基于多特征融合的手语数据高质量筛选系统

摘要

随着时代进步，人们的生活质量逐步提高，但国内聋人的生活质量尚存很大的提升空间。手语翻译研究课题在国内近年越来越流行，但国内的手语数据集规模仍然是一个瓶颈问题，为了扩大国内手语数据集的规模，我校天津理工大学聋人工学院计划并施行收集100人录制2000个不同句子的大规模手语数据集。在收集手语视频的过程当中，为了保证数据的质量，需要对所有的手语视频做从头到尾的检测，这将要花费不少的人力和财力。基于多特征融合的手语数据高质量筛选系统，只需要检测人员在可视化软件上执行一些简单的操作，就可以完成大规模的手语视频数据质量检测。

关键词：中国手语 数据集 评估算法

目录

[第1章 绪论 3](#_Toc18087)

[1.1课题的意义 3](#_Toc30026)

[1.2国内外研究现状 3](#_Toc23324)

[1.3论文结构 3](#_Toc8164)

[第2章 手语语言学 5](#_Toc19592)

[2.1手语的发展历史 5](#_Toc31892)

[2.2手语动词 5](#_Toc5556)

[2.3手语语法 6](#_Toc11511)

[第3章 系统设计 1](#_Toc25594)

[3.1系统的使用场景 1](#_Toc6782)

[3.2系统的设计 1](#_Toc11900)

[3.2.1 Proference配置 1](#_Toc11604)

[3.2.2 Check Data 3](#_Toc7171)

[3.2.2 导出KeyPoints关键点 3](#_Toc31556)

[3.3.2 DataMo 4](#_Toc21706)

[第4章 系统实现 6](#_Toc14023)

[4.1 视频处理 6](#_Toc1998)

[4.1.1 MPEG-4标准 6](#_Toc10162)

[4.1.2 视频解码器以及性能对比 6](#_Toc26446)

[4.2.3视频方向错误的解决方案 7](#_Toc531)

[4.2 骨架提取 9](#_Toc16035)

[4.2.1 骨架提取原理 9](#_Toc21828)

[4.2.2 Mediapipe 9](#_Toc5026)

[4.2.2.1 MediaPipe 主要概念 9](#_Toc23607)

[4.2.2.2 MediaPipe Holistic 10](#_Toc6167)

[4.2.2.4 从原始关键点到Pose类 （改mediapipe的输出介绍，以及对应的身体部位) 11](#_Toc396)

[4.3 特征预处理 12](#_Toc5154)

[4.3.1脏数据产生的原因 12](#_Toc15600)

[4.3.2 MediaPipe可信度 13](#_Toc30848)

[4.3.3偏top方向的角度 13](#_Toc23764)

[4.4 特征重构 16](#_Toc32766)

[4.4.1特征的意义 16](#_Toc9855)

[4.4.2基于不同组合的欧式距离特征 16](#_Toc16812)

[4.4.5插值与调整序列长度 17](#_Toc18437)

[4.5筛选系统的评估方法 18](#_Toc24974)

[4.5.1筛选系统的实际应用场景问题 18](#_Toc12972)

[4.5.2多层感知机(Multilayer Perceptron) 18](#_Toc15315)

[4.5.2.1 多层感知机介绍 18](#_Toc27079)

[4.5.2.2 使用多层感知机 18](#_Toc1765)

[5.5.3 Model Chain 模型链 20](#_Toc28939)

[5.5.3.1 Model Chain模型链的定义 20](#_Toc2761)

[5.5.3.2 Hand数据增强 20](#_Toc7914)

[5.5.3.3 Model Chain 结构 25](#_Toc16552)

[5.5.3.4 Model Chain 评估流程 26](#_Toc31358)

[4.6 筛选系统的界面端 27](#_Toc888)

[4.6.1 Qt框架介绍 27](#_Toc801)

[4.6.2筛选系统界面端的不同组件 27](#_Toc24059)

[4.6.2.1表格目录和视频目录 27](#_Toc30094)

[4.6.2.2参数设置 28](#_Toc18012)

[4.6.2.3 日志系统 28](#_Toc7567)

[4.6.3筛选系统实现细节问题 28](#_Toc19088)

[4.6.3.1 多线程 28](#_Toc15731)

[4.6.3.2 界面端和计算部分的通信 29](#_Toc519)

[第5章 系统测试 30](#_Toc31680)

[5.1系统使用测试 30](#_Toc6820)

[5.2系统评估测试 30](#_Toc12926)

[总结 31](#_Toc849)

[参考文献 32](#_Toc9698)

[致谢 33](#_Toc9343)

# 第1章 绪论

## 1.1课题的意义

随着时代进步，人们的生活质量逐步提高，但国内聋人的生活质量尚存很大的提升空间。手语翻译研究课题在国内近年越来越流行，但国内的手语数据集规模仍然是一个瓶颈问题，为了扩大国内手语数据集的规模，我校天津理工大学聋人工学院计划并施行收集100人录制2000个不同句子的大规模手语数据集[1]。在收集手语视频的过程当中，为了保证数据的质量，需要对所有的手语视频做从头到尾的检测，这将要花费不少的人力和财力。基于多特征融合的手语数据高质量筛选系统，只需要检测人员在可视化软件上执行一些简单的操作，就可以完成大规模的手语视频数据质量检测。

## 1.2国内外研究现状

当在以前的手语识别任务中，以孤立词识别为主。这些年来，研究人员在此任务中引入了许多方法，例如传统的方法，如隐马可夫模型方法，HOG方法,并在5113个大词汇库中取得82.3%准确率。在深度学习的方法上，卷积神经网络CNN等方法的准确率突破94%[2]。

连续手语的翻译任务，在国内外已经存在很多方法，大部分的思路是用分割的方法将连续的手语时序上分成许多短部分，再利用孤立词的识别框架识别出来，并用自然语言处理框架将这些特征转换为文本语言[4-5]。

在这些相关手语研究当中，尚未出现主要对于手语数据的质量检测的研究课题。

目前来看，手语相关的研究在计算机科学领域当中越来越有不少人去关注，包括数字人手语（文字转译手语），手语转译文字等等。

## 1.3论文结构

为了让系统能够对手语视频数据集做出正确的评估，对手语语言学的了解是有必要的。本章将介绍中国手语的发展历史，以及介绍手语动词，在第2章具体介绍。

该系统为了配合天津理工大学聋人工学院的手语数据集采集大任务，考虑到一些应用场景的问题，本章给出手语数据筛选系统的设计思路以及系统的工作步骤。手语数据筛选系统需要处理的数据是重量级的，因为它要处理的一个数据单位是一个视频，再加上后面的算法学习评估，需要的处理时间会更冗长，系统的性能也是需要考虑的一点，了解视频解码器的使用和原理都是有必要的。第3章介绍了系统的使用场景以及系统的一些设计。第4章介绍筛选系统的实现，包括对视频提取的特征的步骤，数据的预处理，特征的重构，特征的的学习，系统的评估方法，以及系统的界面设计。第5章介绍系统的可靠性测试。最后给出总结和参考文献。

国内外研究现状，目前尚未发现对于手语数据集的评估算法的研究成果，这是毕业论文的一个挑战点。为了完成该任务，可以根据一些现成的骨架提取框架的提取结果通过一些合理的处理，配合分类学习的方法实现数据筛选系统。

# 第2章 手语语言学

对于聋人来说，手语是重要的交流工具，以模仿事物的特征和模仿动作为主，表情为辅。其中的表情包括身体的姿态，脸面表情，动作的夸张程度。在广义上，中国手语是指中国聋人使用的手语，研究中国手语发展史，了解聋人手语的特点，有助于手语数据高质量筛选系统的设计。

## 2.1手语的发展历史

从原始生活角度来看，人类祖先在有声语言形成之前经历过一段使用手势进行互相交流。从人类生长角度来看，婴幼儿能够自觉用拇指指出一些事物，表达自我的意愿。可知，对于人类来说，手势是一种表现比较直接的语言，由于有声语言越来越成熟，人类对手势语言的依赖性也随之越来越少。而对与失去听力能力的人群来说，手势语言的地位骤然提升，成为了这些人群的主要沟通语言，甚至是唯一的语言，就是现在所谓的手语。如今，手语的文化成熟程度并非昔日可比。

翻开中国历史资料，中国手语的早期形式出现于公元前100年前的《史记》，“骐骥之跼躅，不如驽马之安步也……虽有舜禹之智吟而不言，不如暗聋之指麾也。”，“暗聋”是指聋人，“指麾”是指聋人的手势比划。可见当时聋人手势亦可作为一种语言。在古代，中国聋人的手势语言多为自然手语。自然手语的特点是大多数手势是以模仿事物的特征形象为主，如:“身体倾斜，将手比作枕头，头落在手上”表示为“睡”，“指嘴、摆摆手”为“哑”。亦有“画字”的手语形式，聋人与他人沟通时可“以手画掌、画地为字、画纸与人语”。中国现代手语的“书空，仿字”，是从这种“画字”形式继承而发展来的。

从近代及现代来看，中国手语发展成两种形式，一是自然手语形式，多为社会上的聋人使用，多数手势类似于古文中的字词，字词简短，表达的意思丰富。手势表现形象，易懂。二是手势汉语，多为学生使用，以手势一对一表示字词，除表现形式之外，语义与汉语似无差别，却非易懂。大多数聋人学生走进大学或走进社会之时，在沟通时，也会改以自然手语为主。

## 2.2手语动词

名词和动词是任何语言的主要成分，在手势语言方面上来看，动词是极其重要的，动以表意，有些名词也以做动作模仿来表示，如“狮子”，以双手在头上乱舞，描述了狮子的头发特征，这种动作也可表示形容词“疯”，也可以表示名词“狮子”。

## 2.3手语语法

语法是语言中的规律，在授予语言学方面上，一些研究者做了一定的研究，对手语的造字词的分析可以借鉴“六书”理论（汉字造字方法），但是不可以覆盖全面的手语语法规律性。手语不同于其它语言形式，因为手语是一种视觉化语言，非同于有声语言序列先后顺序的束缚性，有许多事物和动词可以同时用身体的各个部位以不同的组合方式去表达，有一些研究者对手语做出一些语法的总结，普遍聋人认为这只是手语语法的一部分，到目前来看，对于中国手语语法尚未做出全面的总结。

