Computer Science and Application 计算机科学与应用, 2024, 14(6), 81-93
Published Online June 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.146144

Hans汉斯

智能交互系统的设计与实现：基于人脸识别与语音识别技术

李　玲

北京工商大学计算机与人工智能学院，北京

收稿日期：2024年5月21日；录用日期：2024年6月20日；发布日期：2024年6月27日

摘　要

本文介绍了一种基于人脸识别与语音识别技术相融合的多模态智能交互系统。该系统由人脸识别模块及语音识别模块两大部分组成。通过集成openMV摄像头、麦克风阵列以及openMV IDE软件环境，开发一种多模态系统，该系统能够实现特征点提取与检测，并结合这些功能进行语音增强、语音识别和人脸识别。openMV摄像头进行图像采集，并在openMV IDE软件端执行特征点检测算法，捕捉用户的面部特征，实现身份验证和用户信息的获取。同时，麦克风阵列将负责捕获声音信号。语音增强模块通过运用基于时频卷积网络(TFCN)的轻量级语音增强算法，抑制背景噪声，保持目标语音的失真尽可能低，实现对目标语音的增强。语音识别模块实现了从语音到文本的转换，提升系统的智能化水平。该系统可广泛应用于智能家居领域，具体来说，可以应用于智能门锁，该系统可以自动识别家庭成员的面孔，实现无钥匙进入。此外，语音识别模块可以识别出特定的语音命令，如“开门”或“关门”，从而进一步增加智能门锁的便捷性和安全性。实验结果表明，本智能交互系统通过融合人脸识别与语音识别技术，成功开发了一种多模态智能交互系统。这一集成化的设计不仅体现了系统的高效性和稳定性，更预示了该系统在未来广泛应用中的巨大潜力和实用价值。

关键词

人脸识别，openMV，单通道语音增强，语音识别，智能交互系统

Design and Implementation of Intelligent Interaction System: Based on Face Recognition and Speech Recognition Technology

Ling Li

School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing

文章引用: 李玲. 智能交互系统的设计与实现: 基于人脸识别与语音识别技术[J]. 计算机科学与应用, 2024, 14(6): 81-93. DOI: 10.12677/csa.2024.146144

Received: May 21st, 2024; accepted: Jun. 20th, 2024; published: Jun. 27th, 2024

Abstract

This paper introduces a multi-modal intelligent interaction system based on the integration of face recognition and speech recognition technology. The system consists of two parts: face recognition module and voice recognition module. By integrating the openMV camera, microphone array and openMV IDE software environment, a multi-modal system is developed, which can realize feature point extraction and detection, and combine these functions for voice enhancement, speech recognition and face recognition. The openMV camera collects images and executes a feature point detection algorithm on the openMV IDE software to capture the user's facial features and realize authentication and user information acquisition. At the same time, the microphone array will be responsible for capturing the sound signal. The speech enhancement module uses a lightweight speech enhancement algorithm based on time-frequency convolution network (TFCN) to suppress background noise, keep the distortion of the target voice as low as possible, and realize the enhancement of the target voice. The speech recognition module realizes the conversion from voice to text and improves the intelligent level of the system. The system can be widely used in the field of smart home. Specifically, it can be applied to smart door locks. The system can automatically identify the faces of family members and achieve keyless entry. In addition, the voice recognition module can recognize specific voice commands, such as "open the door" or "close the door", thus further increasing the convenience and security of the smart door lock. The experimental results show that this intelligent interaction system has successfully developed a multi-modal intelligent interaction system by integrating face recognition and speech recognition technology. This integrated design not only reflects the efficiency and stability of the system, but also indicates the great potential and practical value of the system in its wide application in the future.

Keywords

Face Recognition, openMV, Single-Channel Voice Enhancement, Voice Recognition, Intelligent Interaction System

1. 引言

智能交互系统作为人工智能技术的重要应用之一，在日常生活中扮演着越来越重要的角色[1]。通过深度融合人脸识别和语音识别技术，可以实现更加智能化、便捷化的人机交互方式，极大地提高了信息获取和沟通的效率。然而，现有的智能交互系统在处理噪声干扰及复杂场景下的人脸识别时仍面临挑战。背景噪声对语音识别技术的干扰尤为显著，而人脸识别技术在光线不足、面部表情变化或佩戴遮挡物等复杂场景下，其识别准确率可能会受到严重影响。针对这一问题，贾徐鹏和李冬梅提出了基于时频卷积网络(TFCN)的轻量级语音增强算法。TFCN 旨在预测并恢复纯净语音，保证在嘈杂环境下语音的清晰度和识别性。相比其他算法，TFCN 以 93,000 个参数即实现高性能，展现其高效与实用。本文在以上研究的基础上，为了提高语音和人脸识别准确率，基于性价比高的 openMV 和 HMI 串口屏

搭建了一种新的智能交互系统，并基于时频卷积网络(TFCN)的轻量级语音增强算法，抑制背景噪声，提高目标语音的清晰度和可识别性。并采用先进的特征点检测算法，以捕捉更准确的面部特征，提高面部识别的准确率。

