**Invoice Information Extraction System**

*(Theory Explanation in Points)*

**1. Purpose of the System**

* Automates extraction of **structured data** from scanned/digital invoices.
* Converts unstructured invoice images into **machine-readable JSON data**.
* Reduces **manual data entry errors** in accounting/ERP systems.

**2. Key Components**

**A. Optical Character Recognition (OCR) - PaddleOCR**

* Detects and extracts **text, numbers, and symbols** from invoice images.
* Provides **bounding box coordinates** for each detected text element.
* Includes **confidence scores** to measure accuracy.

**B. Machine Learning (LayoutLMv2) - Document Understanding**

* A pre-trained **Transformer-based model** fine-tuned for invoices.
* Classifies extracted text into semantic fields:
  + **Invoice Number**
  + **Date**
  + **Seller Name & Address**
  + **Total Amount**
  + **Line Items (Products, Quantity, Price)**

**C. Fallback Mechanism (Regex-Based Rules)**

* If the **AI model fails**, the system uses **pattern matching** (regex) to find:
  + Invoice No: INV-123 → Extracts INV-123
  + Total: $100.50 → Extracts 100.50
* Ensures **basic extraction** even without AI.

**D. Table Detection for Line Items**

* Groups text by **rows (Y-axis alignment)**.
* Detects **itemized lists** (e.g., product tables).
* Extracts:
  + **Item Number**
  + **Description**
  + **Quantity**
  + **Unit Price**
  + **Total Price**

**3. Workflow Steps**

1. **Image Preprocessing**
   * Converts the invoice to **grayscale**.
   * Enhances text using **adaptive thresholding**.
2. **Text Extraction (OCR)**
   * Detects all text elements with **position and confidence**.
3. **Structured Data Extraction**
   * **Option 1 (AI Mode):** Uses **LayoutLMv2** to classify text into invoice fields.
   * **Option 2 (Fallback Mode):** Uses **regex rules** if AI fails.
4. **Line Item Extraction**
   * Detects **tabular data** (e.g., product lists).
   * Groups items by **rows** and extracts details.
5. **Output Generation**
   * Saves data in **JSON format** with fields like:

json

{

"INVOICE\_NUMBER": "INV-001",

"TOTAL\_AMOUNT": "150.00",

"LINE\_ITEMS": [{"item\_no": "1", "description": "Laptop", ...}]

}

* + Includes **OCR quality metrics** (average confidence).

**4. Advantages**

✔ **Saves Time** – No manual data entry.  
✔ **High Accuracy** – Combines **OCR + AI + Heuristics**.  
✔ **Handles Variations** – Works even if invoice formats differ.  
✔ **Scalable** – Processes **batches of invoices** automatically.

**5. Limitations & Future Improvements**

* **Challenges:**
  + Struggles with **poor-quality scans**.
  + May miss **complex table layouts**.
* **Improvements Needed:**
  + Support for **multilingual invoices**.
  + Better **handwritten text recognition**.
  + Integration with **accounting software (QuickBooks, SAP)**.

**6. Business Applications**

* **Automated Bookkeeping** – Import invoice data directly into accounting tools.
* **Expense Management** – Extract totals and vendor details for reimbursement.
* **ERP Data Entry** – Reduce manual work in inventory/billing systems.