山东师范大学

《人机交互技术》

期中作业

题　　目：针对缺失模态的无障碍人机对话技术

学生姓名：隋　远

学　　号：201911010105

专　　业：计算机科学与技术(非师范)

指导教师：杜　萍

学　　院：信息科学与工程学院

2021 年 11 月 5 日

（一） 项目简介

信息无障碍(information accessibility)是一个学科交叉的技术和应用领域，旨在用信息技术弥补残障人士生理和认知能力的不足，让他们可以顺畅地与他人、物理世界和信息设备进行交互。据中国残联统计，中国现有8500万残疾人，是世界上残疾人口最多的国家。其中，听力残疾2000万人，视力残疾1200万人，各类肢体残疾2500万人，智力残疾和精神残疾1200万人……随着社会老龄化程度加重，残疾人口数量也在持续增长。互联网和用户终端的普及，使得信息无障碍成为一个越来越值得关注的领域，目标是解决残障人士的信息访问甚至是生活服务问题。

本项目借助百度开放平台，搭建一套面向缺失模态人群的信息无障碍人机交互系统。

（二） 用户需求

残障类型多样，用户需要的无障碍技术也不尽相同，本项目面向三类主要的残障类型（视觉障碍、听觉障碍和运动障碍）人群遇到的问题和主要的技术方案。

2.1 视觉障碍用户需求

视力残疾用户的需求包括**独立出行、识别身边物体、与信息设备交互等**。针对独立出行的需求，目前有基于计算机视觉的道路识别技术，通过立体声场或者震动反馈为视力残疾用户指示方向。但是这些设备目前还不能取代盲杖，还需要更多的技术突破。针对识别物体的需求，主要是利用视频/图像转换为文本的技术，包括微软的Seeing AI和谷歌的Lookout都是此类应用。针对使用手机和电脑的需求，主要采用读屏程序screen reader（苹果手机上的VoiceOver或者安卓系统上的Talkback，均为系统默认自带功能），可以通过语音读出获得焦点的控件信息，这样视力残疾用户通过听就能了解设备界面上的信息内容。

2.2 听力障碍用户需求

听力残疾用户面临的主要问题是**与人交流存在障碍，以及观看视频内容时听不到声音**。老年听力障碍是指随着年龄增长，听觉器官的衰老和退变所导致的听觉功能下降，发病率居世界第三位。助听器设备通过放大声音信号，可解决“听不到”的问题；但对于听觉中枢受损的人，声音信号分析能力却难以弥补，解决不了“听得清”的问题。针对听障用户，相关信息无障碍技术包括手语的识别与合成，以及语音识别技术。

2.3 运动控制能力障碍用户需求

运动控制能力缺失的用户，包括上肢残疾，或者患帕金森症、脑瘫、肌肉萎缩、渐冻症等疾病的用户。他们**丧失了灵活控制手指运动的能力，而手指是人表达交互意图的主要运动器官，也是电脑和手机的主要操作器官**。在构建面向这类用户的信息无障碍交互技术时，其中一个难题是用户的差异性，几乎每个用户的可运动部位及其运动能力都是不同的，给构建适合于个体的通用输入技术带来了挑战。相关的技术方案有眼动，但是用“眼动”作为输入方式时，缺少“确认”操作，容易产生误触发，且操作精度有限。

（三） 任务分析

3.1 信息无障碍的主要科学问题

音视频的理解和信息转换（主要针对听障和视障）。视觉和听觉是人们接受信息的主要感官。听障和视障用户因为缺乏某种感官而无法完整理解信息，需要建立音视频的理解技术，用机器算法理解音视频内容的语义，进而转换为用户可用感官能接受的信息类型，包括音频和文字之间的语音识别和文本到语言(Text To Speech, TTS)技术，图像到文字和视频到文字的技术。目前，精度是主要问题，尤其是克服多种噪声条件下的高精度实现，对于这些技术的可用性起到关键作用。

图形用户界面到声音界面的编码转换（主要针对视障）。个人电脑和手机都是图形用户界面，信息以可视的方式传递给用户，而视障用户只能通过听觉（触觉为辅）来接收信息，相比于视觉，不仅信息接收的带宽要低很多，而且信息呈现的模式也发生了变化。视觉提供整体和并行的信息获

取能力，听觉只能提供局部串行的信息。这也会影响用户对于交互界面的心理模型，进而影响到交互决策。因此，需要研究从图形界面到声音界面的编码转换方法，优化“读屏”的方法。

个性化信息输入和意图理解（主要针对视障和运动障碍）。人体的运动控制系统包括运动执行和反馈两部分。运动障碍用户无法精确灵活地控制手指运动，视障用户由于缺少视觉反馈也不能做精确的输入控制，导致物理运动自由度受限和运动控制精度低的问题。前者需要开发具有个性化能力的输入技术，根据用户实际可以控制的输入范围来映射有效的输入；后者需要实现从有噪声的运动控制数据中提取用户的交互意图。

