

**《电子科学创新实验III》**

**课程报告**

**题 目：** 照猫画虎之手语识别系统

**姓 名：** 张旭东

**学 号：** 12011923

**系 别：** 电子与电气工程系

**专 业：** 信息工程

“姓名、学号、指导教师、年级与专业、年月日”均用四号宋体打印，不得手写，各栏目下划线需统一长度

**指导教师：** 游昌盛

2024年 1 月 11 日

照猫画虎之手语识别系统

张旭东

（电子系 指导教师：游昌盛）

[摘要]：本文将密室逃脱中摆出相应姿势来获取线索的关卡与手语结合起来，运用神经网络模型和边缘设备搭建手语识别系统，实时识别英文字母的手势，从而让听障者对游戏的参与程度大大提高和享受到游戏的乐趣。该系统除了提高密室逃脱游戏的趣味性之外，还在促进与听力障碍者交流过程中有一定的作用。

[关键词]：手语识别; 神经网络; 英文字母;

**目录**

**1.背景**..............................**5**

**2.解决方案.............................6**

2.1数据集..................................6

2.2模型..................................7

2.3训练与测试效果..................................8

**3.优化.............................9**

3.1模型..................................9

3.2训练与测试效果..................................11

**4.部署与实时检测.............................11**

4.1手部图像的提取..................................11

4.2实时检测..................................13

**5.项目总结.............................13**

**参考文献(宋体四号，加粗)...............................15**

1. 项目背景

密室逃脱是一系列解谜游戏，玩家需要在一定时间内从封闭的房间中逃脱。这类游戏通常包含各种谜题、密码和隐藏物品，玩家需要通过观察环境、找到线索，并解开谜题来成功逃脱。每个关卡都提供独特的挑战，测试玩家的智力和观察力。不同版本的密室逃脱游戏可能有不同的故事情节和场景设置，但核心玩法是相似的。这种游戏类型在手机应用市场和在线游戏平台上都很流行，玩家可以在其中体验解谜的乐趣，并挑战自己的逻辑思维能力。其中，存在这样一类关卡：玩家需要根据前面得到的线索按照一定的规则摆出相应的姿势以通过这关来获得之后的线索。这类姿势可以是身体姿势，也可以是脸部表情，还可以是手部姿势。

手部姿势，又称之为手语。手语是一种通过手势、动作和表情进行交流的语言形式。它通常被用作一种替代或辅助方式，特别是在与听障者的沟通中。手语不仅仅是字母或数字的手势表示，还包括表达情感和构建句子的复杂手势系统。在密室逃脱等需要大量对话和沟通的解谜游戏中，对听障者而言可能带来挑战。为了让听障者也享受到游戏的乐趣，本文设计了基于手语识别的关卡，通过实时识别相应的手势来获取线索进入下一关卡，让听障者对游戏的参与程度大大提高。

手语主要由手指拼写、单词级符号词汇这两部分组成。手指拼写是一种手动的交流方式，用双手和手指拼写单词。每个字母都用指定的手位置表示。单词级符号词汇需要一个大型视频数据集，用于识别单词或字母的整个手势。本文主要解决第一个部分的问题，所采用的方案是基于视觉数据的神经网络。

1. 解决方案

2.1数据集

本文所采用的数据集是Sign Language MNIST(Fig.1).



Figure 1 手语示例

该数据集中每张图片的大小为28\*28\*3，分为训练集和测试集。训练集中含有27455个案例，测试集中含有7172个案例(Table.1)，其中由于手势运动，并不含有字母“J”、“Z”的案例。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字母 | 训练集中的数量 | 测试集中的数量 | 字母 | 训练集中的数量 | 测试集中的数量 |
| A | 1126 | 331 | B | 1010 | 432 |
| C | 1144 | 310 | D | 1196 | 245 |
| E | 957 | 498 | F | 1204 | 247 |
| G | 1090 | 348 | H | 1013 | 436 |
| I | 1162 | 288 | K | 1114 | 331 |
| L | 1241 | 209 | M | 1055 | 394 |
| N | 1151 | 291 | O | 1196 | 246 |
| P | 1088 | 347 | Q | 1279 | 164 |
| R | 1294 | 144 | S | 1199 | 246 |
| T | 1186 | 248 | U | 1161 | 266 |
| V | 1082 | 346 | W | 1225 | 206 |
| X | 1164 | 267 | Y | 1118 | 332 |