# 第3章 系统设计

为了保证筛选系统在根本的问题上去解决需求，需要对系统的使用场景做出一些分析，包括在使用场景中将会碰到的一些问题。在此基础上做出工作步骤的总结，以做好系统的设计。

## 3.1系统的使用场景

视频数据只有句子分类的标注，以及对应的转译文本，没有视频质量好坏的标注。

系统运行在服务器上，系统使用者没有技术基础。系统筛选的视频数据需要从云服务器下载，并没有视频数据下载的接口提供。且下载的视频数据可能是有限的，并且有对应的详细信息表。

使用者在使用系统时，能够得到系统对视频的评估结果反馈，且可以观察每个视频的评估参数。最终系统可以输出评估筛选结果表，用户可以经过纠正结果后再上传到一个录制的系统，录制的系统会根据筛选结果表驳回不合格的视频，让用户重新录制手语视频。最后给所有的视频做出评估，并且导出评估结果表（Evaluate result table）和视频筛选结果表(Video filter result table)。评估结果表为模型链和分类模型的评估结果，用于用户观察每个视频的质量。视频筛选结果表用于上传手语收集系统驳回不合格的视频。

## 3.2系统的设计

### 3.2.1 Proference配置

考虑到系统的录入数据可能会超出考虑的范围，由此特定了一些系统的评估前置条件。由此给出一些用户参数，如果合乎用户配置的特定条件，系统则可以开始运行程序。参考图3.1，参数设置有

* 最小类别数：识别到的视频类别数需要大于该值。
* 最小类别样本数：每个类别录入的数据的样本数需要大于该值。
* Pose 最小可置信度：可置信度越低，骨架评估工具识别比率越高，质量反而会低。如果提取的Pose关键点质量低于可置信度，MediaPipe则会返回空值（参考4.3.2）。
* Hand最小可置信度：可置信度越低，骨架评估工具识别比率越高，质量反而会低。如果提取的Hand关键点质量低于可置信度，MediaPipe则会返回空值（参考4.3.2）。
* 最大小臂偏北角度：如果识别到的角度低于该角度值，则被录入特征，否则不录入特征（参考4.3.3）。
* 模型特征长度：利用插值的办法使特征的长度统一，如果视频有106帧，特征的长度将会是该值，而不是106（参考4.4.5）。
* Left Hand 最小识别比率：可识别的左手样本数 / 最大小臂偏北角度过滤后的样本数。
* Right Hand 最小识别比率：可识别的右手样本数 / 最大小臂偏北角度过滤后的样本数。
* 学习样本比例：学习样本数 / 总样本数。
* 评估次数：参数设置为N,对于不同的特征，系统均搭建N个学习模型，将会做N次评估（参考4.5.4）。

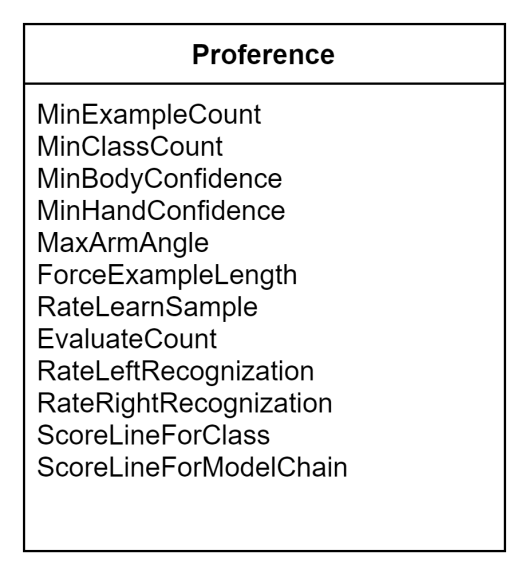


图3.1 Proference

### 3.2.2 Check Data

系统的初始状态为WAIT，设定好视频的根目录和视频详细信息表格之后，系统会自动进行CHECK DATA，检查视频和表格的数据是否吻合状态变为可以允许RUN PROGRAM，如果不吻合，需要打印出不同的地方，以列表展示。如下：

1. 视频（视频根目录下的数据未在表格里发现）

[2022-05-12 06:30:42] : 有3个视频未出现在表格里，以下列出不符合条件的视频：

[2022-05-12 06:30:42] : cj-j-1nOCFNB3-DBmpdd56kG-26194 .mp4

[2022-05-12 06:30:42] : cj-j-2dfvwGddUQ-1KvelR1K-22790 .mp4

[2022-05-12 06:30:42] : cj-j-2dfvwGUQ-2mqWddd3KO-7573.mp4

1. 表格（表格里的数据未在Videos根目录下未发现）

[2022-05-12 06:26:05] : 有3个数据在视频里未找到，以下列出不符合条件的USER\_ID-SUBMIT\_ID:

[2022-05-12 06:26:05] : 2dfvwGUQ-1KvelR1K

[2022-05-12 06:26:05] : 2dfvwGUQ-2mqWd3KO

[2022-05-12 06:26:05] : 1nOCFNB3-DBmp56kG

接下来检查类别个数（CLASS COUNT）是否满足条件，样本个数（Example count per class）是否满足条件。如果这些检查过程（Check）一切正常时，系统将会允许用户导出KeyPoints。

[2022-05-12 06:30:42] : 表格里的数据均符合条件。

[2022-05-12 06:30:42] : class num = 10,大于等于10，满足条件。

[2022-05-12 06:30:42] : 所有分类样本数均满足条件。

### 3.2.2 导出KeyPoints关键点

用户点击“导出关键点”将会开始对每个视频Extract keypoints，并且保存为名为“xxx.keypoints”pickle文件。一些就绪时，系统才会允许用户开启评估任务。

### 3.3.2 DataMo

DataMo封装了每个视频相关的信息，包括Name, ClassName, VideoPath, Features, KeyPoints等等。参考图3.2：

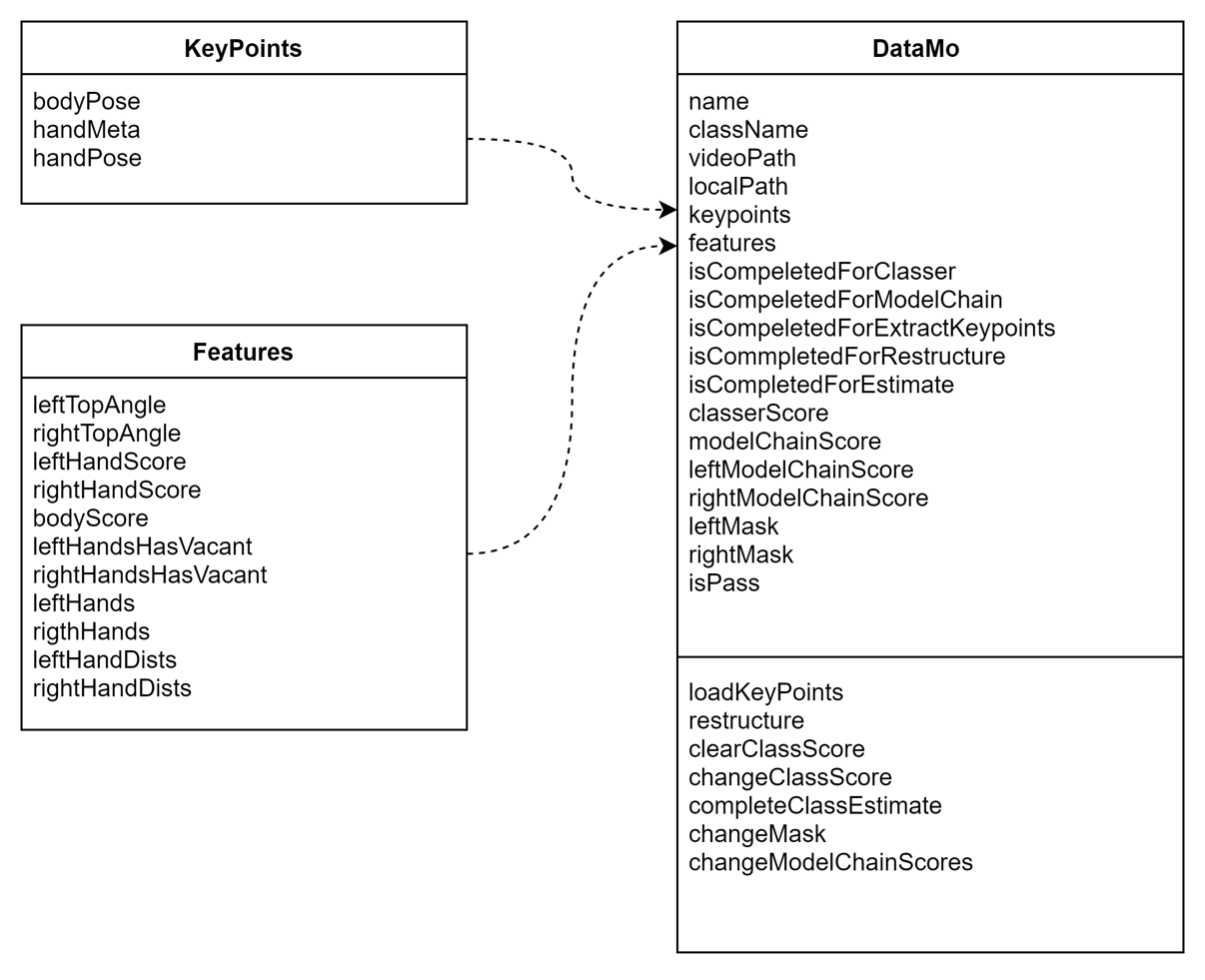


图3.2 DataMo类结构

系统将设计一个DataMo以穿过整个系统的生命周期，将会在输出KeyPoints导出之后被创建。

系统提取KeyPoints之后会更新每一个DataMo实例的keyPoints属性，并且自动本地序列化，以避免以后重新提取KeyPoints。

提取KeyPoints一切就绪时，系统开始重构特征计算，完成计算之后，系统也会更新每一个DataMo实例的features属性，并完成序列化操作。一切就绪之后，系统开始搭建模型(Create models)。系统将开启执行两个子任务，第一个任务为训练分类多层感知机模型以对应整体的视频评估，第二个任务为训练模型链以对每个视频的细节评估。

