04 Optical Character Recognition (OCR) for Text Extraction and Verification

Innovate an OCR-driven solution that seamlessly extracts text from scanned documents, intelligently auto-fills digital forms, and accurately verifies the extracted data against the original source for enhanced reliability and efficiency.

Complexity Level: Medium

Overview:

Manually extracting and verifying data from physical documents is not only tedious but also increases the risk of errors and inconsistencies. Build a solution that leverages **Optical Character Recognition (OCR)** technology to automatically extract text from scanned documents or images and populate digital forms with precision. In addition to text extraction, the solution should include a verification layer that cross-checks the filled data with the source document to ensure accuracy and integrity. It should support diverse document types—such as printed forms, ID cards, and handwritten notes—and offer an intuitive interface that streamlines the workflow, enhances data reliability, and minimises manual intervention.

Recommended Tech Stack:

- Technologies: TrOCR (by Microsoft), JavaScript, HTML/CSS, RESTful APIs
- Data Sets: Publicly available datasets of scanned documents, synthetic data for validation and testing
- Use Open source libraries
- No cloud services should be used

Exact Task

Mandatory Tasks:

Design two separate APIs – one for text extraction and one for data verification – using OCR technology to automate form-filling and ensure data accuracy by comparing the extracted text with given inputs.

1. **API 1: OCR Extraction API:**
    - Accepts a scanned PDF/image.

- Uses OCR to extract relevant fields and populate digital form fields.

- Must support one Latin-based language i.e. English
- Must work with the sample documents provided (e.g., ID card, form, certificate).

2. **API 2: Data Verification API**

    - Accepts form-filled data and the original scanned document.

    - Compares extracted values against submitted data and highlights mismatches or errors.

    - Returns a confidence score or match status for each field.

    Good-to-Have Tasks:

    - Multi-lingual support: Support one non-Latin based language like Arabic.

    - Implement an interface or demo form that consumes the two APIs and showcases extraction + verification.

    - Add support for handwritten text recognition.

    - Enable partial data mapping, where not all fields are expected, but the system extracts available ones.

    - Allow users to manually correct OCR errors before verification.

    - Supports one non-Latin based language like Hindi or Arabic.

    Bonus Tasks:

    - Integrate with MOSIP Modules:
        - Connect the APIs with Pre-registration, Registration Client, or Android Registration Client to show end-to-end flow.
        - Display a capture quality score (e.g., blur detection, lighting quality) after every scan to help users retake if needed.
        - Enable multi-page document support (e.g., application forms spanning 2+ pages).
        - Offer real-time feedback indicating OCR confidence zones on the scanned document.

Deliverables:

- Source code of the solution on GitHub/GitLab.
    - A demonstration video showcasing the solution in action.
    - PPT with the approach, results, and potential impact.
    - Documentation:
        - Detailed workflow documentation on the design and implementation approach of the tool such as Architectural Design, Data Flow Structure of the tool, Integration and installation guidance
        - API documentation
        - Test case/ scenarios and test data (if applicable)

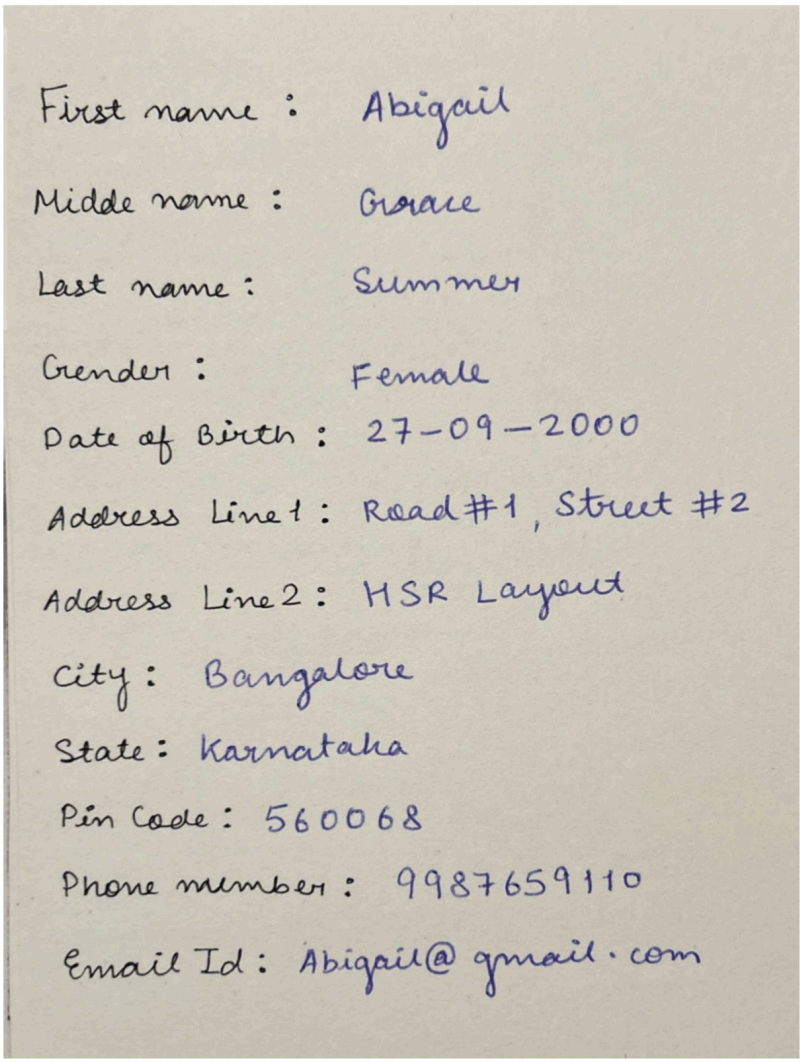
Resources:

1. MOSIP Sandbox environment: https://collab.mosip.net/
2. UIN request form: https://self-register.collab.mosip.net/
3. Collab user Guide link: https://docs.mosip.io/1.2.0/collab-getting-started-guide
4. Resident Services APIs: https://mosip.stoplight.io/docs/resident/au98eyyz-document
5. Sample registration forms:

Registration Form

Name	Ananya Sharma
Age	29
Gender	Female
Address	123, MG Road, Bengaluru, Karnataka - 560001
Email Id	ananya.sharma@example.com
Phone number	+91-9876543210

Name: John Smith
Age: 30
Gender: Male
Address: 123 Elm St
Country: USA
Phone number: 555-12345
Email: john.smith@example.com


First name : Abigail
Midde name : Grace
Last name : Summer
Grender : Female
Date of Birth : 27-09-2000
Address Line1 : Road #1, Street #2
Address Line2 : HSR Layout
City : Bangalore
State : Karnataka
Pin Code : 560068
Phone number : 9987659110
Email Id : Abigail@gmail.com