Table 1 案例统计

2.2 模型

卷积神经网络（CNN）是一种深度学习模型，主要用于处理和分析具有网格结构的数据，如图像和视频。CNN的核心特性是卷积层，通过卷积操作可以有效地提取图像中的特征，同时减少参数量和计算复杂度。这种神经网络结构在图像处理领域取得了巨大成功，广泛应用于图像分类、目标检测、图像生成等任务。我们在本次应用的核心技术中也采用CNN。CNN的结构如Fig.2所示，典型的结构包括卷积层、池化层、激活层、全连接层等，对于输入图像，可以有效抽取图像内容表征，并进行分类或其他处理。卷积层等特殊结构，可以在控制参数量的前提下，保证良好的图像特征提取能力。

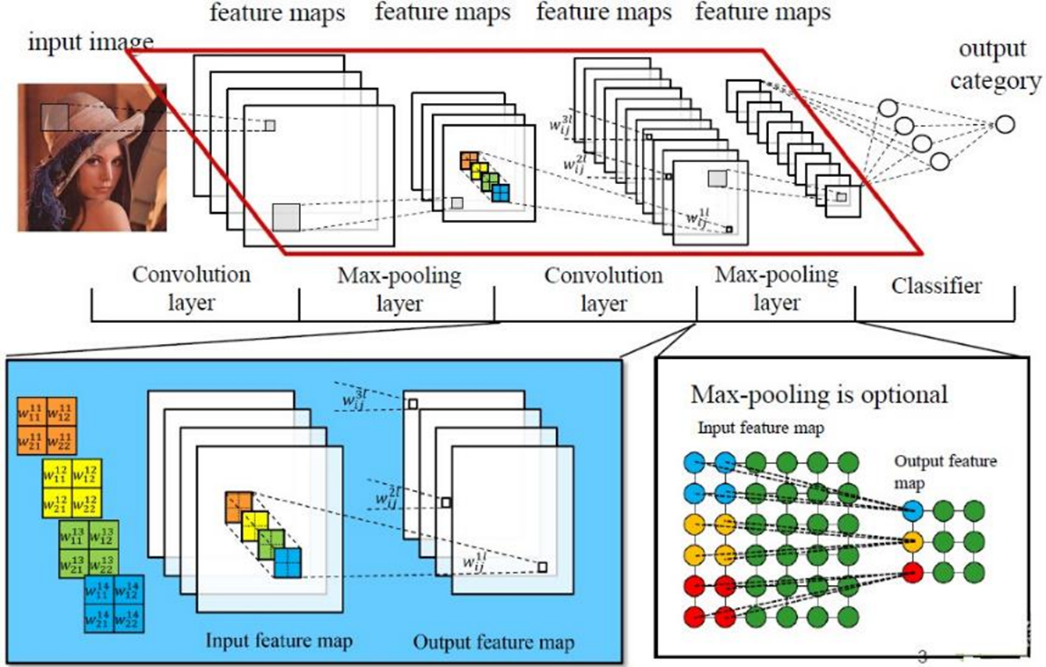


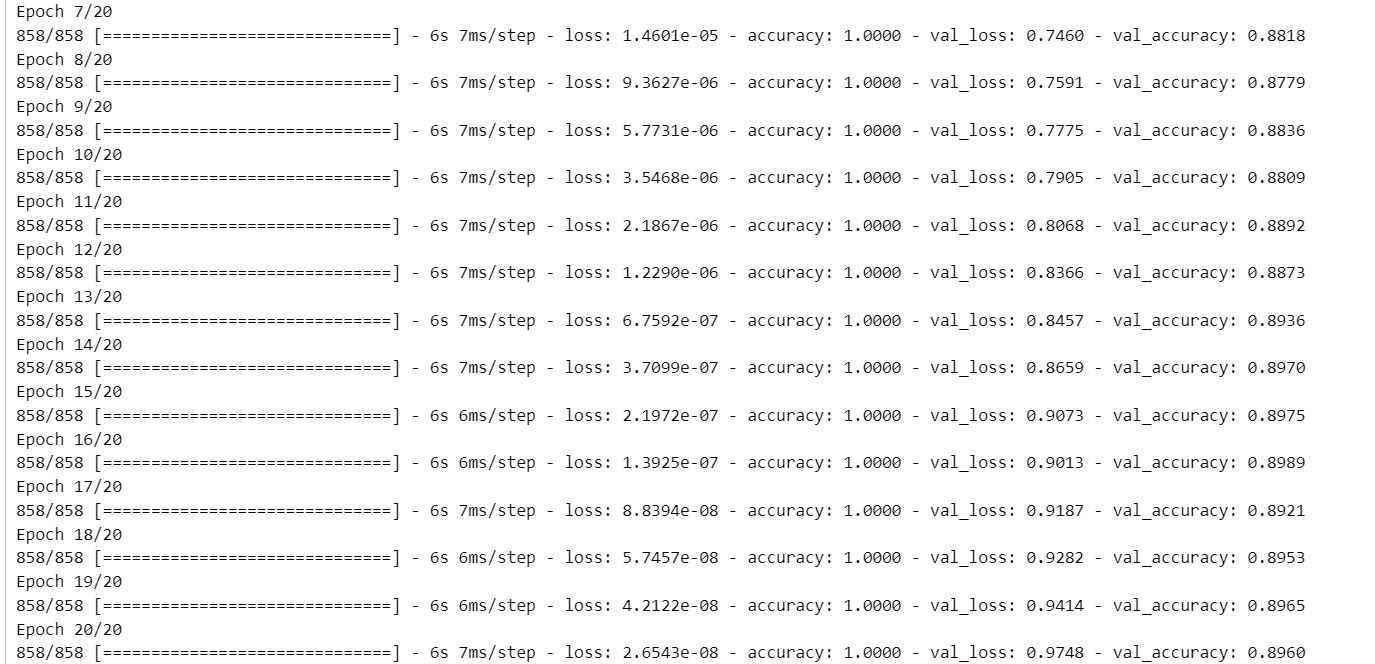
Fig. 2 CNN结构

本文所采用的CNN结构和参数如下所示：