完成Class分类方法评估之后或者完成模型链方法评估之后，分数信息，Mask信息都会更新到每一个DataMo实例。

# 第4章 系统实现

## 4.1 视频处理

### 4.1.1 MPEG-4标准

MPEG-4 是我们常用的视频格式，MPEG-4标准同之前的标准差距在于MPEG-4采用基于对象的编码理念，使若干个音视频对象在时间上和空间上有相互的联系，这些对象能够表示一幅景物，在经过编码处理之后通过可复用传输到接收端，再对不同的对象解码，即可以合成编码前的音视频。

MPEG-4 基于更高的编码效率。 与其他现有或即将推出的标准相比，它基于相同比特率下更高的视觉和听觉质量，这使得在低带宽通道上传输视频和音频成为可能。 同时，MPEG-4 还可以对同时出现的数据流进行编码。

### 4.1.2 视频解码器以及性能对比

视频解码器是一种程序（视频播放器）或设备，用于恢复和解码编码的数字视频。 我们的常见的视频格式大多数是MP4，MP4视频是由MP4特定的格式组成的视频，常见的播放器都会内置很多解码器来解析各种格式的视频。如果没有解码器，系统将无法获取读取特定格式的视频。

对于Python语言来说，MP4视频主流的视频解码器有Open CV, FFmpeg, decord。通过性能对比的实验，参考表4.1。

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Decord**  **Shape(337L,224L,224L,3L)** | **OpenCV**  **Shape(337L,224L,224L,3L)** | **FFmpeg**  **Shape(337L,224L,224L,3L)** |
| **Duration** | 0.42s | 0.93s | 1.3 s |

表4.1 视频解码器的性能对比

可知在同样的视频解码下Decord解码器解码速度最快，比Open CV解码器快Decord解码一个视频过程当中，会一次性的完整读到内存上，解码的缺陷明显在于内存占用过大。考虑到系统上的性能需求不是很大，且程序的内部运行结构似于流水线（PipeLine Structure），在运行的规范和性能上的权衡上，最终采用Open CV解码器。

在系统的实现过程中，使用Open CV作为一个视频的处理库，在解码过程中。Cv2.imencode() 作为一个解码的函数使用，可以使用任何图片格式来解析帧，比如BMP，JPEG在使用过程中发现使用不同的解码方法在性能方面上有明显的差距。参考表4.2，使用JEPG格式解码要比BMP解码慢近于5倍，大小比BMP解码格式的大小204.8倍，可知JPEG格式采用压缩算法导致视频解码性能降低。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Format** | **Resolution** | **Size** | **Duration** |  |
| **BMP** | 800\*600 | 2MB | 0.7 ms |  |
| **JPEG** | 800\*600 | 10KB | 4 ms | 500% slower |

表4.2 不同解码格式的性能对比

#### 4.2.3视频方向错误的解决方案

在录制手语视频的时候，项目参与采录人员会把手机放在面前的一个地方或者让另外一个人来帮忙采录视频。由于不同的手机配置，在一个系统内可能无法正确识别正确的方向。比如说视频的方向可能倒置过来是常见的情况。

为了避免该情况出现，因此设计一个方法，该方法的思路是识别一个视频的5个随机的帧。使用Mediapipe识别这些帧的特征，使用Body的关键点，以两肩的中心点为准，以 l,r 分别标记。有

K = |Y(l,r)| / X(l,r)

|Y(l,r)| 表示l和r的y方向距离

X(l,r)表示 r的x坐标值减l的x坐标值

如果K 接近于0，那么说明这个视频没有被逆时针旋转90°或顺时针旋转90°。

有头的中心点为h，有

K2 = Y(h,l)

Y(h,l)表示h的y坐标值减l的y坐标值

头的中心点h的y坐标值应大于两肩的其中一个点的y坐标值，可知，如果K2大于0，则表示视频没有被倒置。

## 4.2 骨架提取

### 4.2.1 骨架提取原理

人体关键点姿态识别是人体姿态识别技术的重要分支。通过检测人体行为表达过程中，每一帧人体姿态关键部位的位置，将人体姿态简化为人体关键点，并通过这些关键点对人体姿态表达的语义进行分类识别。

人体姿态估计可以分为两种思路：

（1）“top-down”，它指先检测人体区域，再检测区域内的人体关键点。

（2）“bottom-up”，它指先检测图片中所有的人体关键点，然后将这些关键点对应到不同的人物个体。

### 4.2.2 Mediapipe

Mediapipe 整体由具有优化姿势和组件的新管道组成。 每个组件都实时运行。 它的推理后端之间的内存传输最少，增加了这三个组件的互换性支持，这取决于质量/速度的权衡。

#### 4.2.2.1 MediaPipe 主要概念

MediaPipe的核心框架由C++实现，提供对Java、Objective C等语言的支持。MediaPipe的主要概念包括包、流、计算器、图和子图。 数据包是最基本的数据单元。 一个数据包代表特定时间节点的数据，例如一帧图像或一小

段音频信号； 数据流由多个按时间升序排列的数据包组成。 一个数据流的特定时间戳只允许最多存在一个数据包； 数据流在由多个计算单元组成的图中流动。 MediaPipe的图形是定向的——数据包从数据源（源计算Source Calculator或者 图输入流Graph Input Stream）流入图形，直到它们离开汇节点（汇计算器Sink Calculator或图形输出流Graph Output Stream）。

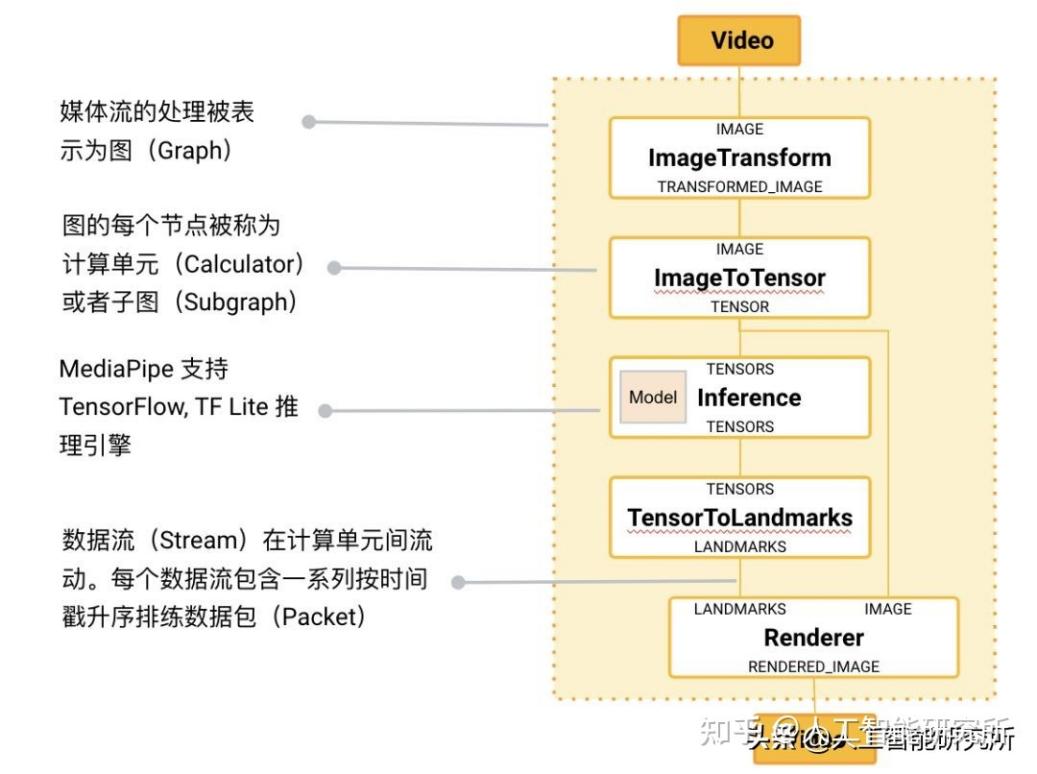


图4.1 mediapipe 结构

#### 4.2.2.2 MediaPipe Holistic

管道和质量媒体管道整体管道集成了姿势、面部和手部组件的单独模型，每个模型都针对其特定领域进行了优化。 但是，由于它们的特殊性不同，一个组件的输入不适用于其他组件。 例如，姿态估计模型使用较低的固定分辨率视频帧 (256x256) 作为输入。 但是，如果要从图像中裁剪手部和面部区域以将其转移到各自的模型中，则图像分辨率将太低而无法执行准确的关节运动。 因此，我们将mediapipe holographic设计为多级流水线，使用适合区域的图像分辨率来处理不同的区域。

#### 4.2.2.3 mediapipe输出结构

系统使用mediapipe提取Keypoints保存为”xxx.keypoints”文件，xxx为视频名，包括了三个主要的信息：body keypoint、Hand meta 和Hand keypoint。

