DAY:01
DATE:09.06.2025

OCR(OPTICAL CHARACTER RECOGNITION)

Task 1: Understand OCR (Optical Character Recognition)

◈ What is OCR?

OCR stands for **Optical Character Recognition**. It is a technology used to **convert different types of documents** (such as scanned paper documents, PDF files, or images captured by a camera) into **editable and searchable text**.

For example:

- A scanned image of a handwritten or printed answer sheet can be converted into machine-readable text using OCR.

How does OCR work?

OCR generally works in the following steps:

1. **Image Preprocessing**
   - Cleaning the image to remove noise
   - Converting to grayscale
   - Aligning the document properly (deskewing)
   - Enhancing contrast or brightness
2. **Text Detection**
   - The system locates where the text appears in the image.
3. **Character Recognition**

- It identifies each character or word using pattern matching or machine learning.

4. **Post-processing**

- Correcting errors (like "1" being detected as "I")

- Applying spell-check or language models to improve accuracy

Common Challenges in OCR

OCR has several challenges, especially in document extraction:

- **Poor image quality** (blurry or noisy scans)

- **Handwritten text** (especially if unclear or cursive)

- **Complex layouts** (columns, tables, or irregular placements)

- **Diagrams or equations** (which are not pure text)

- **Multiple languages or fonts** in a single document

- **Overlapping or skewed text**

- **Symbols** (like in math or scientific notations)

- Closely spaced characters

 Task 2: Explore Answer Sheet Types — with Examples

OCR performance varies depending on the **type of answer sheet**. Let's explore each type with examples and how OCR handles them.

---

 1. Theory-Based Answer Sheets

 Description:
These contain long handwritten or typed paragraphs with explanations, definitions, or essays.

⬜ Example Content:

vbnet

CopyEdit

Q: Explain Newton's First Law.

A: Newton's First Law states that an object remains at rest or in uniform motion unless acted upon by an external force.

🔍 OCR Behavior:

- OCR can read this type relatively well **if the handwriting is neat**.

- Tools like **Tesseract** or **Google Vision OCR** perform well here.

⚠️ Challenges:

- Cursive handwriting or poor penmanship.

- Words joined together or unclear spaces.

✅ Solution:

- Preprocess the image (increase contrast, remove noise).

- Use **handwriting-trained OCR models** (e.g., Microsoft Azure Read API or Google Cloud Vision).

◈ 2. Mathematical Answer Sheets

📝 Description:

These contain handwritten mathematical symbols, equations, derivations, or proofs.

⬜ Example Content:

makefile

CopyEdit

Q: Solve: $\int x^2 \, dx$

A: $\int x^2 \, dx = (x^3 / 3) + C$

🔍 **OCR Behavior**:

- Standard OCR fails or misinterprets symbols like $\int$, $\sqrt{}$, π, superscripts, and fractions.

- Specialized OCR tools are required for good accuracy.

⚠️ **Challenges**:

- Symbols like "x^2" may be read as "x2"

- Integral signs or square roots can be missed

- Spatial layout of fractions or matrices can confuse the OCR

✅ **Solution**:

- Use **MathPix OCR**, **InftyReader**, or **MyScript** (which are designed for math recognition).

- Output formats can include **LaTeX** for rendering equations accurately.

◈ **3. Diagram-Based Answer Sheets**

📝 **Description**:
These include labeled diagrams like circuits, biology drawings, flowcharts, or graphs.

⬚ **Example Content**:

sql

CopyEdit

[Diagram of a Human Heart]

Labels: Aorta, Left Ventricle, Right Atrium, Pulmonary Artery

🔍 OCR Behavior:

- OCR can extract **labels or annotations**, but not the diagram itself.

- Image portions (shapes, arrows) are not recognized as text.

⚠ Challenges:

- OCR may miss curved or rotated text.

- Diagrams with embedded text need **image + text segmentation**.

☑ Solution:

- Use **Computer Vision (CV)** with OCR (e.g., OpenCV + Tesseract).

- Manually or semi-automatically **segment text from the drawing**.

🖼 Example: Labelled Diagram of a Plant

⬛ Question:

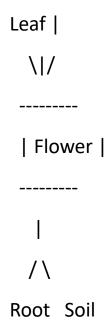vbnet

CopyEdit

Q: Draw and label a typical plant structure.

📝 Student Answer (hand-drawn or printed diagram):

The diagram includes labels such as:

🌿

Stem

|

```
   Leaf |

     \|/

   ---------

   | Flower |

   ---------

      |

     / \

   Root   Soil
```

📷 **In a real answer sheet:**

You would see a **simple sketch** of a plant with arrows pointing to parts like:

- **Root**

- **Stem**

- **Leaf**

- **Flower**

- **Soil**

◈ **4. Mixed Content Answer Sheets**

📝 **Description**:
These contain a combination of paragraphs, equations, and diagrams—typical in real-world answer sheets.

🗒 **Example Content**:

sql

CopyEdit

Q: Define Ohm's Law and draw the circuit diagram.

A: Ohm's Law states V = IR.

[Circuit Diagram Here]

Where V = Voltage, I = Current, R = Resistance

🔍 **OCR Behavior**:

- OCR struggles to determine which region is text, math, or image.

- Requires **layout analysis** or **region-based detection** before applying OCR.

⚠️ **Challenges**:

- Mixed fonts, symbols, handwritten notes, and drawings all together.

- Text placement may vary widely.

☑️ **Solution**:

- Use tools like **LayoutParser**, **Detectron2**, or **DocTR** to split the image into regions (text, math, image).

- Apply appropriate OCR model for each region.

THANKYOU