* 卷积层：输入图像28\*28\*1，激活函数ReLu, 卷积核个数32，卷积核大小3\*3
* 池化层：池化窗口2\*2
* 卷积层：激活函数ReLu, 卷积核个数32，卷积核大小3\*3
* 池化层：池化窗口2\*2
* 全连接层：激活函数ReLU，输出维度512
* 分类层：激活函数softmax，输出维度26

2.3 训练与测试效果

在整个数据集上迭代20个轮次，结果如Fig.3所示。随着迭代的进行，在训练集上的准确率可以达到100%，但在测试集上的准确率却低于90%，达不到预期期望。同时，根据查阅资料得知，训练深度CNN会存在梯度消失的问题。反向传播的梯度连乘带来的问题会使非常深的神经网络的梯度很快变为0，这最终会导致整个梯度下降变慢。



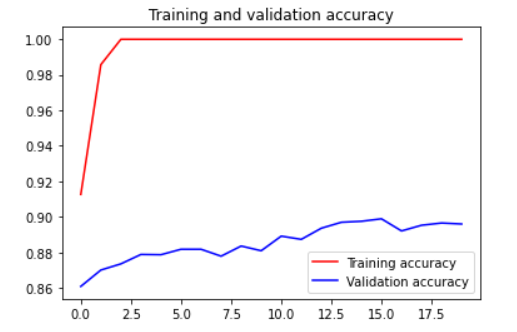
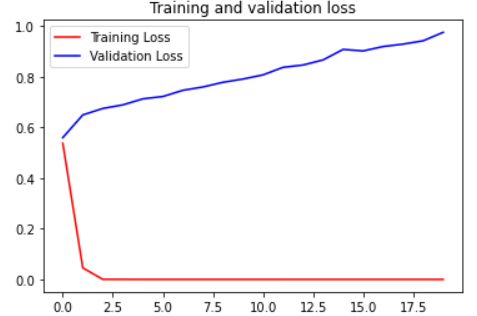
 