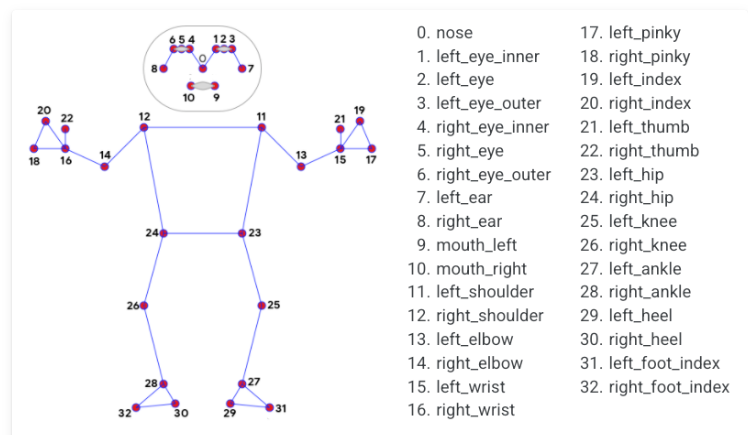


图4.2 body keypoint结构

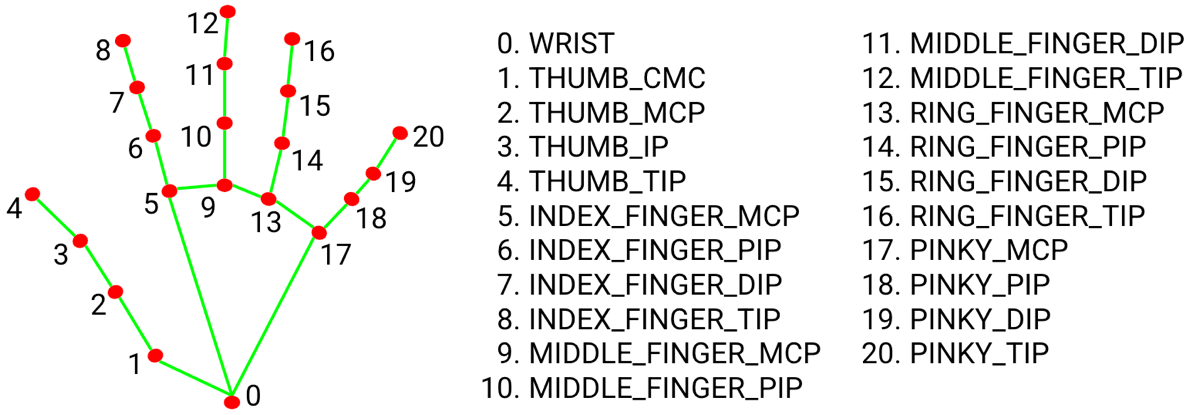


图4.3 hand keypoint结构

## 4.3 特征预处理

在学习模型中，特征的质量好坏将会对模型的准确率和鲁棒性产生不可忽视的影响。预处理是对特征处理的第一个重要的步骤。

### 4.3.1脏数据产生的原因

在评估身体姿态和手势关键点的过程中发现提取的特征质量低下，在实验的过程中，总结出以下几点原因：

1. 模糊性

模糊性不只是在图像上的质量程度，在手势者从表达一个关键手势到另一个手势，会有快速的移动和变化，由此会产生动作模糊不清楚导致的脏数据，关键点提取的精度将大大折扣。

1. 背景颜色接近

手势的附近背景颜色接近于肉色，会导致提取关键点的精确度会下降。

1. 手语静止状态

在手势者的表达过程中，或者在开始打手语之前和结束打手语后会把手臂放下来。这些的手势并没有任何的意义，如果学习模型的训练数据包含这些手势的话，可能会影响学习模型的训练效果。

### 4.3.2 MediaPipe可信度

系统使用的是MediaPipe提取视频的骨架关键点。MediaPipe是一个功能强大的工具，它提供了一个可信度配置的功能，对于一个视频的每一个帧，MediaPipe识别的过程当中，如果可信度配置低于用户配置的可信度，返回的识别结果会为空。对于上述的第一、二个问题，这些原因下产生的识别结果可信度必然会很低，使用这点的功能很好的解决这两个问题。

### 4.3.3偏top方向的角度

对于第三个问题，在实验观察中，从小臂的角度来看更容易判断手语者是否正在打手势。以角度来分析，选择一个角度为零角度，因为手势者打手势的过程中，会举起手势，且大多数的时候，手是向上的，或者向左向右。因此选择top方向作为参考角度，参考角度可以视为零角度。

x,y分别表示手腕和肘关节的位置，x到y的方向偏top方向的角度为s，如果s大于180°，有

angle = s - 180

如果s 小于180°时，有

angle = s

无论小臂如何朝左朝右，其angle值会接近于90°，小臂朝上时，angle值接近于0°，当停止打手势的时后，小臂朝下，angle值接近于180°。

见图2，可知大概从第50多帧真正的手势才开始表演。在处理数据帧，我们将过滤掉s>150的帧。

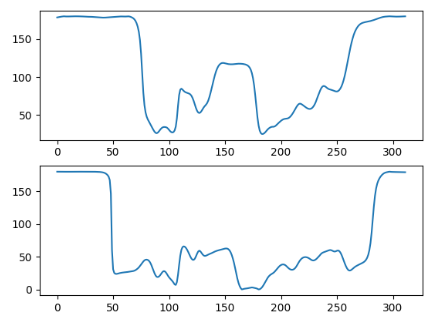


图6.1 偏top方向的角度图

上图表为左小臂的angle值的变化图，下图表为右小臂的angle值的变化图。

以下给出一个视频的特征经过预处理前后的对比：

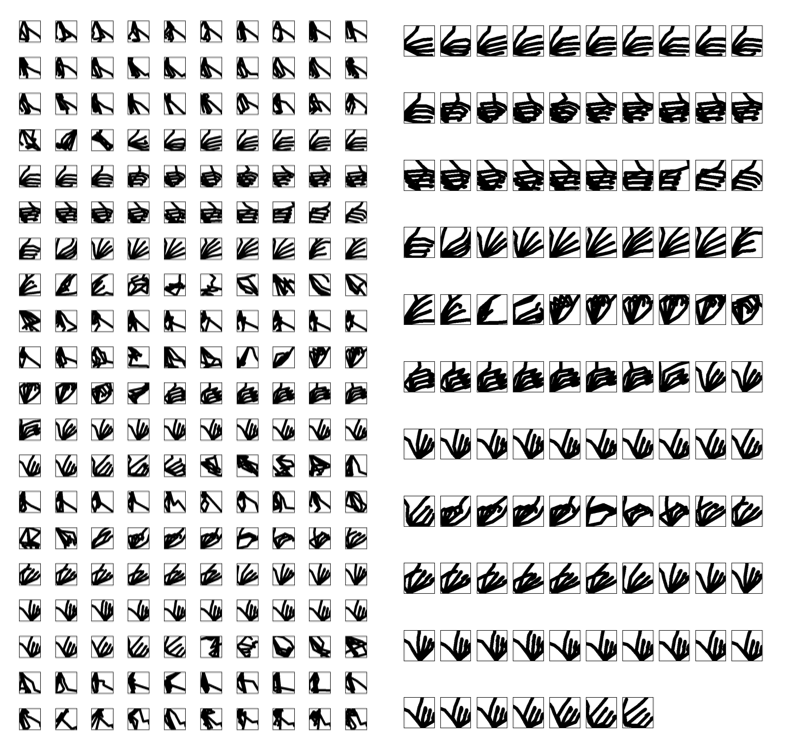


图 6.2预处理前后特征对比

图6.2左侧的是预处理前的图，开头或结尾部分的手语没有在右侧出现，因为这些是在人手放下的状态的识别特征。右侧的手势状态均为清晰。

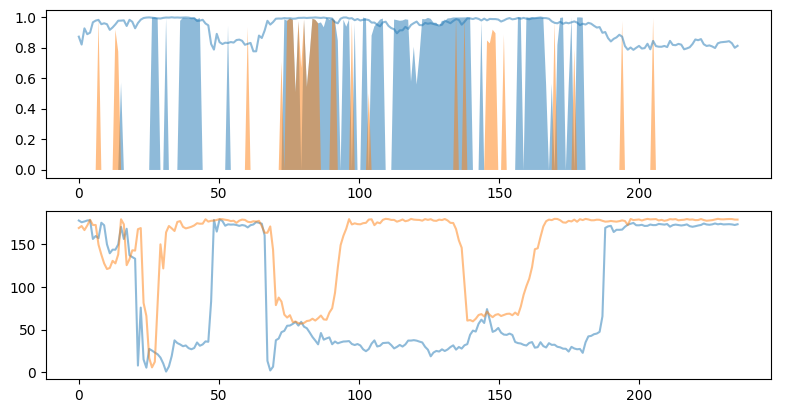


图6.3 mediapipe输出质量和Top角度对比图

参考图6.3，上子图的蓝线表示body的上半身的visibility，蓝色的阴影表示左手的scose，实际上为手语对象的右手score，黄色的阴影表示为右手的score，实际上为手语对象的左手score。下子图是对应的Top Angel值。通过观察，top angel值越低时，对应的score越高。

## 4.4 特征重构

### 4.4.1特征的意义

在机器学习中，特征学习或表示学习是一组技术，允许系统从原始数据中自动发现特征检测或分类所需的表示。这取代了手动特征工程，并允许机器学习特征并使用它们来执行特定任务。

特征学习的动机是机器学习任务（例如分类）通常需要在数学和计算上便于处理的输入。然而，图像、视频和传感器数据等现实世界的数据并没有屈服于通过算法定义特定特征的尝试。另一种方法是通过检查发现这些特征或表示，而不依赖于显式算法。

从中国人聋人习性的角度来看，手型的方向和手的形式决定了某个特殊的含义。对于某个关键点的微小变化或者甚至某个手势的变化，我们将从一个新的角度去解析手势的特征，当手型处于某个形状时即处于某一个状态，手型的每一个关键点的相对位置具有价值的意义。从这些关键点之间建立一些特征值来描述特殊的含义，即基于不同组合的欧式距离特征。

### 4.4.2基于不同组合的欧式距离特征

由手势中不同的21个关键点组成的210个不同的任意两个关键点距离组成的距离向量，我们将使用基本的距离方法-欧式距离去衡量手型的相似度。对于下述：







基于不同组合的欧式距离特征的特殊意义在于它能够避免不同视觉角度带来的识别问题。对于一个3D对象来说，在一个固定的视觉角度，3D对象不同角度的旋转可能会带来不同的视觉效果。在手语翻译场景中，手势不同的角度变化可能带来不同的识别效果，例如在示范“你”时，有的人偏好指向前方，有的人偏好指向上方，类似于“1”。对于这两个不同的指向方向，但应视为同一个手势。那应该抓住两个不同角度的对象的共同的特征，即3d对象不同视觉角度的不变性：它们的KeyPoint点之间的关系不会改变。

为了验证特征对有效性，设计出一个简单的模型，录入26个字母的不同的手势，每个字母录入10秒钟。参考图4.7，同一个字母a的手势，模型均识别为正确的结果，呈现出特征可以避免不同角度带来的影响。参考图4.8，从字母u到字母微小变化，模型呈现出对于人手不同节点之间的关系是敏感的。