2. 国内外研究现状

人脸识别与语音识别技术，作为现代智能交互系统的核心技术，其研究历史可追溯至较早的时期，并已经积累了显著的科研成果。特别是在 21 世纪的技术浪潮中，随着深度学习等先进技术的不断推动，这两项技术均获得了巨大的进步。人脸识别技术已经从初始的二维图像识别阶段，逐步演进至更为精确的三维识别技术，显著提升了识别的准确性和可靠性。同时，语音识别技术也在语音增强、语音识别等关键领域取得了重要突破，极大地提升了语音信息的处理效率和准确性。

尽管人脸识别与语音识别技术各自的研究已经相对成熟，但两者融合的智能交互系统研究尚需进一步深入。将人脸识别与语音识别技术相结合，不仅能够提升智能交互系统的准确性和效率，还能够为用户带来更加便捷、个性化的体验。例如，在家庭智能助手、智能客服等场景中，通过人脸识别技术，系统可以识别用户的身份，并据此提供个性化的服务；而语音识别技术则允许用户通过语音指令与系统交互，使得操作更加简单直观。因此，加强人脸识别与语音识别技术的融合研究，对于推动智能交互系统的发展具有重要意义。

2.1. 基于人脸识别的智能交互系统

基于人脸识别技术的智能交互系统主要依靠的是人脸识别算法。人脸识别技术在光线不足、面部表情变化或佩戴遮挡物等复杂场景下，其识别准确率可能会受到严重影响。当前，人脸识别的方法可分为四种：基于几何特征的方法、基于模板匹配的方法、基于统计学习的方法及基于深度学习的方法。近年来，深度学习技术被广泛应用于人脸识别领域，并取得了显著的成果。基于深度学习的人脸识别算法主要包括卷积神经网络(CNN)、深度信念网络(Deep Belief Networks, DBN)等。由于本文中使用的是基于卷积神经网络(Convolutional Neural Network, CNN) [2]的深度学习方法，下面则着重介绍基于深度学习的卷积神经网络(CNN)和循环神经网络(RNN)的研究。这些方法通过训练大量的人脸数据，可以学习到人脸的深层次特征表示，从而提高人脸识别的准确性和鲁棒性。

2.1.1. 基于深度信念网络(Deep Belief Networks, DBN)的人脸识别技术

深度置信网络(DBN)，由受限玻尔兹曼机(RBM)堆叠而成，是深度学习领域的早期模型之一。在人脸识别领域，DBN 展现了其独特的优势。通过姿态映射和姿态分类，DBN 能够学习到侧面人脸图像到正面人脸图像的全局映射，并达到良好的分类性能。然而，直接使用人脸图像像素作为输入时，DBN 可能忽略人像的局部特征，并受到姿态、光线、噪声等因素的干扰。为了优化 DBN 在人脸识别中的性能，研究者们提出了多种方法。赵远东[3]结合了 Gabor 小波与 DBN，有效提取了人像的抽象特征，降低了姿态、光线等因素对识别率的影响，实现了对人像的准确识别。

2.1.2. 基于卷积神经网络(Convolutional Neural Network, CNN)的人脸识别技术

CNN 通过模拟人脑视觉皮层的层次结构，自动学习图像中的特征表示，无需复杂预处理，直接以图像像素作为输入，降低了数据重建的复杂度。通过多层卷积、激活函数和池化运算，CNN 能够学习到图像中的高级特征，实现准确的人脸识别。已有研究将 CNN 应用于人脸识别，并取得了优异的性能。经过对 CNN 模型的进一步优化，通过引入学习非线性特征变换的策略，该方法成功减小了类内变化，并确保了不同身份的人像间距保持稳定。这一改进显著提升了人脸识别性能，使得在 LFW 数据库上的识别率提

升至 99.15% [4]，不仅超越了原有记录，也超越了当前领先的人脸识别算法的性能。然而，CNN 对光照、姿态等变化敏感，且训练需要大量标注数据。未来研究可针对这些问题进行优化，以提高人脸识别技术的鲁棒性和实用性。

2.2. 基于语音识别的智能交互系统

背景噪声对语音识别技术的干扰尤为显著，因此语音识别当中的语音增强技术显得尤为重要。语音增强技术的核心在于从含噪语音中提取有用的语音信号，抑制或降低噪声干扰。语音增强算法可以分为三类：基于滤波器的方法、基于统计模型的方法及基于神经网络的方法[5]。近年来，深度学习技术逐渐被应用于语音增强领域。本文侧重介绍卷积神经网络模型的研究。

卷积神经网络模型

卷积神经网络是一种前馈神经网络，其人工神经元可以响应一部分覆盖范围内的周围单元，对于大型图像处理有出色表现。目前，卷积神经网络已经取得了许多令人瞩目的成果，但仍然存在一些挑战和问题。深度残差网络在语音增强领域表现突出。Chen Yinghao 等人设计多层网络[6]，通过预测补偿和特征映射解决梯度消失，增强模型稳定性与鲁棒性。Mohammed Bahoura 在 FPGA 上实现谱减法，为硬件加速提供新思路。魏磊则在 FPGA 上实现 CNN 语音增强，为未来研究提供参考[7]。李斌验证了硬件平台上神经网络在语音增强中的优越性能。这些研究推动了语音增强技术的发展。