Fig. 3 训练和测试结果

解决梯度消失的方法主要有使用激活函数ReLU、批标准化、残差连接等。有一些特殊结构的神经网络，也可以大程度缓解这个问题，如ResNet。

1. 优化

3.1模型

ResNet 是 Residual Networks 的简称，由残差块组成，残差块的核心组件是跳跃连接。跳跃连接，也称为快捷连接，让神经网络跳过某些层并将一层的输出馈送到神经网络中另一层的输入。它能帮助模型避免乘以中间跳过的那些层的权重，从而有助于解决梯度消失的问题。ResNet结构示意图如Fig.4所示。ResNet核心效果有两个：一是跳过层来解决梯度消失的问题，二是通过让模型学习恒等函数来确保最高层的性能至少与最低层一样好。

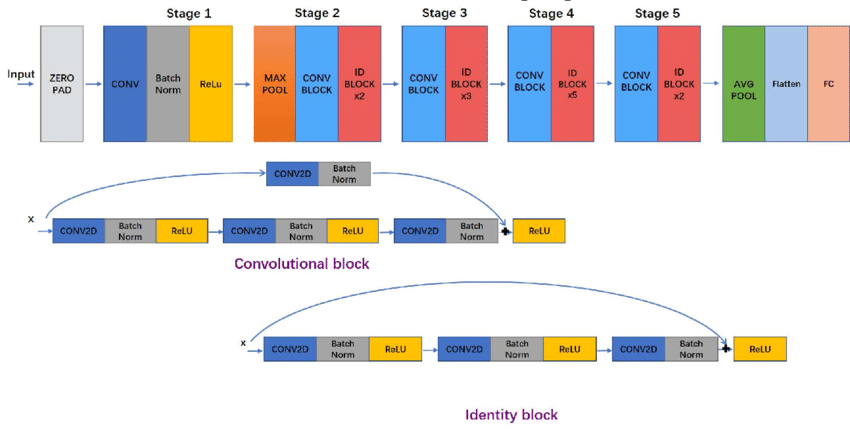


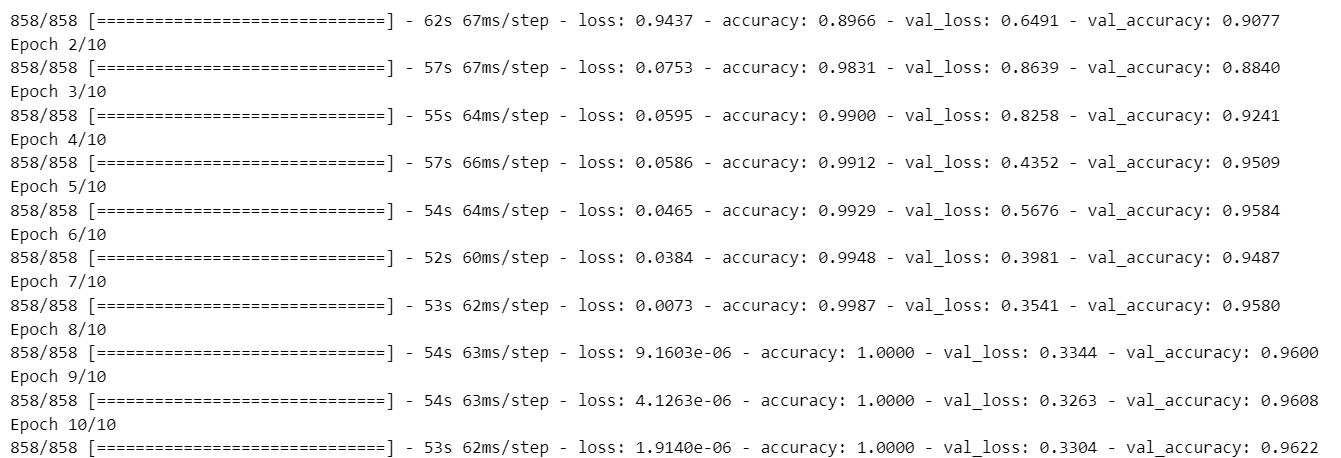
Fig.4 ResNet结构示意图

本文所采用的ResNet结构和参数如下所示：

* 第一阶段：卷积层(核个数64，核大小5\*5，步长为1)➡批标准化➡激活层(ReLU)➡池化层（核大小3\*3，步长为2）
* 第二阶段：卷积块(核个数64、64、256，核大小3\*3，步长为1)➡恒等块(核个数64、64、256，核大小3\*3，步长为1)➡恒等快(核个数64、64、256，核大小3\*3，步长为1)
* 最后阶段：池化层(核大小2\*2)➡展平➡分类层(激活函数softmax，输出维度26)

3.2 训练与测试效果

在整个数据集上迭代10个轮次，结果如Fig.5所示。



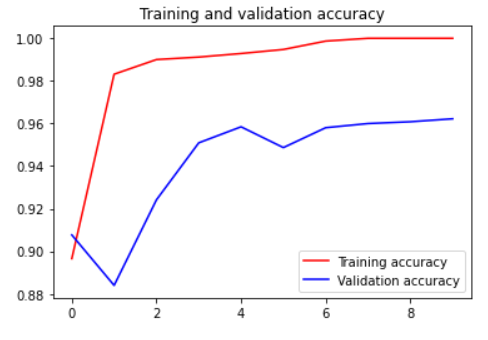


