# **基于MediaPipe的实时手势识别与非接触式人机交互系统**

林之博

西南大学 计算机与信息科学学院 软件学院

中国重庆市北碚区天生路2号

[LeoLin.5@outlook.c](mailto:firstauthor@i1.org)om

**Abstract（摘要）**

*在这个项目中，我将利用普通的电脑摄像头，搭建一个能通过手势实时控制电脑的系统。平时我们用鼠标键盘或者触摸屏，都需要物理接触，不仅在一些公共场合不卫生，有时也不够方便。我的想法是，通过计算机视觉技术来解决这个问题。具体来说，我会使用Google的MediaPipe框架来捕捉和跟踪手部的关键点，然后在此基础上，编写我自己的一套逻辑算法来识别不同的手势，比如“握拳”是静音，“张开五指”是取消静音，以及通过两个手指的开合距离来平滑地调节音量。最终的目标是做出来一个能用的原型，验证这个想法的可行性。*

# **Introduction（概述）**

## **Problem statement （问题描述）**

让电脑通过摄像头“看懂”我们的动作是一件很酷的事情。我对这种人机交互方式很感兴趣，它让我们可以用更自然的方式去控制机器。特别是现在，非接触式操作因为卫生和便利性等原因，应用前景越来越广。课程项目要求把学到的知识应用到感兴趣的问题上，这就是一个很好的机会，我的项目就是这个方向的一个具体应用实践。我不打算从零去训练一个复杂的视觉算法或模型，而是想利用现有的技术（Google MediaPipe)，利用它成熟、高效的手部跟踪能力，专注于如何设计出手势指令，并将其与实际的电脑控制功能（如音量调节）结合起来，做一个小而实用的应用，从而解决实际问题 。

## **Expected results（预期结果）**

这个项目主要想解决的问题是：我们能否只用一个普通摄像头，就实现一套流畅、实时的手势控制系统？

我觉得这个问题很有研究价值，因为：

* 成本低、门槛低：不需要任何昂贵的特殊设备，人人都有的摄像头就能用。
* 交互直观：用手势来控制音量大小、播放暂停，比找鼠标点击要自然得多。
* 应用场景广：未来可以用在智能家居、公共展示屏、无菌手术室，或者帮助一些行动不便的人士。

# **Related work（相关工作）**

在做这个项目之前，我主要参考了Google MediaPipe的官方论文[1]与MediaPipe 解决方案指南[2]。

我的项目和MediaPipe的关系是“利用”而不是“竞争”。

相同点是，我们都处理手部的21个关键点坐标。

不同点是，MediaPipe的工作是提供这些坐标，它是一个底层的“数据提供者”。而我的工作，是在拿到这些数据后，进行“二次加工”，设计出上层的“手势语言”和“控制逻辑”。可以说，MediaPipe解决了“看”的问题，而我专注于解决“理解”和“执行”的问题。

# **Method（方法）**

我的代码会像一个流水线一样工作：

· 用OpenCV捕捉画面：让程序能看到摄像头拍到的东西。

· 把画面喂给MediaPipe：我会直接调用mediapipe.solutions.hands这个现成的工具。它会帮我完成最难的工作——在画面里找到手，并返回21个关键点的坐标。

· 自己写算法判断手势：这是我的核心工作。我会根据那些坐标点的位置关系，写出判断规则。

比如判断“握拳”，我的算法会去检查四个手指的指尖是不是都收到了手心里。

比如调节音量，我的算法会去计算大拇指和食指指尖的距离，距离远音量就大，距离近音量就小。为了防止控制时音量条乱跳，我还会对这个距离值做一个“滑动平均”的平滑处理。

· 把手势“翻译”成电脑指令：最后一步，我会用pycaw这样的库，把识别出的手势，比如“音量80”，变成真正的系统音量调节命令。

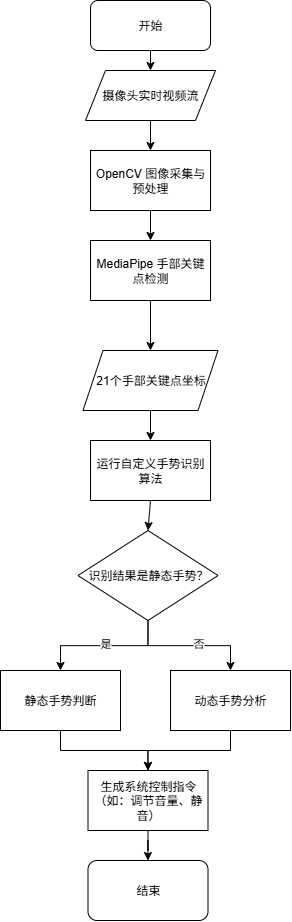


Figure 1: 项目整体流程图

# **Experiments（实验）**

## **Data （数据）**

本项目不使用任何预先录制好的、公开的数据集。系统处理的数据是通过标准电脑摄像头（笔记本内置或USB摄像头）实时采集的视频流。这种方式能更好地评估算法在真实、动态环境下的性能和实时性。

