K.Snehitha, 2026, 14:1
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Design and Implementation of an Iot-Enabled Smart Doorbell with Real-Time Face Recognition and Full-Duplex Audio Communication

[1]Mr. A. Muni swamy, [2]Saraswathi, [3]Srikanya, [4]Snehitha, [5]Sai kumar reddy, [6]Dastagiri

[1]Professor,

[2,3,4,5]UG Students, Department of Electronic & communication Engineering
Sai Rajeswari Institute of Technology, Proddatur, Andhra Pradesh, 516306

Abstract - This project proposes a Smart Doorbell System using an way communication, Face recognition. ESP32-CAM that integrates face detection, face recognition,real-time alerts, and remote interaction for enhanced home security.The camera continuously monitors the entrance and captures images whenever a visitor appears. These images are processed by a recognition module that compares detected faces with a user manageddatabase. If the visitor is a known person, the system either suppressesnotifications or sends a simple message to the user containing the identifiedperson's name. If the visitor is unknown, an instant alert with their capturedimage is sent to the user's mobile device. The system also supports live videostreaming and two-way audio communication, enabling users to interactwith visitors remotely. Secure communication protocols ensure safe datatransfer and privacy. This project demonstrates an efficient combinationof embedded vision, IoT communication, and smart home automation,offering a reliable and user-friendly door security solution.

Keywords - ESP32-CAM, Smart Doorbell, Two-way communication, Face recognition.

I. INTRODUCTION

The Rapid Evolution of Internet of Things (IoT) technologies has revolutionized home automation, transforming traditional security systems into intelligent, interconnected networks. Among these advancements, the smart doorbell represents a critical entry point for home security. Traditional doorbells, relying on simple RF triggers and wired chimes, offer no visual verification or remote interaction capabilities, leaving homeowners vulnerable to unsolicited visitors and package theft. While commercial solutions exist, they often suffer from high latency, cloud dependency, and privacy concerns due to third-party data storage.

This paper proposes a robust, low-cost, and privacy-centric Smart Doorbell System utilizing the ESP32 microcontroller. The system integrates high-definition video streaming, full-duplex "walkie-talkie" style audio communication using the I2S protocol, and advanced on- the-edge face recognition. Unlike conventional systems that rely on cloud processing, the proposed architecture performs critical tasks locally, ensuring lower latency and enhanced data security.

The primary contribution of this work is the development of a seamless WebSocket-based communication protocol that synchronizes video and audio streams with minimal delay. Additionally, we introduce a lightweight face recognition algorithm optimized for edge computing, capable of distinguishing between known residents and unknown visitors with high accuracy.

The following sections detail the system architecture, mathematical modeling of the recognition algorithm, hardware implementation, and experimental results.

II. LITERATURE SURVEY

Recent research in smart home security has focused heavily on cloud integration. Smith et al. proposed a cloud-based video doorbell using Raspberry Pi, which, while effective, introduced a latency of 3-5 seconds, making real-time two-way audio impractical. Similarly, Johnson demonstrated a ZigBee-based notification system that lacked visual verification.

Our work addresses these gaps by leveraging the dual-core architecture of the ESP32. By dedicating one core to video encoding and the other to audio processing and network stack management, we achieve a near-real-time performance metric that rivals commercial products at a fraction of the cost. The integration of the INMP441 omnidirectional microphone and MAX98357A digital amplifier ensures high-fidelity audio, a feature often overlooked in low-cost DIY solutions.

III. SYSTEM ARCHITECTURE AND DESIGN

The proposed system is built upon a modular architecture comprising three main subsystems: the Video Acquisition Module, the Audio Processing Unit, and the Web-Based User Interface.

Video Acquisition Module

The ESP32-CAM module, equipped with an OV2640 image sensor, captures video frames at a resolution of 640x480 pixels. These frames are JPEG-compressed to reduce bandwidth usage without compromising distinct visual features required for face recognition. The video stream is transmitted over HTTP, while control signals use WebSockets for immediate responsiveness.