2.3. 基于人脸识别与语音识别的智能交互系统

目前对于基于人脸识别与语音识别[8]的智能交互系统研究较少，两者融合的智能交互系统研究尚需进一步深入。本文选择将基于卷积神经网络的人脸识别技术与基于时频卷积网络(TFCN)的轻量级语音增强算法融合开发一种新的智能交互系统。TFCN 旨在预测并恢复纯净语音，保证在嘈杂环境下语音的清晰度和识别性。相比其他算法，TFCN 以 93,000 个参数即实现高性能，展现其高效与实用。抑制背景噪声，提高目标语音的清晰度和可识别性。并采用先进的特征点检测算法，以捕捉更准确的面部特征，提高面部识别的准确率。

3. 系统总体设计

对智能交互系统整体设计，为保证系统运行的稳定性、准确性和实时性，需要软硬件的联合运行。本设计系统以 openMV 和 HMI 智能串口屏作为控制器，来实现照片采集、人脸识别及显示内容的功能，openMV 通过 IDE 软件端实现照片采集及人脸识别功能，通过麦克风阵列对人声进行录音，保存语音文件来做语音增强和语音识别，再将识别的文字提取出来通过智能显示屏显示。利用该智能交互系统实现人脸识别及语音转文字输出，实现数据的准确显示、存储和反馈。系统总体设计框架如图 1 所示。

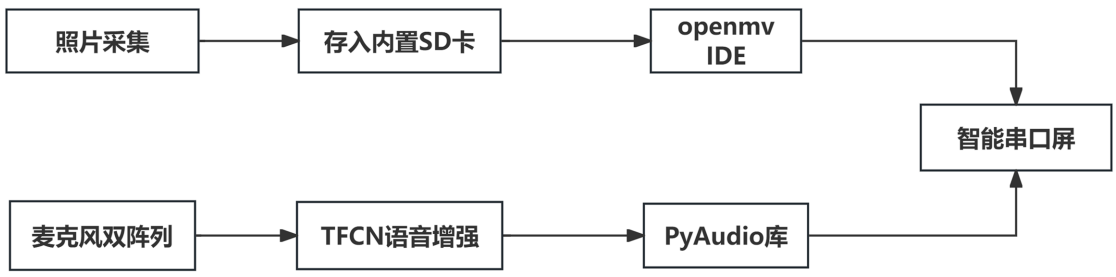

Figure 1. Block diagram of the overall system design
图 1. 系统总体设计框图

4. 系统软件功能设计

系统实现的功能主要有利用 openMV 和 TFCN 语音增强技术与 HMI 智能串口屏通信实现语音增强、语音识别、人脸识别功能，将 openMV 人脸识别的结果及语音转文字的结果通过 HMI 智能串口屏显示的功能。对此拟展开研究论述如下[9]。

4.1. 人脸识别功能

人脸识别特指利用分析比较人脸视觉特征信息进行身份鉴别的计算机技术。人脸识别模块是系统的重要组成部分，其设计目的是识别系统用户的面部特征，实现自然的用户认证和交互[10]。利用 openMV 官网提供的项目实例中的图像预处理、特征点检测及人脸区域定位的人脸识别模块，需要配合手动设计特征，进行特征提取机特征匹配，之后运用基于卷积神经网络(Convolutional Neural Network, CNN)的深度学习算法，完成对人脸的识别。同时需要在 openMV 上接入 HMI 智能串口屏以便显示说话者名字。程序设计为基于局部二值模式(LBP)的人脸识别程序[11]。

首先，OpenMV 从人脸数据库中加载一组预定义的人脸图像，并计算每个图像的 LBP 特征。然后，它将这些特征与当前捕获的人脸进行比较，计算特征差异度。最后，它输出最匹配的人脸的特征差异度和对应的说话者名字。

计算 lbp 特征

$$d1 = \text{img.find\_lbp}\left(\left(0,\ 0,\ \text{img.width}(\ ),\ \text{img.height}(\ )\right)\right) \tag{1}$$

#d1 为第 s 文件夹中的第 i 张图片的 lbp 特征

$$\text{dist} += \text{image.match\_descriptor}\left(d0,\ d1\right) \tag{2}$$

#计算 d0, d1 即样本图像与被检测人脸的特征差异度。

可知特征差异度越小，被检测人脸与此样本更相似更匹配。

4.2. 语音识别功能

4.2.1. 语音增强

在处理自然场景下的语音信号时，由于外界环境噪声、设备自身缺陷以及传输通道的限制等因素的影响，原始语音数据通常会受到一定程度的污染[12]。因此，在进一步分析之前，进行有效的降噪处理是至关重要的。单通道语音增强技术的目标在于最大限度地压制背景噪声，同时保留目标语音信号，以最小化信号失真。通过这种处理，不仅可以提高语音的感知质量和清晰度，而且对于自动语音识别系统来说，还能显著降低单词识别错误率。

