

Intelligent Document Image Processing System

1. Introduction

The rapid digitization of business processes has increased the need for intelligent systems capable of converting large volumes of scanned and image-based documents into structured, actionable data. Organizations across finance, healthcare, legal, education, and government sectors continue to rely on documents such as invoices, application forms, contracts, and reports, many of which exist only in scanned or image format. Manual data extraction from these documents is inefficient, costly, and susceptible to human error.

The **Intelligent Document Image Processing System** is designed to address these challenges by employing advanced computer vision and document understanding techniques. Unlike traditional OCR solutions that focus solely on text extraction, this system emphasizes layout awareness, structural integrity, and contextual understanding of document components. By integrating OCR, layout detection, table and form extraction, and handwriting recognition into a unified platform, the system enables end-to-end document digitization with minimal human intervention, supported by a web-based document management interface.

2. Problem Statement

Despite significant advances in OCR technology, fully automated document understanding remains a complex problem. Conventional OCR systems often extract text as unstructured blocks, losing critical information related to layout, tables, and relationships between elements. Documents containing complex formatting, multi-column layouts, forms, or handwritten inputs pose additional challenges and often require manual post-processing.

Key problems include:

- Inaccurate text extraction from low-quality or skewed scans
- Inability to preserve document structure such as tables, key-value pairs, and sections
- Limited support for handwritten or mixed-content documents
- Lack of standardized structured outputs for integration with enterprise systems
- Absence of centralized platforms to manage, track, and audit document processing workflows

These limitations prevent organizations from achieving full automation in document-driven processes and hinder scalability, compliance, and operational efficiency.

3. Methodology

The proposed system will follow a comprehensive, multi-stage architecture designed for accuracy, scalability, and extensibility.

a. Document Ingestion and Image Preprocessing

The system will support scanned PDFs and common image formats (JPEG, PNG, TIFF). Preprocessing techniques such as deskewing, noise removal, contrast enhancement, and resolution normalization will be applied to improve OCR performance. Automatic orientation detection will ensure consistent alignment across documents.

b. OCR and Layout-Aware Text Extraction

Advanced OCR engines, including transformer-based OCR models, will be employed to extract textual content. Layout detection models such as LayoutLM or CNN-based layout detectors will identify document components including headings, paragraphs, tables, and form fields. This ensures that extracted text retains spatial and structural context.

c. Table, Form, and Handwritten Content Processing

Dedicated table extraction models will reconstruct row-column structures while preserving data integrity. Form understanding modules will identify key-value pairs and field boundaries. For documents containing handwritten inputs, handwriting recognition models will be integrated to extract text accurately alongside printed content.

d. Structured Data Generation and Web-Based Management

Extracted content will be converted into structured representations such as JSON and CSV, maintaining hierarchical and relational information. A web-based document management system will allow users to upload documents, monitor processing status, review extracted results, and download structured outputs. Secure authentication and audit logging will ensure enterprise readiness.

4. Outcomes

The successful completion of this project will result in a comprehensive document intelligence platform capable of handling diverse and complex document types. Key outcomes include:

- Accurate, layout-aware text extraction from scanned and image-based documents
- Reliable detection and reconstruction of tables, forms, and handwritten content
- Automated generation of structured outputs suitable for databases and analytics systems

- A scalable, user-friendly web platform for document processing and management
- Significant reduction in manual data entry effort and processing time

This system provides a strong foundation for enterprise document automation, digital transformation initiatives, and advanced AI-driven document understanding solutions.