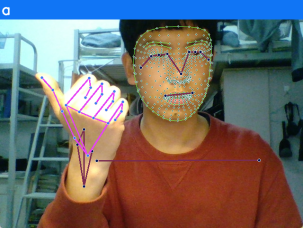
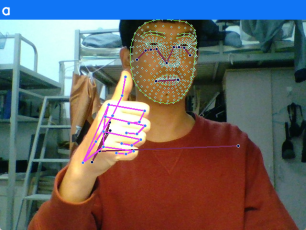
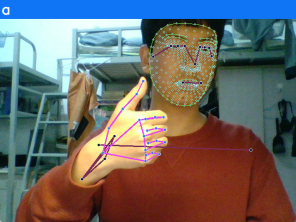
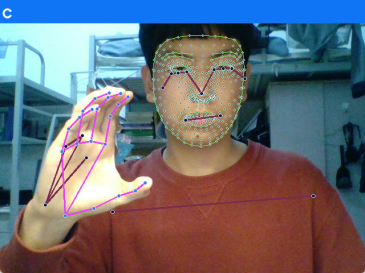
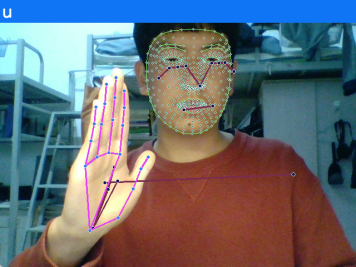
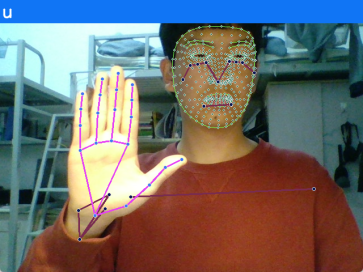


图4.7 不同角度识别a手势的效果图

图4.8 u手势到c手势的识别效果图

### 4.4.5插值与调整序列长度

视频的时间长短各有不同，每个人的手语速度，变化，开始录制时间和结束录制时间都是决定特征向量的长度不同的因素。

系统将要使用这些特征作为学习的特征，使用MLP多层感知器训练模型，在这步骤有一个比较棘手的问题，在设计学习网络结构的时候，在第一层输入单元需要确定固定的向量Shape，面对这样的问题，流行的解决方案有使用padding填充方法，使用循环神经网络进行学习。该系统使用一个插值的方法将所有的长度调整为统一。

插值法是根据已知的离散点计算模拟函数，即构造的函数能够模拟出特征。再给出一系列的x，能够返回对应的函数值。如果已知点有限，可以通过构造多项式函数，给出更多的x点集合，获取更多的点集。 这样即解决了长度短长不一的问题。

插值方法有许多种类，有Lagrange插值法，Newton插值法，Neville插值法，分段线性插值法。该系统使用的scipy模块下的interpolate插值函数，该插值函数默认使用linear插值方法。

## 4.5筛选系统的评估方法

### 4.5.1筛选系统的实际应用场景问题

特征学习可以是有监督的，也可以是无监督的。

* 在监督特征学习中，使用标记的输入数据来学习特征。示例包括监督神经网络、多层感知器和（监督）字典学习。
* 在无监督特征学习中，使用未标记的输入数据来学习特征。示例包括字典学习、独立分量分析、自动编码器、矩阵分解和各种形式的聚类。

该系统的评估方法面对的问题主要在于对于这些的视频没有好坏的标注。从这个角度来看，它属于无监督学习任务。

从采集手语视频数据集的结构角度来看，将会有不同的句子任务类，每个任务类都有不同的人去录制视频。句子任务类别是一个视频的唯一标签。即为了解决评估的方法，我们只有唯一可以利用的标签就是句子任务分类标签。从这个角度来看，它又属于有监督学习。

### 4.5.2多层感知机(Multilayer Perceptron)

#### 4.5.2.1 多层感知机介绍

多层感知器 (MLP) 是一类前馈人工神经网络 (ANN)。 MLP 一词的使用含糊不清，有时泛指任何前馈 ANN，有时严格指由多层感知器组成的网络（具有阈值激活。 多层感知器有时通俗地称为“vanilla”神经网络，尤其是当它们具有单个隐藏层时。

#### 4.5.2.2 使用多层感知机

该系统使用简单的MLP模型作为评估模型，根据输出的计算可以发现，其实隐层的每个神经元是由输入特征x的线性组合构成。然而如果仅仅是线性组合，那么不管这个神经网络有多少层，结果都将与特征线性相关。于是我们在每个神经元结果z之后，添加一个激活函数(Activation Function)，改变线性规则，比如使用Sigmoid函数，Tahn函数，最后我们展开参数，使用全连接层与sigmoid函数配合输出。MLP模型的结构如图8.1：

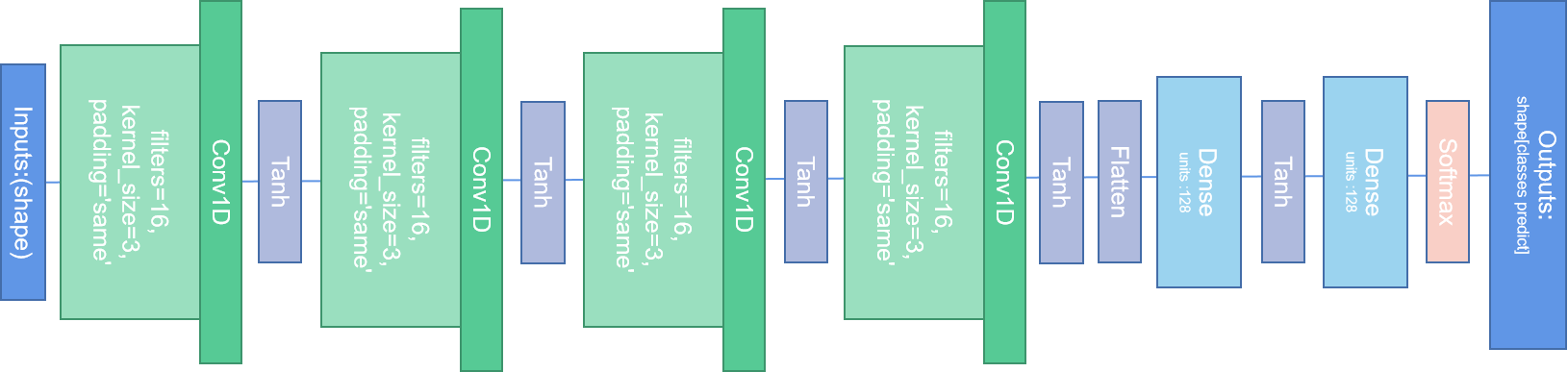


图8.1 MLP模型的结构图

参考4.5.5，一个视频的特征长度为M,也就是经过数据预处理之后的特征的长度，使用线性插值的方法使视频的特征长度变为用户配置的值N。特征长度为N,那么对于上述的多层感知器，不同的特征在输入层都有不一样的Input Shape。如：

* left\_hand\_dist\_net : (441, 100) //左手基于不同组合的欧式距离特征Shape
* right\_hand\_dist\_net : (441, 100) //右手基于不同组合的欧式距离特征Shape

最后输出的形式为：

[class 1 predict, class 2 predict, class 3 predict ..., class N predict]

N为输入数据的手语视频类别个数。Class x Predict 为该视频特定的的特征在这个模型的预测的概率值，即视频属于这些概率值中最大的类。如果概率值越大的话，则视频的类似程度相对较高。

为了提高评估结果的可靠性，系统可以根据用户配置对于不同的特征搭建N次MLP多层感知机模型。同时也为了提高数据的鲁棒性，训练的样本数量有限，且在不同的模型都会不同的训练学习样本。

### 5.5.3 Model Chain 模型链

#### 5.5.3.1 Model Chain模型链的定义

上述的多层感知机做视频评估有局限性，他只能感知到视频的整体的质量，不能够在视频上的时间序列上察觉到手势者错打或者少打或者多打的问题。为了能够解决这次问题，论文设计出一个方案，即在学习视频序列的时候，产生一个自适应Model Chain 模型链，于上面的多层感知机方法不同，它不依赖于别的视频类别的数据。模型链从字面的意思可知他是由多个模型形成的链。

模型还是多层感知机，定义



注意，MLP的输入与之前多层感知机不同，只接受视频序列上的任何一个帧的特征，而不是所有的帧。

#### 5.5.3.2 Hand数据增强

##### 1.**数据增强随机方法**

数据增强的方法，可以避免过拟合，提高模型的鲁棒性，即解决了Model没有学习的样本。在4.4可知，所有的特征都是基于手势的每个关键点的状态。

数据增强的方法定义：





HAugument是对Hand手势的增强函数，HPAugument是对手势其中的任一个关键点的增强的函数。和𝐴𝑋是用户配置的可以接受的值变化范围。

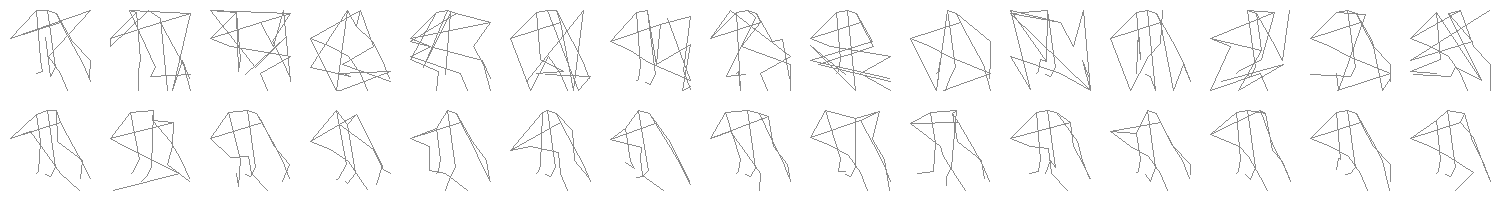
在数据增强的方法上，借鉴对比学习的思路，不仅要在正样本上增强，也要在负样本上增强。正样本增强数据时，应设置为较小的值，且应使值与值的差距较小。负样本增强数据时，应设置为较小的值，且应使值与值的差距较大。由于被归一化处理，和𝐴𝑋的最小值为0，最大值为1。

##### 2.**数据增强随机方法实现**

实现数据增强随机方法为：

HandRandomAugment(hand,times, goodRange, badRange)

HandAugment方法会根据hand、goodRange、badRange输出times个正样本和负样本。如goodRange设定为(0,0.2),badRange设定为(0.3,1),times设定为15，会输出以下的结果。左图为输入的hand，右图分别上行为数据增强后的负样本，下行为数据增强的正样本。