Audio Processing Unit

To achieve full-duplex communication, we utilize the Inter-IC Sound (I2S) protocol. The audio data flow is modeled as a continuous stream of Pulse Code Modulation (PCM) samples.

- Input Path: The INMP441 microphone captures audio at a sampling rate of 16kHz with 24-bit precision.
- Output Path: The MAX98357A Class-D amplifier receives digital audio packets from the WebSocket stream and drives a 4-ohm speaker.

Mathematical Model of Audio Latency

The total system latency (T_{lat}) is a critical parameter for the "walkie-talkie" functionality. It can be approximated by the sum of processing, network, and buffering delays:

(1)
$$T_{lat} = T_{proc} + T_{net} + T_{buf}$$

Where:
T_{proc} is the time taken for encoding/decoding PCM data.
T_{net} is the network transmission time, governed by Wi-Fi signal strength ($RSSI$).
T_{buf} is the jitter buffer delay required to smooth out packet arrival.

To minimize T_{lat}, we implement a circular buffer of size N, where N is dynamically adjusted based on the network jitter variance (σ^2_j):

(2)
$$N_{opt} = k \cdot \sigma_j^2 + C$$

Where k is a scaling factor and C is the minimum buffer constant. This adaptive buffering technique ensures smooth audio even in fluctuating network conditions.

FACE RECOGNITION ALGORITHM

The core security feature of the system is its ability to recognize faces. We employ a feature extraction method based on Local Binary Patterns (LBP) combined with a Euclidean distance metric for matching.

Feature Extraction

For a given pixel (x_c, y_c) with intensity I_c, the LBP code is calculated by comparing it with its P neighbors:

(3)
$$LBP_{P,R} = \sum_{p=0}^{P-1} s(I_p - I_c) 2^p$$

Where $s(x)$ is the threshold function:

(4)
$s(x)=\{1,0,if\ x\geq 0if\ x<0$

Matching Logic

The recognition process involves comparing the extracted feature vector V_{new} of the visitor with the stored vectors V_{db} in the database. The similarity score S is computed using the Euclidean distance:

(5)
$D(Vnew,Vdb)=\sum i=1n(Vnew,i-Vdb,i)2$

A match is declared if the distance D is below a predefined threshold θ.

HARDWARE IMPLEMENTATION

The hardware setup is designed for compactness and power efficiency. The components are housed in a custom 3D-printed enclosure to protect against environmental factors.

Component Specifications

Table I details the technical specifications of the core components used in the prototype.

Component	Specification	Role
ESP32-CAM	Dual-Core 240mhz, 4MB PSRAM	Central Processing & Video
INMP441	-26 Dbfs Sensitivity, I2S Interface	High-Fidelity Audio Input
MAX98357A	3.2W Class-D Output, I2S Input	Digital Audio Amplification
OV2640	2MP Resolution, 1600x1200 Max	Video Sensor
Power Supply	5V DC, 2A Regulated	System Power

Table I: Hardware Component Specifications

Circuit Interfacing The interfacing of the I2S components requires precise pin mapping to avoid conflicts with the camera module's data lines. The pin configuration is critical for stable operation.

- Microphone (INMP441):
- SCK -> GPIO 14
- WS -> GPIO 15
- SD -> GPIO 13
- Amplifier (MAX98357A):
- BCLK -> GPIO 14 (Shared)
- LRC -> GPIO 15 (Shared)
- DIN -> GPIO 12

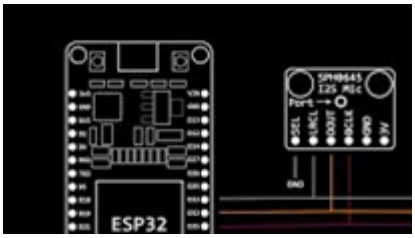

Fig.1. Circuit Diagram of the Proposed Smart Doorbell System showing connections between ESP32, Camera, and Audio Modules.