为了显著提升语音文件的质量并提取出清晰无杂质的语音信号，本系统引入了一种前沿的语音增强技术，即基于[13]时频卷积网络(TFCN)的轻量级语音增强算法。这一创新方法巧妙地运用了优化设计的扩展卷积与深度可分离卷积结构，不仅显著降低了模型的复杂度，使得参数量控制在约 93,000 的较低水平，而且在性能上达到了与当前主流算法相抗衡的水平。

在语音增强模块。采用 TFCN 语音增强算法，先将输入的原始带噪语音信号经过分帧、加窗、FFT [14](快速傅里叶变换)的预处理，将时域信号转为频域信号，通过扩展卷积和深度可分离卷积结构来提取语音信号中的关键时域特征，对提取的时域特征进行学习和处理，之后基于学习到的特征对原始带噪语音信号进行增强，去除噪声，保留并增强语音信号。语音增强后，对信号进行 IFFT(逆快速傅里叶变换)，将频域信号转为时域信号，最后，输出增强后的去噪语音信号。图 2 是语音增强算法流程图。

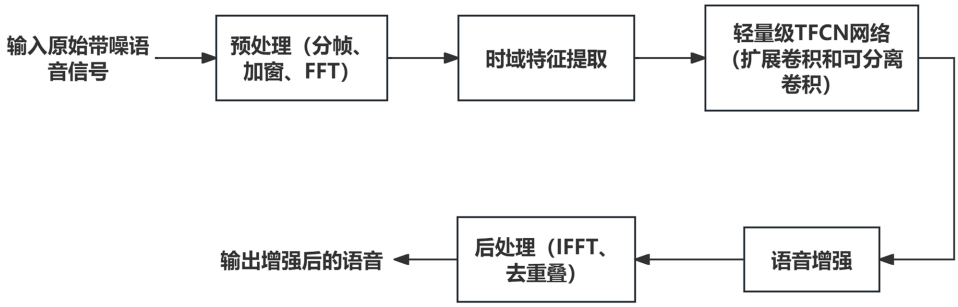

Figure 2. Speech enhancement algorithm flow
图 2. 语音增强算法流程

这种结合 TFCN 和二维卷积的结构设计，不仅保留了 TFCN 在时域建模上的优势，还通过二维卷积显著增强了其在频域上的处理能力，有效地捕捉时间和频率的复杂交互，从而实现了对语音信号的增强。在性能方面有着显著提升，可以有效去除噪声，同时保留语音信号中的关键信息，增加了语音信号的可读性与清晰度。

在该系统中，为了对语音增强模块的性能效果进行测试，利用了备受认可的 VCTK 数据集。VCTK 数据集包含了 28 位男女说话者的纯净语音样本及其相应的噪声干扰版本，提供了一个全面而多样的测试平台。为了确保测试的一致性和准确性，首先将所有数据统一重采样至 16 KHz。该数据集当中涵盖了从人工模拟的嘈杂声音到源自 Demand 数据库的真实环境噪声，进一步丰富了测试环境的复杂性。

在数据集的分配上，遵循了标准的机器学习流程，将数据划分为训练集、验证集和测试集。其中，训练集包含了 9567 个话语样本，用于模型的训练和优化；验证集则包含 1245 个样本，用于在训练过程中监控模型的性能，以避免过拟合现象。为了更贴近实际应用场景，选取了两位未参与训练的说话者，并结合 5 种典型的噪声类型和多个不同的信噪比水平，构建了独立的测试集[15]。

为了对语音增强模块的性能进行评估，选用了四种业内广泛认可的客观评价指标：STOI、PESQ、Csig 和 Cbak。这些指标涵盖了语音的可懂性、感知质量、信号失真程度、背景噪声干扰以及整体语音质量等多个方面，提供了全面而细致的评估视角。指标分数越高，代表该指标在性能方面效果较好。

通过这一系列的评估流程，这些评估结果不仅验证了算法的有效性和鲁棒性，还为后续的优化和改进提供了宝贵的参考。

4.2.2. 语音识别

语音识别将人类语音转化为机器可识别和理解的文本或命令信息，实现人与机器的无障碍交流。本智能交互系统调用了 Python 的 PyAudio 库。作为音频处理领域的核心工具，PyAudio 库不仅具备音频录制、保存和播放的基础功能，更拥有实时处理音频数据的强大能力，从而确保系统功能的完善与高效。

为了确保用户语音指令的精准捕捉，本系统特别采用了双阵列麦克风设计进行录音。这种先进的录音技术不仅显著提升了录音的清晰度，使得语音指令更加清晰可辨，而且在复杂环境中也能有效过滤背景噪音，确保指令的准确识别。此外，PyAudio 库还提供了丰富的参数调整选项，用户可以根据实际需求灵活设置采样率、位深度和声道数，从而适应不同的应用场景，满足多样化的需求。此外，PyAudio 库还引入了回调函数和事件驱动机制，为系统带来了更高的交互灵活性和响应速度。