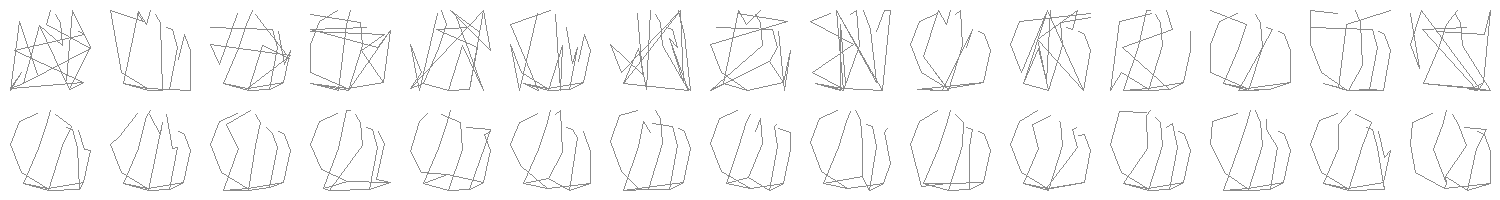
 

图4.5 hand数据增强效果图

##### 3.基于骨骼约束的数据增强

随机方法产生的正负样本，使用多层感知器的效果并非理想。做出一个小的测试，结果如下：



图5.5 hand序列图

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Handx | Hand1 | Hand2 | Hand3 | Hand4 | Hand5 | Hand6 | Hand7 | Hand8 | Hand9 | Hand10 |
| Result | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 | 0.99 | 0.98 | 0.97 | 0.97 | 0.98 |

表5.3 hand序列评估结果表

第1个到第6个的hand与数据增强的输入hand相似，评估分数理应接近1。但是之后的hand并非相似，评估分数不应该接近1。

产生这个问题的原因可能是模型并没有对数据增强输入的hand产生模拟的能力，需要进一步产生别于数据增强的输入hand，即产生更大范围的基于骨骼约束的手势。

手部可以建立如图5.6所示的人手骨骼的约束模型，参考图5.7了解手部骨骼，可知人手由月骨，三角骨，掌骨和指骨四大部分构成。由骨骼的生理结构可知，小拇指，无名指，中指，食指的拇指只能伸直或内收，无法反向向外伸开。

指骨部分的关节处的夹角只能受限于关节处附近两个拇指节的两条线的共面，即只有一个活动自由角度。掌骨与手心的夹角（橙色部分）最小值为90°，最大值为180°。蓝色部分的夹角最小值为80°，最大值为180°，指骨尖端部分的夹角（粉红色部分）最小值为100°，最大值为180°。

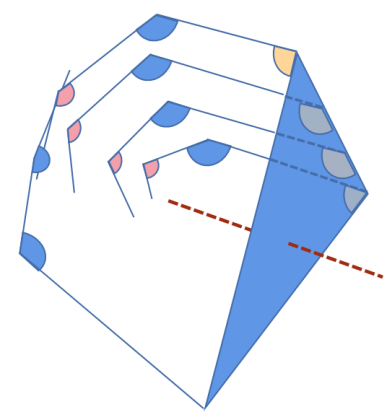


图5.6 人手骨骼约束模型

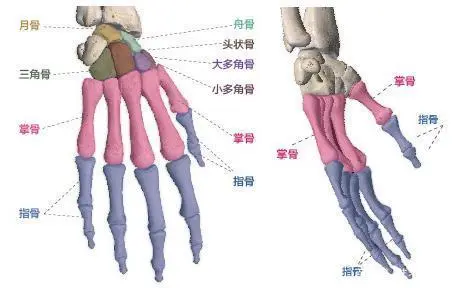


图5.7 手部骨骼解剖结构

指骨部分的关节活动可以抽象为在一个空间上，设以点O为中心长度为OA的圆C，如果给定∠AOB的角度为θ，点B只能活动在圆M的边上，求出满足条件的点B的坐标。设O的坐标为(),设A的坐标为(),B的坐标为(x,y,z)求解方程为：

其中是OC和OA的共面的平面的法向量的，是AO两点之间的距离。



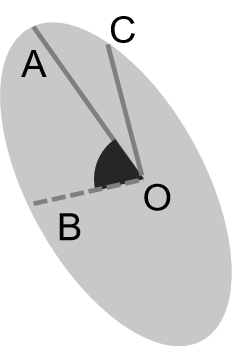


图5.8 指骨部分的关节活动计算抽象图

掌骨与手心的关节活动可以抽象为在一个空间上，OC围绕AD旋转到OB。

使用上诉的旋转方法旋转手指很难到达离校的效果，如图5.9，旋转有更多的细节问题需要考虑，如多个指节之间的关系等等。

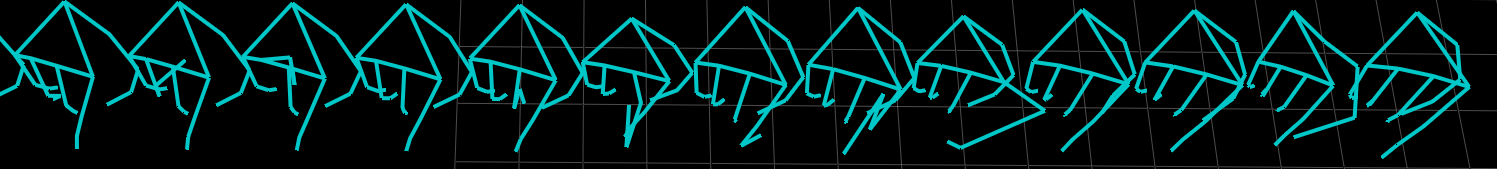


图5.9 手势旋转操随机生成效果图

##### 4.现有样本抽取方法生成正负样本

由于时间限制，我们将从现有的数据集里抽取随机样本以增强模型的模拟能力。为了观察人手不同指节之间的关系，列出以下关系:系统将会计算表格里的以A, B, B三点组成的∠AOB角度。表格4.1是人手所有两个连接的指节的角度。表格4.2是人手指尖和手心以掌骨部分和指骨部分的连接处为中心的夹角关系。

|  |  |  |
| --- | --- | --- |
| A | O | B |
| 0 | 1 | 2 |
| 1 | 2 | 3 |
| 2 | 3 | 4 |
| 0 | 1 | 6 |
| 5 | 2 | 7 |
| 6 | 3 | 8 |
| 0 | 1 | 10 |
| 9 | 2 | 11 |
| 10 | 3 | 12 |
| 0 | 1 | 14 |
| 13 | 2 | 15 |
| 14 | 3 | 16 |
| 0 | 1 | 18 |
| 17 | 2 | 19 |
| 18 | 3 | 20 |

表格4.1

|  |  |  |
| --- | --- | --- |
| A | O | B |
| 0 | 2 | 4 |
| 0 | 3 | 4 |
| 0 | 5 | 8 |
| 0 | 6 | 8 |
| 0 | 9 | 12 |
| 0 | 10 | 12 |
| 0 | 13 | 16 |
| 0 | 14 | 16 |
| 0 | 17 | 20 |
| 0 | 18 | 20 |

表格4.2

以10个类，每一个类均有70多个样本，统计结果以图5.8展示。据观察，表格4.1的关系除了大拇指指节之间的角度均为0°到60°,其他指节之间的角度均为60°到180°。表格4.2的组成的关系分布均为0°到100°。

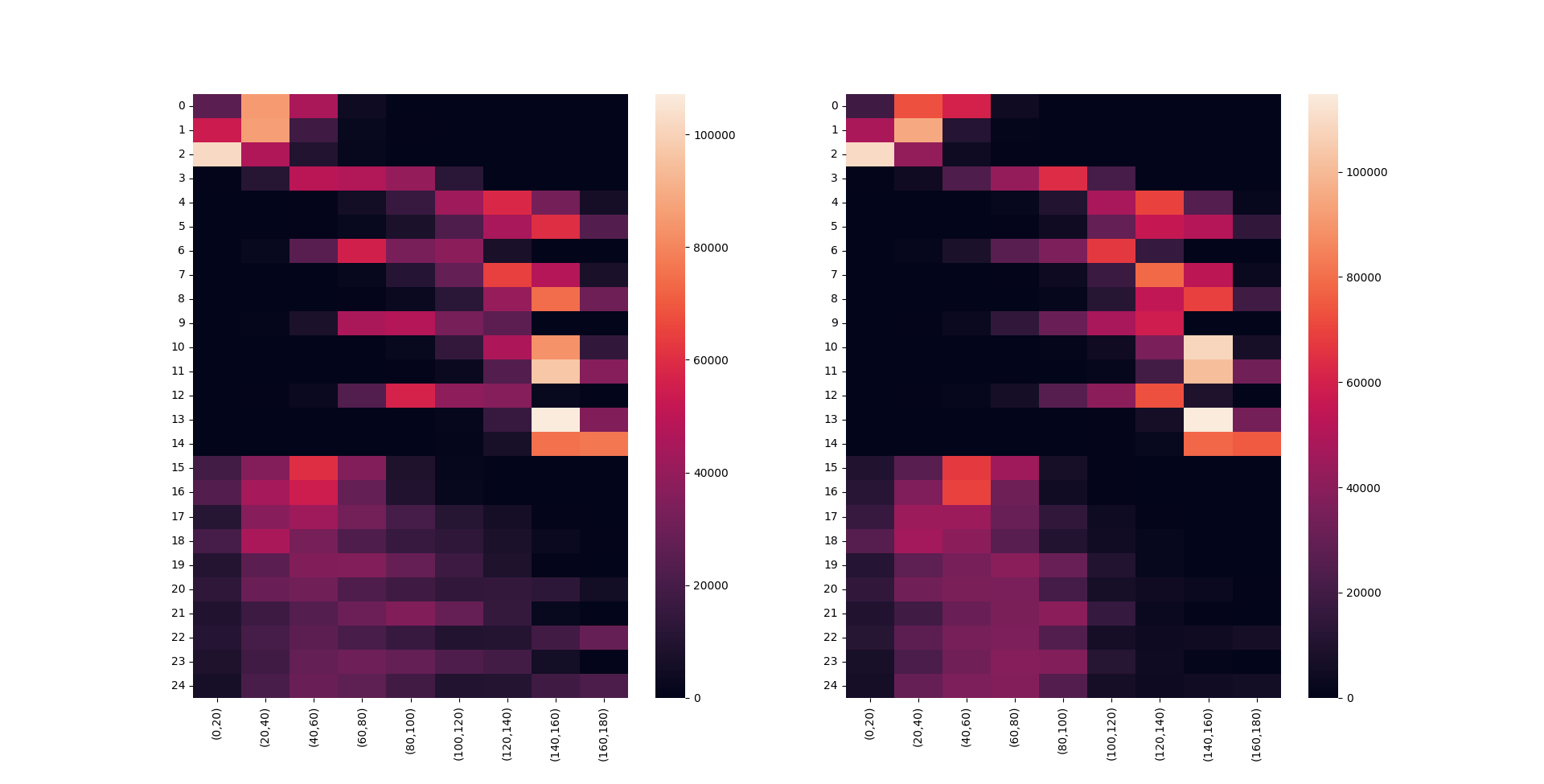


图5.8 角度分布热图

正负样本生成的思路为以一个参考的手势，计算他的不同关系的角度，第一部分的角度为表格4.1的关系，第二部分的角度分布为表格4.2的关系。在现成的样本里抽取任意一个手势，第一部分的角度与参考手势的角度最大差异如果小于20°，且第二部分的角度与参考手势的差异小于40°，添加进正样本集合里，并且使用上述的数据随机增强方法增强正样本。如果这两个条件有一个不满足的话，则放进负样本集合里，不适用随即增强的方法。使用该思路的方法测试效果如图5.9和表格5.3，图5.9的手势均来自与参考手势同一个的视频，第1个到第6个的结果均高于80％，其他的手势均接近于0％。在测试其他的手势，均到达理想的效果,如图5.10和表格5.4。



