DESIGN DOCUMENT: AI-Powered Form Filling Assistant

Project Title: Indian Citizen Services Automation

Deliverable: Technical Design & Architecture Documentation

---

1. Project Overview

This project addresses the manual effort and error-prone nature of filling government service forms at Seva Kendras. It utilizes an AI-driven pipeline to auto-populate form templates using data extracted from uploaded identity documents such as Aadhaar, PAN, Voter ID, and Driving Licenses.

+1

2. Proposed Architecture

The system follows a modular pipeline designed for high accuracy and low latency:

1. **Ingestion Layer**: Supports PDF and image uploads, as well as real-time voice-based input for field filling.

   +4

2. **Processing Layer (OCR)**: Scans documents using Tesseract with support for multiple Indian languages (English, Hindi, Marathi).

   +3

3. **Entity Extraction Engine**: Uses Regex and NLP to identify and isolate specific fields like Name, DOB, and Address.

   +3

4. **Mapping & Output**: Maps extracted data to standardized templates and generates a downloadable "Verified Citizen" PDF.

   +3

3. Technical Implementation: The Lookaround Regex

A critical challenge in extracting Indian IDs (especially Aadhaar) is the **Virtual ID (VID) Problem**. OCR often captures footer metadata (like the 16-digit VID) as part of the address string.

- **Solution**: We implemented **Positive/Negative Lookaround Regex**.
- **Logic**: The extraction engine uses lookarounds to set a "hard stop" at the address field. It searches for keywords like "Mobile," "Phone," "Help," or "www" and stops capturing text immediately before these markers, ensuring the VID or footer text is never included in the address field.

4. Key Features

- **Multilingual Support**: Optimized for diverse Indian document scripts.
- **Data Review**: A user-centric interface allows for manual editing and verification before final PDF generation.

  +1

- **Voice Correction**: Integrated microphone support for accessibility-focused data entry.

5. Performance Benchmarks

The system is tuned to meet the following hardware and performance targets:

- **Accuracy**: >90% precision in entity extraction.
- **Latency**: ≤ 3-5 seconds per document processing cycle on Intel hardware.

6. Roadmap & Stretch Goals

- **Handwritten Recognition**: Future integration to handle hand-filled forms.
- **API Integration**: Direct document fetching via DigiLocker or official government portals.