在项目的定量评价阶段，我需要采集实验数据来计算准确率。我的采集计划如下：

采集环境：在一个光照条件稳定（如室内灯光）、背景相对简洁的环境中进行。

采集设备：使用固定的笔记本电脑摄像头或USB摄像头，确保采集过程中摄像头位置和角度不变。

采集规程：我会预先定义好一组（例如5个）需要测试的手势。对于每一个手势，我会重复执行30次。每一次执行，系统会记录下它的识别结果（是识别正确，还是识别错误，或是识别成了哪种其他手势）。这30次的结果就构成了对单个手势进行评价的“数据集”。

## **Evaluation （评价）**

我将结合定性与定量两种方式来评价和展示我的结果，以全面地证明我解决了“通过手势实时控制电脑”这一问题。

Qualitative results (定性结果):

定性评价的目标是直观地展示“我的系统能用，而且用起来效果如何”。具体展示方式包括：

实验效果图 (Experimental Effect Diagrams)：这是主要的定性展示方式。我会提交程序的运行截图和录制的演示视频。截图中会清晰标注出手、识别出的骨骼点、以及程序判断出的手势结果。视频中则会完整地演示我做出一个手势（如两指开合），以及电脑系统做出相应反应（如系统音量条实时变化）的全过程，以此证明系统的功能完整性和流畅度。

曲线图 (Curve/Line Graphs)：对于“控制音量”这类动态手势，我可以绘制一条曲线图。图的X轴是时间，Y轴是“两指指尖的距离”和“系统音量”这两个值。通过这条曲线可以直观地分析我的算法映射是否平滑、响应是否及时。

Quantitative results (定量结果):

定量评价的目标是用客观的数据来证明“我的系统有多准”。

1.评价指标 (Evaluation Metrics)：

识别准确率 (Recognition Accuracy)：这是最核心的评价指标。对于每一个测试的手势，其准确率的计算公式为：

2.分析方法 (Analysis Methods)：

我会使用在6.1节中采集到的数据。对每一个手势的30次测试结果进行统计，计算出各自的识别准确率。

最终，我会将所有手势的准确率结果汇总在一张表格中，并使用柱状图 (Bar Chart) 进行可视化，方便直观地比较不同手势的识别效果。同时，我也会展示混淆矩阵来分析具体的错误分布情况。

# **Plan (进度安排)**

本项目的开发将围绕课程设定的三个进度报告周期进行，具体目标如下：

项目进度报告1 (截止日期：~10月7日)

完成开发环境的搭建，配置好OpenCV、MediaPipe等所需库。

实现基础的手部关键点实时检测，并在摄像头画面中成功可视化。

将核心检测功能封装为独立的handDetector模块，完成项目基础架构。

项目进度报告2 (截止日期：~10月28日)

完善handDetector模块，实现对关键点坐标的稳定提取。

设计并实现至少3种静态手势的识别算法（如：握拳、五指张开）。

完成对已实现手势的初步定量测试，计算并用图表展示识别准确率。

项目进度报告3 (截止日期：~11月18日)

实现动态手势识别算法（如：通过两指距离控制音量），并与系统功能绑定。

为提升用户体验，对动态控制逻辑加入平滑滤波处理。

完成计划中的其他部分，并录制演示视频。

在第三份进度报告提交后，剩余时间将全部用于汇总所有实验数据、撰写最终的结题报告、完善代码并上传至GitHub，以及准备最终的项目答辩。

# **Miscellaneous (其他说明)**

本项目是为《机器视觉与行业大数据应用》课程独立完成的，与我参与的其他课程或研究没有内容上的重叠 。

# **References**

1. C. Lugaresi, J. Tang, H. Nash, et al. MediaPipe: A Framework for Building Perception Pipelines. arXiv preprint arXiv:1906.08172, 2019.
2. Google. MediaPipe 解决方案指南. Google AI for Developers.[Online].Available:<https://ai.google.dev/edge/medipe/solutions/guide?hl=zh-cn>
3. G. R. Bradski. The OpenCV Library. Dr. Dobb's Journal of Software Tools, 2000.