Fig.5 训练和测试结果

对比CNN和ResNet的测试效果可知，ResNet在测试集上的准确率比前者高，为96.22%。这也在一定程度上说明了ResNet 的结构确实能够带来效果上的提升。

1. 部署与实时测试

4.1手部图像的提取

由于在实际运用的环境中摄像头所捕获的图像或视频流中不单单仅含有手部这一个对象，还可能会含有头部等其他对象。此前训练模型的数据集中只含有手部这一个对象，因此在实际运用情况中需要从捕获的图像中提取出手部图像再进行手势识别。在提取手部图像这个过程中，本文借用了MediaPipe这个开源框架。

MediaPipe在训练手掌模型中，使用的是单阶段目标检测算法SSD。同时利用三个操作对其进行了优化：1.NMS；2.encoder-decoder feature extractor；3.focal loss。NMS主要是用于抑制算法识别到了单个对象的多个重复框，得到置信度最高的检测框；encoder-decoder feature extractor主要用于更大的场景上下文感知，甚至是小对象(类似于retanet方法)；focal loss是有RetinaNet上提取的，主要解决的是正负样本不平衡的问题，这对于开放环境下的目标检测是一个可以涨点的技巧。利用上述技术，MediaPie在手掌检测中达到了95.7%的平均精度。在没有使用2和3的情况下，得到的基线仅为86.22%。增长了9.48个点，说明模型是可以准确识别出手掌的。

MediaPipe通过对整个图像进行手掌检测后，使用手部关键点模型通过回归对被检测手部区域内的21个三维手部关节坐标进行精确的关键点定位，即直接进行坐标预测。标记点如Fig.6所示。