在环境适宜，周围环境噪音弱的情况下，本智能交互系统能够迅速而准确地将用户的口头指令转换为可执行的文本命令。这一功能的实现，不仅简化了用户的操作过程，提高了系统的易用性，还进一步展现了 PyAudio 库在音频处理方面的卓越性能。通过充分利用 PyAudio 库的各项功能，本系统实现了高效、准确的音频处理与用户指令识别[16]。

5. 系统硬件设计

为了实现对说话者名字和语音识别的结果的显示，本设计系统选用 HMI 智能串口屏作为显示设备。借助 USART HMI 软件设计平台，设计程序实现输出说话者名字和说话内容。设计界面展示如图 3。实物效果如图 4 所示。

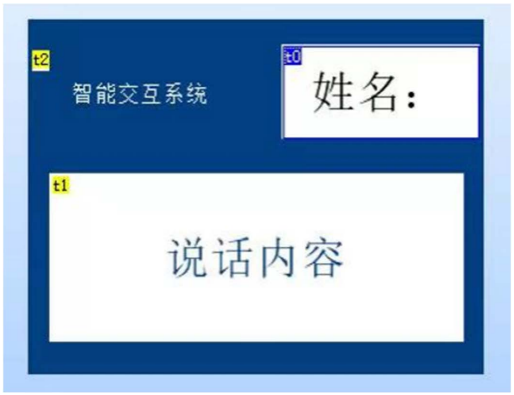

Figure 3. HMI smart serial screen design interface
图 3. HMI 智能串口屏设计界面

Figure 4. Intelligent serial screen physical effect
图 4. 智能串口屏实物效果

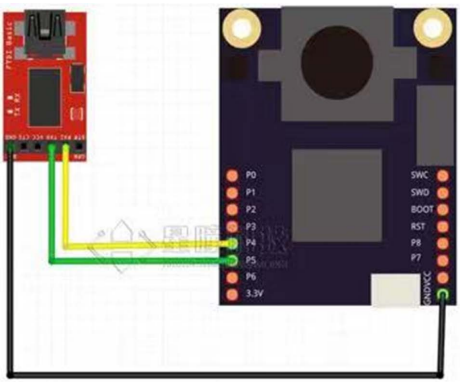

Figure 5. Connecting line diagram
图 5. 连线示意图

人脸识别功能使用 openMV 作为视觉传感器，电脑设备提供的扬声器作为录音设备。openMV 视觉传感器与 HMI 智能串口屏之间使用串口通信进行连接，连线示意图如图 5 所示。openMV 有两个电源输入端，VIN 输入为 3.6 V~5 V，推荐 5 V，USB 和 VIN 可以同时供电。openMV 视觉传感器使用 openMV IDE 软件端编写人脸识别程序，openMV 视觉传感器示意如图 6 所示。UART (Universal Asynchronous Receiver/Transmitter，通用异步收发器)是一种常用的串行通信协议，广泛应用于单片机或各种嵌入式设备之间的通信。使用 UART 协议进行串口通信需要设定工作模式、波特率。本实验中使用串口 3，波特率为 9600。

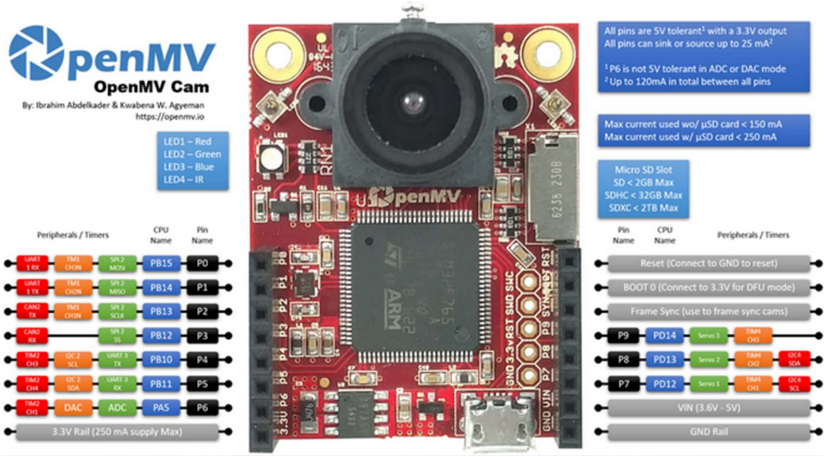

Figure 6. Schematic diagram of openMV vision sensor
图 6. openMV 视觉传感器示意图

5.1. openMV 视觉传感器的选择

在众多的视觉传感器模块中，openMV 是一种基于 Python 的低成本、高性能视觉传感器模块，具有许多优点，因此被选择作为本系统的视觉传感器模块。首先，openMV 的设计简单，易于使用，可以快速地进行开发和测试。其次，openMV 具有高分辨率、高帧速率和多种图像处理功能，可以满足人脸识别的需求。此外，openMV 还具有丰富的软件库和开源代码，可以方便地进行二次开发和定制化。在本系统中，openMV 的高性能和易用性使其成为了一个理想的选择。同时，openMV 还具有较小的尺寸和重量，可以方便地嵌入到智能交互设备中，使得其在实际应用中具有广泛的适用性。

