OCR System Architecture for Automated Invoice Processing

EURON

OCR System Architecture for Automated Invoice Processing

One common and impactful use case for OCR is **automated invoice processing**, where invoices from different vendors are scanned, digitized, and stored in a structured database for further analytics.

◆ Architecture Components and Flow

1. **Document Ingestion**
    - **Sources**: PDFs, Scanned Images, Emails, Mobile App Uploads, FTP/S3 Buckets
    - **Tools**: Apache NiFi / AWS Lambda / Azure Logic Apps
    - **Preprocessing**: Image enhancement, noise reduction (OpenCV, PIL, Tesseract)
2. **OCR Engine**
    - **Text Extraction**: Google Tesseract / AWS Textract / Azure OCR / Google Vision API
    - **Handwriting Recognition** (if required): TensorFlow/Keras-based CNN-LSTM model
    - **Table Detection & Structure Extraction**: LayoutLM / Detectron2 for structured invoices
3. **Post-OCR Processing & Validation**
    - **Regex-based or NLP-based Text Cleaning** (Removing unwanted text, special characters)
    - **Field Mapping & Entity Recognition**: NER (Named Entity Recognition) using SpaCy/BERT
    - **Spell Check & Correction**: SymSpell / BERT-based sequence correction
4. **Database & Storage**
    - **NoSQL Database (MongoDB/Elasticsearch) for fast search & indexing**
    - **SQL Database (PostgreSQL/MySQL) for structured storage**
    - **Data Lake (AWS S3/Azure Blob Storage) for raw documents**
5. **Fraud Detection & Anomaly Detection**
    - **Rule-Based and ML-Based** (Random Forest/XGBoost)
    - **Graph-Based Anomaly Detection** (Neo4j for detecting duplicate/fraudulent invoices)
6. **Integration with Downstream Systems**
    - **ERP Systems (SAP, Oracle, QuickBooks, NetSuite, Salesforce, Zoho, etc.)**
    - **BI Dashboards (Tableau/Power BI) for reporting**
7. **APIs & Microservices Layer**
    - **FastAPI / Flask / Django for exposing OCR services**
    - **Kafka / RabbitMQ for event-driven processing**
    - **Webhook Integration for real-time notifications**