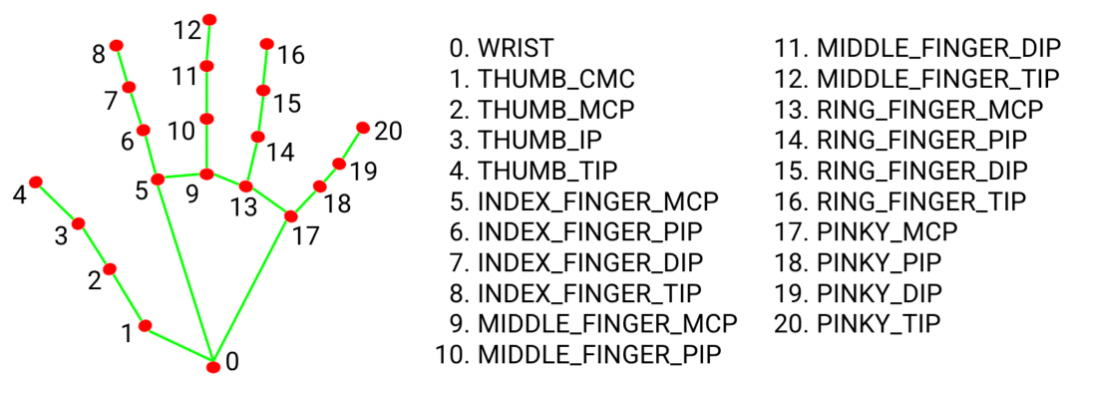


Fig.6 21个首部标记点示意图

效果图如Fig.7所示

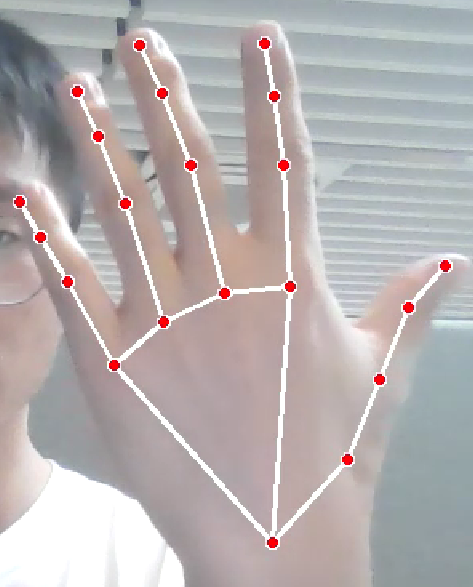


Fig.7 效果图

在得到21个三维手部关键点的坐标后，便可以根据这些坐标在捕获的图像或视频流中截取出手部图像之后进行手势识别。

4.2实时测试

通过 python 收集我们的摄像头的镜头采集的图像并进行实时预测。具体过程是解析捕获的每一帧图像并提取出手部图像，将其处理为灰度图，并用之前训练好的模型进行检测。为了更轻松地对预估结果查看，预测的字母将会显示在实时画面上以下是两个案例的效果图(Fig.8)。