图5.9 测试手势

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Handx | Hand1 | Hand2 | Hand3 | Hand4 | Hand5 | Hand6 | Hand7 | Hand8 | Hand9 | Hand10 |
| Result | 0.81 | 0.91 | 0.93 | 0.59 | 0.72 | 0.90 | 0.00 | 0.11 | 0.00 | 0.00 |

表格5.3 测试结果



图5.10 测试手势

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Handx | Hand1 | Hand2 | Hand3 | Hand4 | Hand5 | Hand6 | Hand7 | Hand8 | Hand9 | Hand10 |
| Result | 0.95 | 0.72 | 0.33 | 0.51 | 0.00 | 0.74 | 0.00 | 0.93 | 0.96 | 0.97 |

表格5.4 测试结果

使用图5.10中的第一个类似的手势测试其他同一个类的视频时，发现一个问题。图5.12左部分的手势却没有较高的分数，然而右部分的手势分数较高。仔细观察可发现，图5.10的参考手势的无名指和小拇指是伸张的，图5.12中的手势是收缩的。另外，图5.12的手势里的食指和中指处于紧接状态，和图5.10的手势状态接近，但存在于角度的问题。这两个图中的手势需要用角度判断来区分。

图5.12 测试其他视频的手势

#### 5.5.3.3 Model Chain 结构

为了清楚的介绍Model Chain的结构和原理，给出图5：

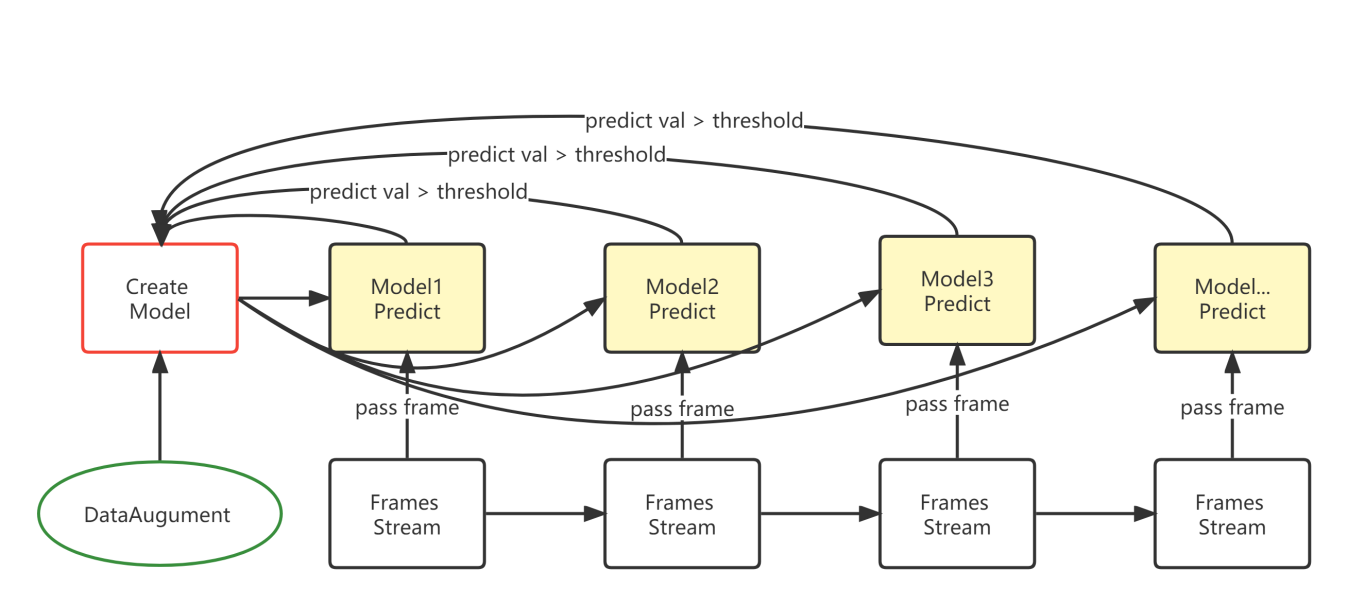


图5. Model Chain 结构图

在开始的阶段，Create Model 组件会创建一个新的模型，并且从Frames Stream 获取下一个帧的特征，在此特征的基础上使用数据增强的方法增加训练样本，提高模型的模拟能力，经过一定的迭代训练次数之后，从Frames Steam获取下一帧的特征做评估，如果此次的评估值大于用户设定的Threshold值，则再创建一个新的模型，开始创建的Model模型和此次创建的模型形成含有两个模型的模型链，如果识别再次使用数据增强的方法，继续上述的步骤。如果低于用户设定的threshold值，则继续从Frames Stream获取下一帧的特征。

#### 5.5.3.4 Model predict Mask

为了加快Model Chain 对每一个视频的评估，系统将聚集同一类的视频所有手势，同时建立Mask。Mask主要是作用于记录每一个手势的状态，如图5.15，Mask的每一个单元都对应这一帧的手势的状态，初始值为0，表示为未被处理，如果该帧没有对应的keypoint，则记为-1。获取第一个值为0的mask，并且去使用这个手势作为参考手势train mode。完成train model，去predict 所有的手势，如果被识别为相同的手势则标记为同样的记号，如蓝色的方块为1，他们都是同一类手势。标记完成之后，model chain会获取下一个标记为0的手势继续执行。

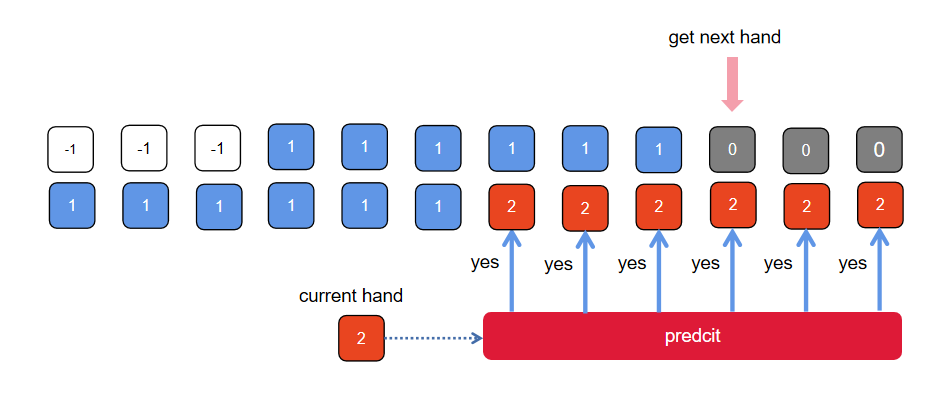


图5.15 Mask get next hand

同一类的73个样本，聚集之后使用第一个手势作为参考手势，执行predict之后的结果如图5.13，白色表示对应的手势predict结果较好，黑色表示对应的手势与参考手势无关，结果分布图如5.14，可见无关的手势占大多数，相关的手势值接近于1。仔细观察图片，更多的同类手势是连续的，此方法可以作为手势分割的思路，如手势标记，手势搜索这些场景上。