5.2. HMI 智能串口屏的选择

HMI 智能串口屏具有高度集成、易于使用、可定制化等优点，可以方便地实现图形化界面的设计和显示输出。同时，HMI 智能串口屏还具有多种通信接口和协议支持，可以与各种主控芯片和嵌入式系统进行无缝连接，实现高效的数据传输和显示输出。

6. 系统实现及功能介绍

软件开发平台设计了采用 openMVIDE 平台开发、软件 USART HMI 和 Pycharm 联合开发模式。其中，openMVIDE 平台与 openMV 硬件配合实现人脸照片收集存至内置 SD 卡和人脸识别，并将结果输出至 HMI 智能串口屏，将结果可视化输出。Pycharm 软件负责接入麦克风输入的语音文件，实现语音输入、TFCN 语音增强及语音转文字功能。HMI 智能串口屏实现对说话者名字、说话者说话内容的输出显示。程序总体设计流程如图 7 所示。

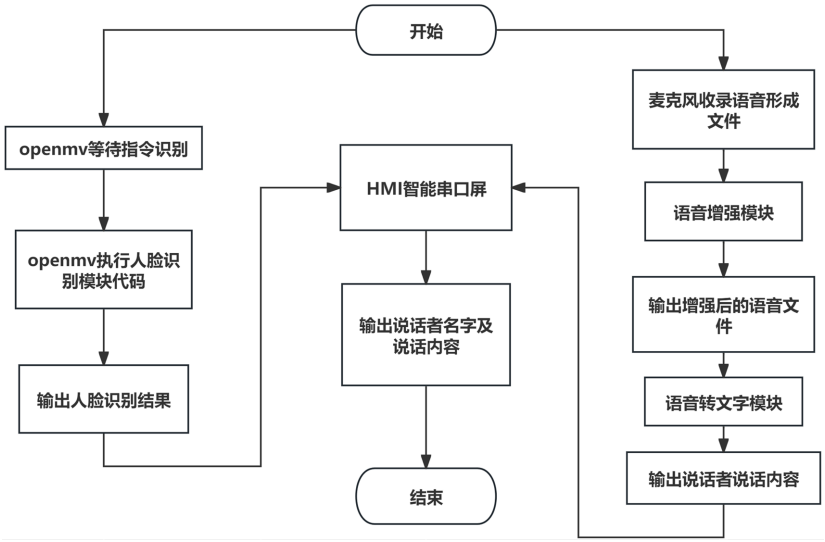

Figure 7. Flowchart of the general design of the programme
图 7. 程序总体设计流程图

人脸识别程序设计

从人脸数据库中加载预定义的人脸图像，并计算每个图像的 LBP 特征。然后，它将捕获当前的人脸，计算 LBP 特征，并与数据库中的特征进行比较。最后，它将输出匹配的人脸名称。设计流程如图 8 所示。

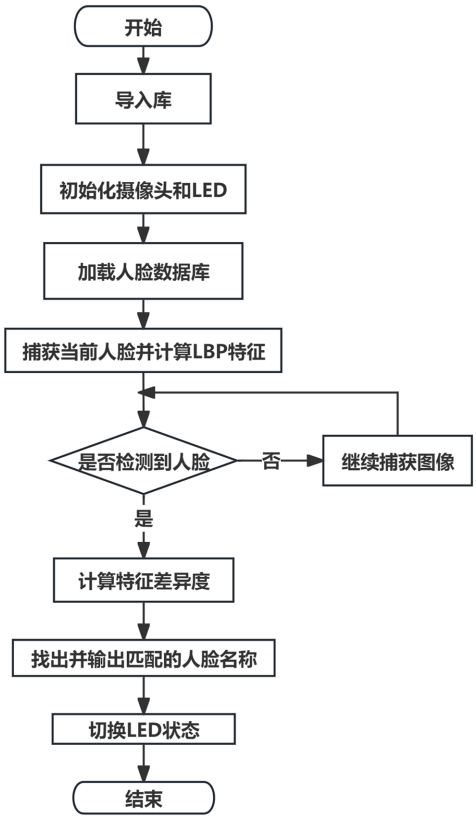

Figure 8. Flowchart of face recognition procedure
图 8. 人脸识别程序流程图

7. 测试方案与测试结果

7.1. 测试方案

1) 硬件测试阶段：检验系统的机械组件，确保其在预定的操作范围内能够顺畅运行。同时，电路焊接的细致审核确保了所有模块间的连接精准无误。

2) 软件测试阶段：对系统程序进行全面且细致的验证，以保证其能够稳定、无异常地执行。

3) 软硬件集成测试：将经过细致审查的程序烧录至硬件中，并在离线状态下进行了全面的运行测试。这一阶段的重点在于评估人脸识别与语音识别两大核心模块在实际应用中的表现是否精准可靠。

