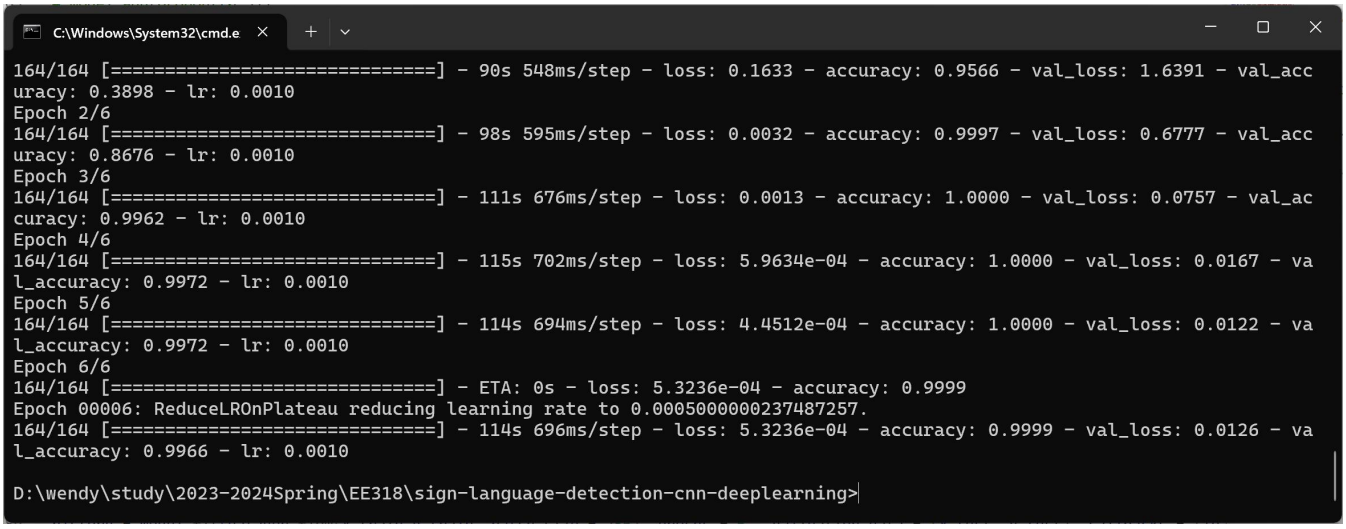


预处理后重新还原的图像数据

采用 CNN 深度学习模型识别手势-使用上述制作的数据测试

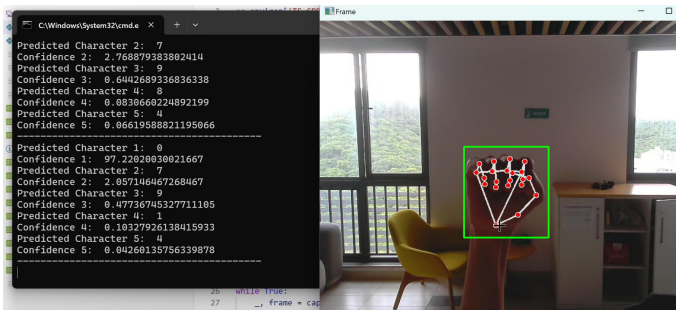
使用了 tensorflow 中的 CNN 模型，使用了三层卷积层，卷积核大小分别为（7,7），（5,5），（3,3），输出通道数分别为 35,20,10。激活函数使用了 LeakyReLU。训练了 6 个 epoch。