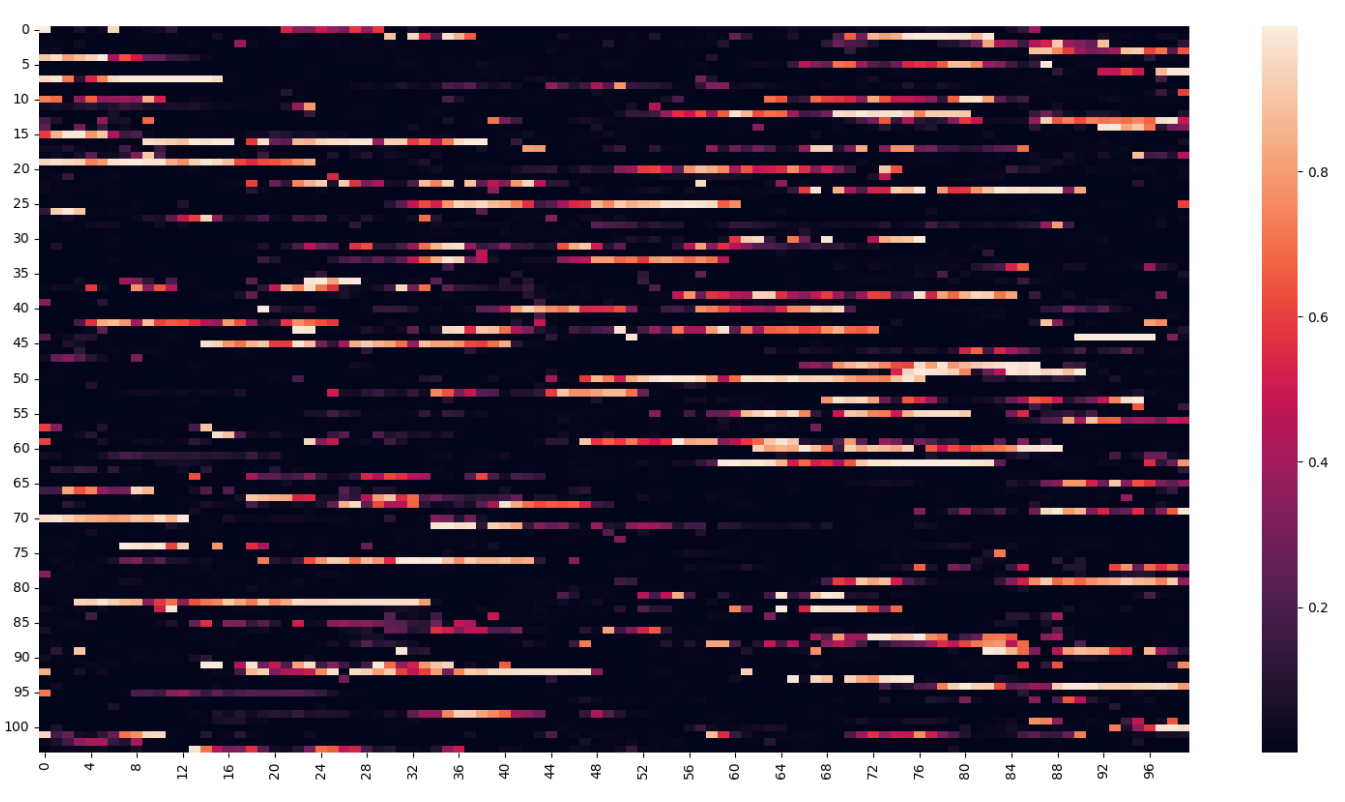


图5.13 modle predict 所有手势的热图

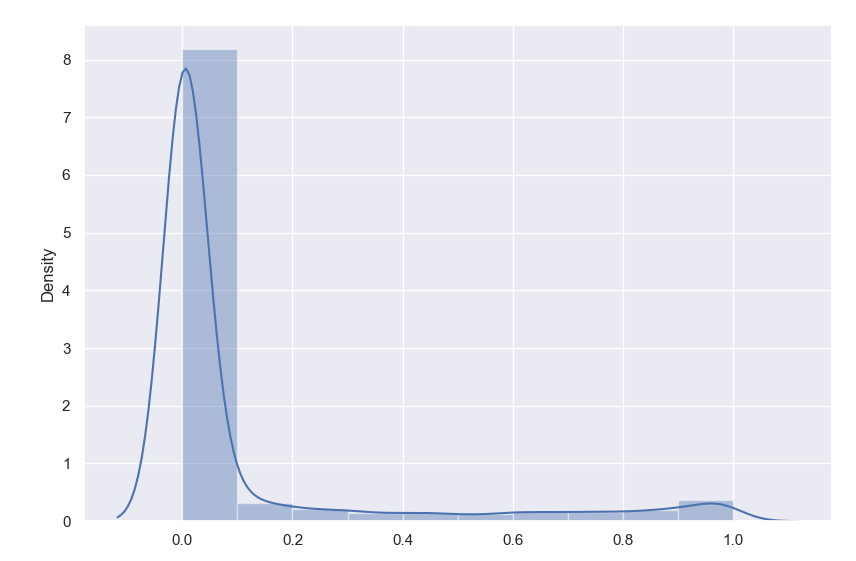


图5.14 modle predict 所有手势的热图

## 4.6 筛选系统的界面端

### 4.6.1 Qt框架介绍

Qt 它是一个跨平台的 C++ GUI 应用程序框架，为应用程序开发人员提供构建最先进的图形用户界面所需的所有功能。Qt 易于扩展，并允许真正的组件编程。Qt 和 GTK、KDE、MFC、 OWL、 VCL、ATL 是同一个GUI库。

QT优势如下：

1. 优秀的跨平台特性

Qt 支持以下操作系统：Microsoft Windows,Linux, Mac OSX等。

2. 面向对象

Qt 良好的封装机制使得 Qt 非常模块化，复用性好，非常方便用户开发。

3. 丰富的API

Qt 包括多达 250 多个 C++ 类，它还提供基于模板的集合、序列化、文件、I/O 设备、目录管理、日期/时间类。

### 4.6.2筛选系统界面端的不同组件

#### 4.6.2.1表格目录和视频目录

从手语采录系统上下载的数据包括表格目录和视频数据。表格数据的信息包含视频的一些标签，如task id，submit id，user id 等等。为了读取这些数据，需要用户输入表格的目录和视频的目录。

#### 4.6.2.2参数设置

#### 4.6.2.3 日志系统

数据的类别个数未满足用户的要求类似的信息将会在日志系统输出，告知用户程序未运行的原因。

### 4.6.3筛选系统实现细节问题

#### 4.6.3.1 多线程

筛选系统的运行结构类似于LinePipe流水线，在实现前端界面有一个严重的问题。仅仅是简单的实现流水线，会导致界面端的无法正常显示，也无法正常使用。

这需要多线程将计算的部分异步化，这样即不会影响到用户界面端的正常使用。下面给出实现思路：

计算的部分：

def Run():

//run code...

Xxx

界面端的部分：

def StartRun():

runThread = Thread(target=Run) //建立一个线程

runThread.start() //开启线程的运行

假使Run()是计算的部分，那么只需要建立一个线程，把这个分配给这个线程，开启这个线程的运行，这时候界面端的代码不会在这等待Run的计算，而是继续执行接下来的代码。

#### 4.6.3.2 界面端和计算部分的通信

上述解决了界面端的卡顿现象，但是多线程将界面端和计算组件完全给分开了，假使用户开启运行的时候，计算组件开启运算，一旦完成运算，界面端无法获知计算组件是否完成。Qt 提供了一种叫做信号/槽的类型来代替回调，它使得组件之间的协作变得非常简单。信号是用于发生的时间，槽是用于接受信号的发送。

实现思路如下：

def Run():

//run code...

Xxx

EndSignal.emit(End) // 发出信号

信号绑定：

EndSignal.connect(HandleEngSignal)

接受信号将执行的部分：

def HandleEngSignal():

//Process code ...

Run()完成计算之后发出EndSignal信号，在界面端的部分需要定义槽，即信号绑定。一旦发出信号，界面端会执行HandleEndSignal部分做相关的处理。

# 总结

手语数据筛选系统最开端从Openpose工具安装，尝试使用DTW匹配算法去检测，使用小波过滤过滤特征，到现在使用模型链的方法去评估每一个视频的手势。

中间碰到了不少的瓶颈，比如Openpose的安装比较繁琐，算法已经过旧。DTW算法比较耗时，且不适合用于收拾轨迹的匹配。为了到达效果，不断地重新设计方法，之前的特征有手势轨迹，还有手势的角度Net等等特征，最终只保留毕业论文中的基于不同组合的欧式距离特征。由于多层感知机分类的方法评估只能感知视频的相似程度，为了更进一步观察手语视频的细节。在毕业设计期间，设计了一个新的思路，ModelChain模型链评估方法。在具体设计模型链方法之前，有一个目标，希望能过做出能够察觉视频少打手语或者多打手语，还要能够判断手势顺序的对错。在具体实现过程当中，遇到了不少的问题，如ModelChain的Model的参数是如何训练，如何能够认识手势。借鉴了对比学习的思路，增强正负样本，在设计手势的增强方法时，随机方法显然不能让模型对手势有准确的认识。为了解决这个问题，想出基于骨骼约束的数据增强方法，这个方法更先进，但是实现难度大。由于时间有限，只能先用现成的样本里抽取样本增强正负样本，到达了认识相似手势的效果。在这里有一些问题要提出，ModelChain可以判断视频是否少打手势，但是不能判断视频是否多打手势。因为多大手势很有可能是视频的帧率和分辨率导致，解决这个问题需要更多的语义智能判断，如果做到了一个语义的判断，那就离翻译更进一步了。再次提出，ModelChain方法也可以用于手势自动标记，手势搜索这些场景。

# 参考文献

1. 袁甜甜,赵伟,杨学,胡彬.大规模连续中国手语数据集的创建与分析[J].计算机工程与应用,2019,55(11):110-116.
2. 黄杰.基于深度学习的手语识别技术研究[D].中国科学技术大学,2018.
3. 江勇军.基于Kinect的孤立词手语识别系统研究[D].中国科学技术大学,2015.
4. 杨明羽,叶春明.结合Bi-2DPCA与CNN的美式手语识别[J].计算机工程,2021,47(12):278-284.
5. 王婧瑶,范飞,刘豪宇,蒋钰雯.基于机器视觉的聋哑人手语识别——语音交互系统[J].物联网技术,2021,11(12):3-5.
6. 雷明焕,何高明.基于深度学习的中国标准手语视频分类方法[J].电子技术与软件工程,2021(23):162-165.
7. 杨淑莹,赵敏,郭杨杨,田迪.基于改进的EfficientDet的手语识别算法[J/OL].微电子学与计算机:1-10.
8. 朱晨冰，李建英.Qt5.12实战[M].清华大学出版社.2020.
9. [1]高宇翔,顾定倩.中国手语的发展历史回顾[J].当代语言学,2013,15(01):94-100+126.
10. [1]吴铃.汉语手语语法研究[J].中国特殊教育,2005(08):15-22.
11. [1]余晓婷,贺荟中.国内手语研究综述[J].中国特殊教育,2009(04):36-41+29.
12. [1]倪兰.中国手语动词方向性研究[D].复旦大学,2007.
13. [1]张晓梅.中国手语与美国手语的比较[J].长春大学学报,2009,19(12):125-126.
14. [1]姜春雷. 基于MPEG-4的视频压缩编码的设计与实现[D].大庆石油学院,2006.
15. 人体姿态估计--OpenPose算法解析 <https://www.jianshu.com/p/98c11545d4fb>
16. MediaPipe 集成人脸识别，人体姿态评估，人手检测模型
17. <https://zhuanlan.zhihu.com/p/353861059>
18. MediaPipe Holistic谷歌面部，手部与姿势预测
19. <https://www.toutiao.com/i6909453618743804428/?group_id=6909453618743804428&wid=1644821142596>
20. <https://blog.csdn.net/weixin_45288557/article/details/109278843>
21. <https://blog.csdn.net/codingpy/article/details/109913018>
22. <https://www.bilibili.com/read/cv4531460/>

# 致谢