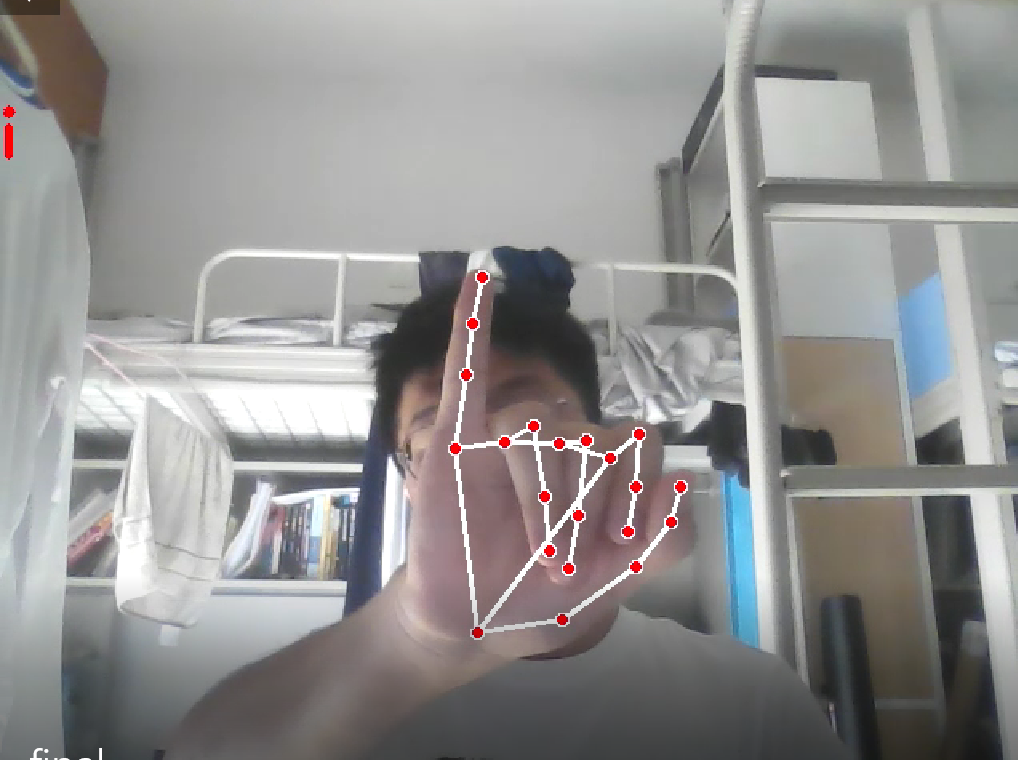
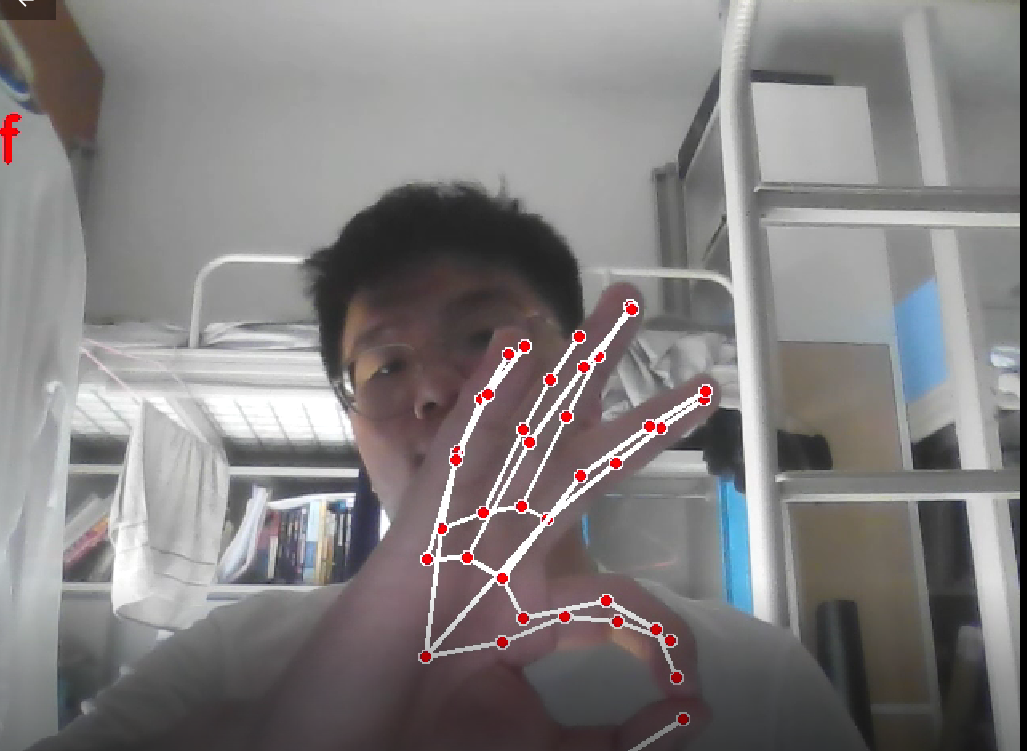
 

Fig.8 i和f的示例图

1. 项目总结

即使模型在测试集上的准确率很高，但在实际运用中效果还是比较差。原因主要有以下几点：

* 摄像头拍摄的角度。摄像头拍摄的角度会造成实际捕获的图像与数据集之间存在差异，影响模型实际运行效果。
* 在实际环境中，实际捕获的手部图像背景与数据集图像背景之间也存在较大的差异，且在实际情况中背景不能够控制。
* 模型在数据集上训练的效果很好，不代表在实际情况中效果也会很好。实际情况与训练测试情况有不可避免的偏差。

参考文献

[1] 韩信子．深度学习教程 | 吴恩达专项课程 · 全套笔记解读．（2022-04-12）．[2024-01-11]．https://showmeai.tech/tutorials/35．

[2] 思绪无限．基于深度学习的手势识别系统．（2023-06-27）．[2024-01-11]．https://blog.csdn.net/qq\_32892383/article/details/124155373．

[3] 开鑫9575．手部21个关键点检测+手势识别．（2021-04-05）．[2024-01-11]．https://blog.csdn.net/weixin\_45930948/article/details/115444916.