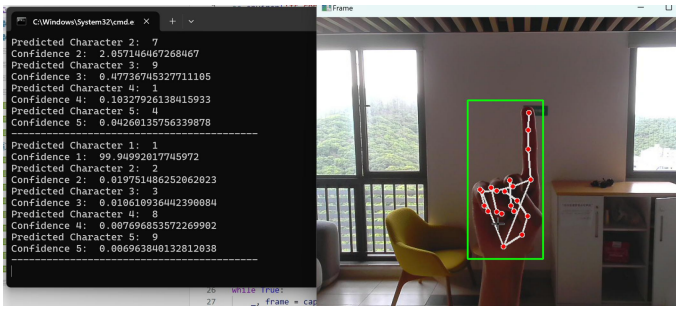
训练结果如下：

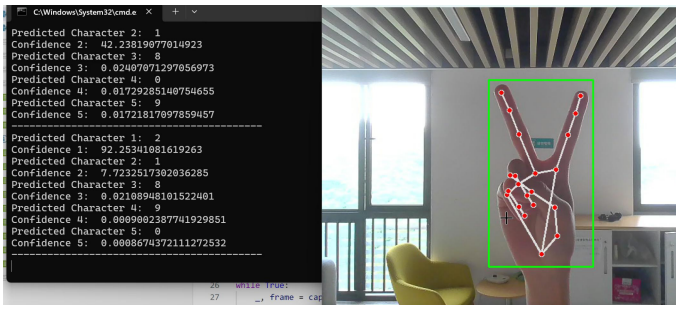


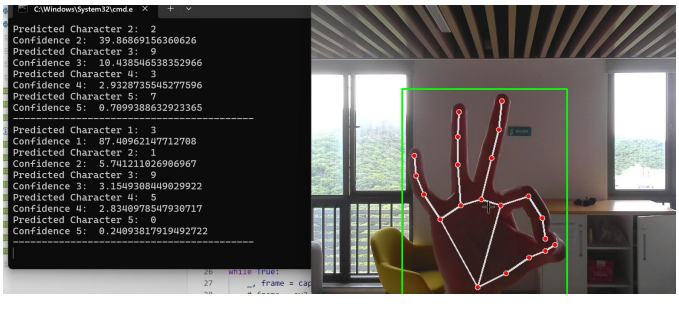
可以观察到，取当前参数时，准确率已经可以接受。

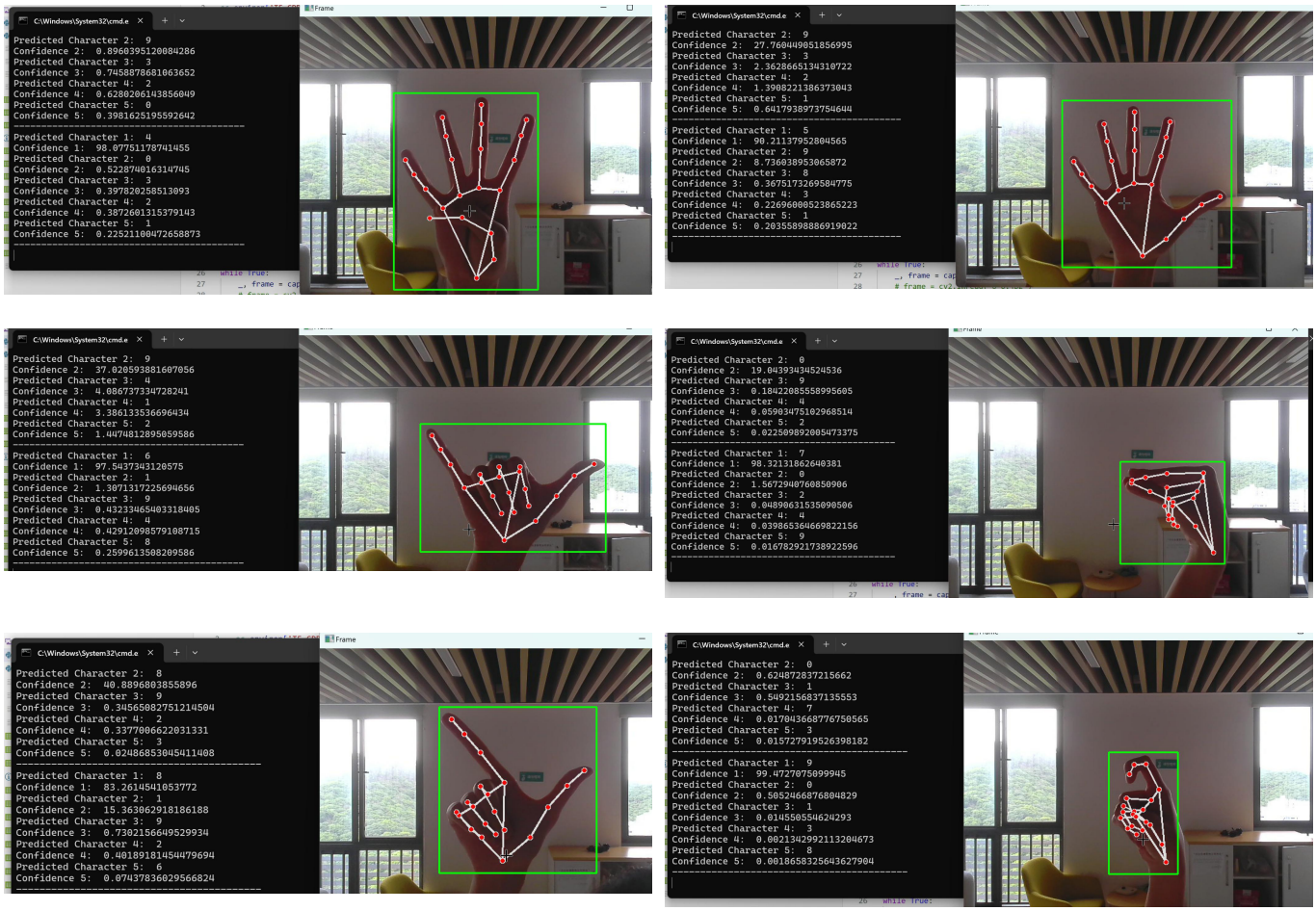
使用训练的模型，运行图像识别测试程序，结果如下：











数字 0-9 已都可以识别，且能够一定程度的抗环境干扰，但准确率并非每次都能很高，偶有出错情况。出错可能由于手势相近，在仅有 56*56 像素的图像上，叠加了背景干扰因素转成灰度值后，部分图像容易相近。而模型可能过拟合，故相近的手势容易被估计为其他情况。响应速度已可以接受。采用他人手势试验，绝大部分能得到正确结果，部分手势仍存在问题。

存在的问题

（1）由于目前的 CNN 匹配的是裁切的正方形图像的灰度值，当背景颜色（白色变黑色）和光照条件变化时，模型的鲁棒性差。

（2）训练可能存在过拟合，训练参数与样本量不匹配，还未找到合适的值，识别结果准确度欠佳。

（3）偶尔识别会失败，置信度最高输出结果不一定是正确结果，无法确认唯一识别结果，不便进行下一步处理（文字和语音实时显示）。

改进方向

（1）手部颜色通常与背景颜色有明显差异，可以利用 RGB 通道对手部边界提取处理，去除背景，提高背景鲁棒性。

（2）可以使用手部骨架信息处理，匹配手指骨架，减少光照和背景对识别结果的影响。

（3）实际的手语不止手掌信息，还需要叠加手臂信息以传达正确意义。考虑结合手臂信息，期望支持更广泛的手语识别。

（4）CNN 模型仅在比较像素值，且目前只比较了灰度值，准确度较低。考虑优化深度学习模型，考虑其他深度学习模型、其他识别算法，提高识别准确率。

（5）不同人的手型有差异，容易降低识别准确度。考虑增加不同人、不同环境、不同姿态的数据，增加数据多样性。