7.2. 测试结果与分析

7.2.1. 测试结果

1) openMV 首先从人脸数据库中加载了一组预设的人脸图像样本，如图 9 所示。

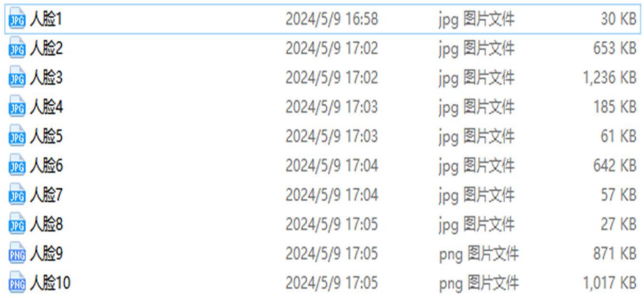

Figure 9. Sample face images
图 9. 人脸图像样本

随后，系统计算了这些样本图像的局部二值模式(LBP)特征，并将这些特征用于与实时捕获的人脸图像进行对比分析。在比较过程中，系统通过计算特征差异度来评估两者之间的相似性或差异性。在人脸识别领域，特征差异度是衡量从图像中提取的特征与数据库中存储特征之间不一致程度的指标。通常，较小的特征差异度表示较高的匹配度和识别准确率。测试结果如表 1 所示。

Table 1. Face recognition test result data
表 1. 人脸识别测试结果数据

原始 LBP 特征	识别后的平均特征差异度
[7, 102, 4, 244, 1, 177, 115, 24]	0.012006552
[133, 210, 236, 239, 81, 222, 157, 169]	0.147933194
[153, 63, 33, 212, 131, 57, 217, 22]	0.146742407
[95, 93, 19, 54, 218, 199, 75, 241]	0.190116911
[138, 180, 185, 119, 56, 254, 30, 103]	0.015203617
[151, 96, 25, 93, 105, 1, 104, 32]	0.112193672
[49, 189, 221, 89, 254, 101, 236, 63]	0.026721764
[121, 83, 226, 1, 225, 80, 223, 248]	0.129160915
[77, 112, 254, 138, 172, 52, 206, 191]	0.058295348
[84, 86, 64, 203, 45, 55, 92, 167]	0.076124257

2) 在评估语音识别的精确性时，采用 VCTK 数据集，选取三个指标：置信度评分、识别速率及识别精确率。置信度评分作为一个量化指标，为每次识别结果赋予了一个可信度值，使得能够直观地评估识别结果的质量，其值在 0 到 1 之间。此外，识别速率和识别精确率提供了系统性能的统计概览，识别精确率值在 0 到 1 之间。这些数据可用于评估语音识别系统的效能。选取 VCTK 数据集上的 12 条语音测试，系统在 VCTK 数据集上的基准测试结果如表 2 所示。

Table 2. Speech recognition test result data
表 2. 语音识别测试结果数据

置信度得分(0~1)	识别速度(s)	识别准确率(0~1)
0.934383571	0.982673254	0.949104326
0.894109223	0.721796058	0.930002915
0.865096015	1.322629627	0.977962824
0.952170376	0.724573153	0.942898111
0.880853832	0.831845492	0.945628762
0.918547499	1.13618196	0.940541809
0.816881479	0.531539082	0.903656381
0.829455433	1.307162358	0.972608562
0.898420586	0.759491213	0.960806306
0.820603031	1.346085313	0.965269088
0.845003902	0.55793521	0.955581055
0.817696095	0.750305922	0.959613345

3) 多模态融合验证实验：为了验证人脸识别与语音识别技术融合的有效性，比较仅使用单一模态(仅使用人脸识别或仅使用语音识别)与多模态融合后的系统响应时间。

Table 3. Multi-modal fusion validation experiment data
表 3. 多模态融合验证实验数据

响应时间	仅使用人脸识别(s)	仅使用语音识别(s)	人脸识别与语音识别技术融合(s)
1	0.365	0.946	0.984
2	0.432	0.932	1.032
3	0.386	0.956	0.968
4	0.378	0.948	1.056
5	0.412	0.943	0.978

经过对表 3 中数据的对比，观察到技术融合后的系统响应时间相较于单一模态下的系统响应时间略有增加。然而，进一步分计算后发现，多模态融合系统的平均响应时间仅比单纯使用语音识别的系统多 58 毫秒，这一差异在用户体验和系统性能层面均被视为可接受的范围内。

7.2.2. 测试结果分析

经过测试，该智能交互系统在人脸识别和语音识别两方面均展现出了显著优势。在人脸识别方面，系统呈现出极低的平均特征差异度，低于 0.1，这显著彰显了其高度的匹配性和识别准确性。同时，在语

音识别方面，系统的置信度平均得分高达 0.87，这一高水平评分验证了其识别结果的高度可靠性。此外，该系统在识别速度上也表现出色，能够在约 1 秒内迅速完成语音识别，展现了其高效的性能。更为重要的是，系统对于含噪语音的平均识别准确率高达 95%，充分证明了其出色的降噪能力和识别效果。经多模态融合验证测试，由结果可知该系统的有效性。综上所述，该智能交互系统性能得到了有效保障，展现出了其强大的实用价值。