3.2 通过智能交互技术实现信息无障碍

信息无障碍是以用户为中心的交互方案，是对人的交互性能的优化。优秀的信息无障碍技术要适应用户的生理和认知能力，而不是让用户适应技术。为此，要采用智能交互方法来开展研究，从用户角度来设计和创新适用的交互模式，通过智能传感、智能用户意图推理和智能信息呈现来构建信息无障碍的交互界面。

对用户行为和认知能力的准确建模：需要建立用户动作能力和心理模型的计算模型，建立并引入生理、心理的先验知识来描述用户的信息输入输出能力，对用户意图、表达方式、动作控制能力做统计建模和描述。残障用户的一个重要特点就是个性化，每个用户的信息输入输出能力都不同，除了需要研究合适的模型函数，还要研究个性化参数的计算方法，力求能准确地描述个体残障用户的能力。

智能的感知技术：需要研究高精度的感知技术，感知用户的外在动作，也感知用户的内在心理状态。信息无障碍中，智能感知的应用场景非常丰富，针对不同类型的残障用户、不同的交互任务、不同的使用情景，都需要适合的传感方案。哪些动作或者心理状态是有交互价值的，如何采用成本可控、易于部署的硬件方案都是需要考虑的问题。

智能的意图推理技术：残障用户通过自然动作（手、眼或身体其他部位）表达交互意图。优化信息无障碍的交互体验，需要使用户在表达意图

时的生理和心理开销最小化，但这样将导致用户的表达方式和表达动作都是不精确的，在时间和空间上都存在随机性。如何从连续随机的行为数据中提取用户的输入意图，是需要解决的问题。作为操控型的交互界面，需要具有高精度、可理解、结果可预测的特点。

智能的信息呈现：感官残障用户对于交互界面的理解是不完整的，交互决策的心理模型带有随机性。高可用的信息无障碍交互界面，首先需要有对用户信息需求的预测能力，确定信息输出的目标；然后根据用户的信息接收能力将目标信息编码到具体模态上，编码方式涉及多模态融合；最后根据交互情景，对信息呈现的编码方式做动态优化调整，保证用户接收信息的有效性。

（四） 功能模块

4.1 输入模块

针对不同的用户需求，需要设计不同的输入模块，如针对听障和视障人群，播放音视频时用户因缺乏某种感官而导致无法完整理解信息，需要建立音视频的理解技术，此时输入模块应当根据用户缺乏模态信息，将特征输入机器学习算法理解音视频语义，进而转换为用户可用感官能接受的信息类型。

4.2 实时语音识别模块

通过语音识别实现人机对话，解决视障用户需求。将语音对话实时识别为文字，实现自然流畅的人机对话。

输入：输入接口同时接受两种输入格式：（a）实时音频流输入，要求上传实时，不能过快，即整体耗时略多于原始音频流，若因为网络不稳定不过导致需要重新发起请求续传的，允许超发一段 XXms 的音频，待网络恢复后将全部音频传给服务端。（b）音频文件输入，支持 pcm 格式的音频文件，每 160ms 为一帧发送，间隔 1-2ms，整体耗时短于音频流输入。