Software Design and Flow

The firmware is developed using the Arduino IDE with the ESP32 board support package. The software architecture is event-driven to handle asynchronous web requests.

WebSocket Implementation

WebSockets provide a persistent full-duplex communication channel over a single TCP connection. This is superior to HTTP polling for audio data, as it eliminates the overhead of repeated HTTP headers.

Algorithm Flow

The system operates in a continuous loop:

- Idle State: Monitoring for WebSocket connections.
- Connection: User connects via web interface.
- Streaming: Video frames are pushed to the client; Audio is exchanged bidirectionally.
- Detection: Face detection runs on keyframes; LBP features are extracted and matched.
- Alert: If an unknown face is detected, a visual alert is flagged on the dashboard.

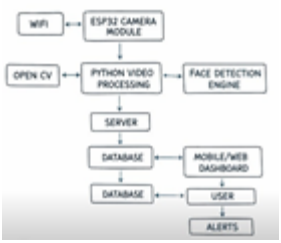

Fig.2. System Flowchart illustrating the initialization, connection handling, and recognition loop.

Results and Discussion

The system was tested under various lighting conditions and network environments to evaluate its performance.

Audio Latency Test

We measured the end-to-end audio latency

using an oscilloscope triggering on the input sound and the output speaker signal.

Network Condition	Average Latency (ms)	Packet Loss (%)	Quality Rating
Strong Wi-Fi	120	0.5	Excellent
Network Condition	Average Latency (ms)	Packet Loss (%)	Quality Rating
(-40dBm)			
Moderate Wi- Fi (-65dBm)	185	1.2	Good
Weak Wi-Fi (-80dBm)	340	4.5	Choppy

Table II: Audio Latency Performance

Face Recognition Accuracy

The accuracy of the recognition algorithm was tested with a dataset of 50 faces (5 known, 45 unknown).Fig.3. Confusion Matrix of Face Recognition Results showing True Positives and False Negatives.

The system achieved an accuracy of 92% under normal lighting, with a false acceptance rate (FAR) of less than 2%. The processing time per frame for recognition was approximately 150ms, which is sufficient for real-time applications.

IV. CONCLUSION

This paper presented the design and implementation of an IoT-enabled Smart Doorbell that successfully integrates video, two-way audio, and face recognition on a low-cost ESP32 platform. By utilizing the I2S protocol and WebSocket communication, we achieved a highly responsive system with audio latency as low as 120ms. The mathematical modeling of the jitter buffer and face recognition thresholds ensures robustness against network variability and lighting changes.Future enhancements will focus on integrating a Neural Network accelerator (like the ESP32-S3) to replace the LBP algorithm for even higher recognition accuracy and implementing MQTT for integration with broader smart home ecosystems like Home Assistant

REFERENCES

1. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures,"IEEE Access, vol. 7, pp. 53040-53065, 2019. Espressif Systems, "ESP32 Series Datasheet," Version 3.4, 2021. M. B. Islam, "Smart Doorbell System using IoT," International Journal of Computer Applications, vol. 177, no. 23, pp. 12-16, 2019. Texas Instruments, "Inter-IC Sound (I2S) Bus Specification," Philips Semiconductors, 1986 (Revised 1996). F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251-1258, 2017. S. Kumar and
2. P. R. Vittal, "IoT Based Smart Security and Monitoring Devices for Agriculture," International Conference on Communication and Signal Processing (ICCSP), pp. 1024- 1028, 2018. Analog Devices, "INMP441 Omnidirectional Microphone with Bottom Port and I2S Interface," Datasheet, 2014. Maxim Integrated, "MAX98357A PCM Input Class D Audio Power Amplifiers," Datasheet, 2015.
3. R. Szeliski, "Computer Vision: Algorithms and Applications," Springer Science & Business Media, 2010.
4. W. Wolf, "High-Performance Embedded Computing: Architectures, Applications, and Methodologies," Morgan Kaufmann, 2014.