8. 结束语

现有的智能交互系统单一运用人脸识别与语音识别技术，且在处理噪声干扰及复杂场景下的人脸识别时仍面临挑战，本文因此介绍了一种基于深度融合人脸识别和语音识别技术的智能交互系统，人脸识别模块基于卷积神经网络，语音增强模块通过运用基于时频卷积网络(TFCN)的轻量级语音增强算法，抑制背景噪声，该系统可以实现更加智能化、便捷化的人机交互方式。系统通过 openMV 摄像头在 openMV IDE 软件平台上运行人脸识别算法，实现精准的人脸识别功能。同时，配备的麦克风模块能够捕捉用户语音信息，经过 TFCN 语音增强模块的优化处理，显著提升了语音信号的清晰度和质量，进而增强了语音识别模块的识别准确率和系统的整体性能。为语音识别模块提供了更优质的输入数据。这一处理过程使得人与硬件设备之间的语音交互更加高效、准确。该技术的应用为智能交互系统在实际应用中提供了更为可靠和优质的语音交互体验。

此外，系统还集成了 HMI 智能串口屏，能够实时显示说话者的姓名和所说内容。这一设计使得整个系统的操作更加直观、便捷，用户能够在短时间内轻松掌握使用方法。同时，该系统功耗低，可通过移动电源供电，安装过程简便快捷，为用户快速搭建智能交互系统提供了极大的便利。

在智能家居领域，该系统具有广泛的应用前景。具体来说，在智能门锁应用中，系统能够自动识别家庭成员的面孔，实现无钥匙进入的便捷体验。同时，通过语音识别模块，系统能够识别特定的语音命令，如“开门”或“关门”，从而进一步增强了智能门锁的安全性和便捷性。实验结果显示，该系统在人脸识别和语音识别方面均具有较高的精度，展现了良好的实用性和应用前景。

展望未来，将继续深入探索智能交互系统在不同场景下的应用，并研究更多先进的语音增强方法以提升系统性能。随着技术的不断进步，坚信智能交互系统将在更多领域展现出巨大的应用潜力，为人们的生活带来更加便捷、智能的体验。

参考文献

[1] 黄玲, 王霄, 邵健, 胡娟, 张译. 基于 NodeMCU 智能语音交互家居系统设计[J]. 智能计算机与应用, 2021, 11(2): 164-168, 173.

[2] 李玲俐. 基于深度学习理论的人脸识别技术应用综述[J]. 计算机与数字工程, 2021, 49(9): 1912-1914, 1929.

[3] 赵远东. 基于深度神经网络的人脸识别方法研究[D]: [硕士学位论文]. 沈阳: 沈阳工业大学, 2019.

[4] 闫新宝, 蒋正锋. 基于 VGGNet 深度卷积神经网络的人脸识别方法研究[J]. 电脑知识与技术, 2023, 19(25): 34-37. https://doi.org/10.14004/j.cnki.ckt.2023.1370

[5] 杨涛. 基于机器学习的语音增强技术[J]. 电声技术, 2024, 48(3): 39-41. https://doi.org/10.16311/j.audioe.2024.03.013

[6] 黄修正. 基于可编程 SoC 的卷积神经网络数字解调信号语音增强[D]: [硕士学位论文]. 北京: 北京交通大学, 2023. https://doi.org/10.26944/d.cnki.gbfju.2023.000575

[7] 魏磊. 基于 CNN 的语音增强算法的研究与 FPGA 实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2021.

[8] 屈瑾. 基于语音识别的智能交互系统设计[J]. 自动化与仪器仪表, 2023(1): 221-225. https://doi.org/10.14016/j.cnki.1001-9227.2023.01.221

[9] 刘搏飞, 刘春池, 邢晓鹏, 隋盛誉, 孙嘉成, 李广凯, 谢印庆. 基于人工智能与物联网技术的家居门禁系统[J]. 物联网技术, 2022, 12(9): 117-118, 121.

[10] 廖玥灵, 马敏耀, 令狐蓉, 等. 基于面部识别的新型智能门禁系统设计与实现[J]. 无线互联科技, 2022, 19(20): 49-51.

[11] 赵慧, 张伟, 郝喆. 基于 OpenMV 视觉模块的人脸识别监控系统研究[J]. 信息化研究, 2022, 48(1): 55-58.

[12] 王宝妮, 包艳艳, 倪子越. 基于 STM32 的语音信号处理与传输技术研究[J]. 产业创新研究, 2024(6): 112-114.

[13] 相增辉, 张国梁, 庞渊源, 等. 基于深度卷积神经网络的智能机器人语音自动识别方法[J]. 自动化技术与应用, 2024, 43(4): 43-46. https://doi.org/10.20033/j.1003-7241.(2024)04-0043-04

[14] 孙思雨, 张海剑, 陈佳佳. 基于傅里叶卷积的多通道语音增强[J]. 无线电工程, 2024, 54(3): 580-588. http://kns.cnki.net/kcms/detail/13.1097.TN.20230901.1943.018.html

[15] Jia, X. and Li, D. (2022) TFCN: Temporal-Frequential Convolutional Network for Single-Channel Speech Enhancement. arXiv: 2201.00480. https://doi.org/10.48550/arXiv.2201.00480

[16] 王臣. 基于深度学习的人脸识别方法的探究[J]. 数字通信世界, 2020(7): 169-170.