处理：调用百度开放平台实时语音识别 websocket API 接口，进行实时语音识别，并将转换得到的文字信号做进一步文字解析后，传递给语音

生成模块。

4.3 语音合成模块

通过语音合成将机器生成的文字信号转换为语音信号传递给视障用户，完成信息交互。

输入：输入接口接受各类文本格式文件，使用 UTF-8 编码，要求文本长度必须小于 1024 字节；若文本较长，可以采用多次请求的方式。

处理:调用百度开放平台在线语音合成 SDK 接口,获得语音合成能力。

输出：输出指定音频格式的语音信号，借助播放设备播放语音，完成流程的人机交互。

4.4 图像识别模块

针对失声障碍人群，通过图像识别模块，识别手语信息，并转换为文字信号，实现人机交互。

输入:输入接口接受手语图像,要求图片可正常解码、长宽比例合适，背景相对模糊。

处理：借助百度开放平台搭建手语识别模块，对常见的手语信号进行文字转换。

输出：输出手语信息对应的文本信号，完成人机交互。

4.5 眼动识别模块

针对运动控制能力障碍人群，通过眼动识别模块，实现与机器的交互操作。

输入：输入接口利用移动端设备摄像头捕捉用户眼动信号

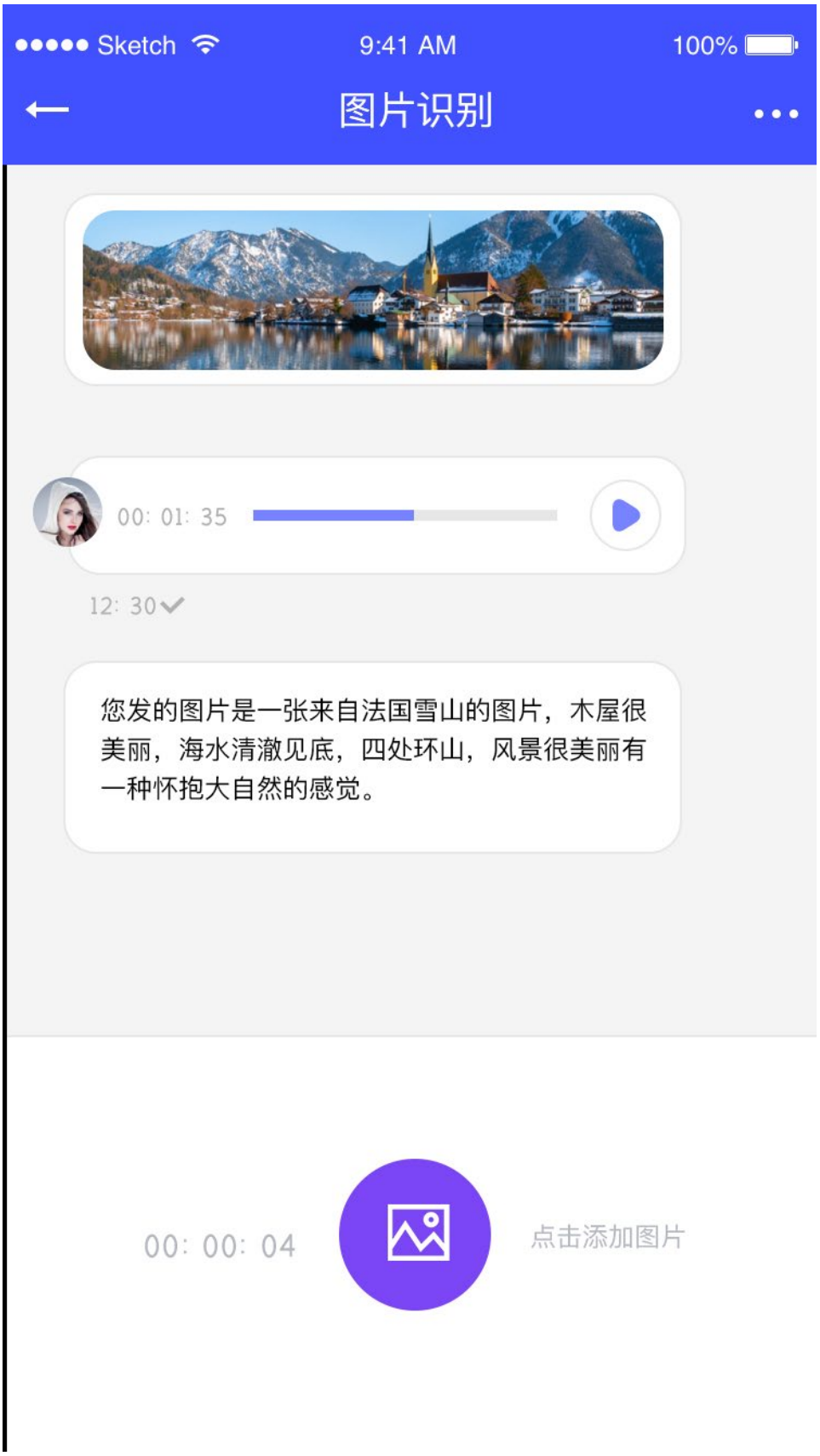
处理：通过系统内嵌眼动识别模块，对用户眼动信号进行分析，推断用户眼动信息，从而实现与机器进行交互。

输出：一系列机器交互指令。

（五） 界面设计


Sketch 9:41 AM 100%
请说出您的需求
00:54
图像识别
语音识别
眼动识别
Sketch 9:41 AM 100%
图片识别
00: 01: 35
12: 30
您发的图片是一张来自法国雪山的图片，木屋很美丽，海水清澈见底，四处环山，风景很美丽有一种怀抱大自然的感觉。
00: 00: 04
点击添加图片
Sketch 9:41 AM 100%
眼动识别
请上下转动眼球，实现翻页功能
00:54
00: 00: 04
点击后立即对准镜头
Sketch 9:41 AM 100%
语音识别
马上快到元旦了，李俊福祝福您老身体健康，福寿安康，万事如意，带我向家人问好，给予新春的祝贺！
12: 00
谢谢你的问候，也待我向你家人问好
元旦过后相见！
祝元旦快乐！！！
12: 15
00: 01: 35
12: 30
00: 00: 04
按住语音说话

首页:用户可通过点按的方式或语音输入的方式，选择各个功能模块，如图像识别、语音识别以及眼动识别。



图片识别模块：用户可以通过下方的点击添加图片，添加图片，机器将同时返回语音信息和文字信息。

眼动识别模块：用户通过点按下方按钮，点击后可以立即对准镜头，此时机器将根据用户眼动信号完成一系列指令（如翻页等功能）；针对视觉障碍人群，也可以通过语音输入的方式初始化机器，进行一系列指令的输入。

语音识别模块：用户通过点按下方按钮，按住语音说话，机器自动将识别到的音频信号转换为文字信号并输出到显示屏幕上，完成人机对话。