DAY:1
DATE:09.06.2025

Task 1 : Understand OCR

OCR (Optical Character Recognition) is the process of converting images of text into actual digital text that can be searched, edited, and processed by computers.

For example, when a scanned image of an **exam paper** is processed through OCR, the text of **questions and answers** can be extracted and stored digitally.

WORKING OF OCR:

Image Acquisition :

The document is scanned or photographed → resulting in an image file (JPG, PNG, or PDF).

Image quality is critical → better images result in more accurate text extraction.

Image Pre-processing :

OCR accuracy heavily depends on the quality of the input image. Pre-processing techniques are applied to scanned exam papers to enhance text visibility and improve recognition results. The key pre-processing steps used are listed below:

Grayscale conversion
Converts the image to a single intensity channel (gray), reducing complexity and improving text-background separation.

Binarization (black & white conversion)
Converts the image to pure black and white, enhancing contrast between text and background for better recognition.

Noise removal
Removes unwanted specks, stains, and background noise that can interfere with text extraction.

Deskewing
Corrects tilted or rotated text, ensuring that text lines are properly aligned for accurate segmentation.

Resolution enhancement
Ensures the scanned image meets the minimum resolution requirement (usually ≥300 DPI), which is critical for capturing small or fine text clearly.

Text Detection / Layout Analysis :

In this step, the OCR system analyzes the structure of the document and identifies where text is located.

It segments the image into distinct regions such as:

- Paragraphs
- Headings
- Lines of text
- Words
- Tables
- Non-text regions (images, diagrams)

By performing layout analysis, the system ensures that the original document structure is preserved, and text elements are extracted in the correct reading order.

This step is especially important for exam papers, where Questions, answers, tables, and diagrams may appear in different parts of the page.

Character Segmentation :

In this step, the OCR system breaks down the detected text regions into smaller units:

Lines ->Words -> Individual Characters

This process is known as segmentation.

Proper segmentation ensures that each character is isolated correctly before recognition.

If segmentation fails (due to connected letters, noise, or poor image quality), OCR accuracy will drop.

This step is especially challenging for:

- Handwritten text
- Mathematical equations
- Closely spaced characters

Accurate character segmentation is crucial for ensuring high recognition quality in exam papers, which often contain mixed handwriting styles and complex layouts.

Character Recognition :

In this step, the OCR system reads each character from the image and tries to figure out which letter, number, or symbol it is.

Modern OCR tools use AI models to recognize characters, even if they are written in different styles or fonts.

Example: The system looks at the shape of a letter and decides if it is an A, B, 3, 5, or any other symbol.

Post-processing & Output Generation :

After the text is recognized, the OCR system checks the results and tries to fix common errors (like confusing 0 with O, or 1 with l). It also arranges the text properly:

- Keeps paragraphs, headings, and tables in order.
- Applies spell-checking to improve accuracy.

Finally, the cleaned text is saved in a digital format such as:

- TXT (plain text)
- PDF (searchable)
- DOCX (editable document)

Now the text can be used for searching, editing, or AI-based grading which is exactly how it helps in exam paper extraction.

Common challenges in OCR Documentation :

Poor Image Quality :

Low-resolution scans, blurry images, skewed pages, and faded text significantly reduce OCR accuracy.

Handwritten Text :

Variability in handwriting styles, inconsistent spacing, and connected characters make handwriting recognition highly challenging.

Complex Layouts :

Documents with multiple columns, tables, diagrams, and mixed content often confuse the OCR system's layout analysis.

Mathematical Equations and Symbols :

Standard OCR engines are not designed to handle complex mathematical expressions, special symbols, or scientific notation.

Task 2 : Explore Answer Sheet Types

1.Theory-Based Answer Sheets

Structure:

- Mostly continuous text (paragraphs)
- Handwritten or typed text
- May include headings, underlines, bullet points

Impact on OCR:

✓ Works well when handwriting is clear.
✗ Errors occur if handwriting is cursive, faded, or slanted.
✗ Underlines and strikethroughs may interfere with character recognition.

2.Mathematical Answer Sheets (with Equations)

Structure:

- Mix of text + mathematical equations
- Handwritten or printed formulas
- Symbols like √, ∑, π, fractions, exponents

Impact on OCR:

✗ Regular OCR engines struggle with math content.
✗ Symbols are often misrecognized or skipped.
✗ Handwritten equations are particularly difficult to process.
✓ Text portions of answers may still be extracted accurately.

3.Diagram-Based Answer Sheets

Structure:

- Contains diagrams, flowcharts, circuit diagrams
- Text inside diagrams (labels)
- Text outside diagrams (captions, explanations)

Impact on OCR:

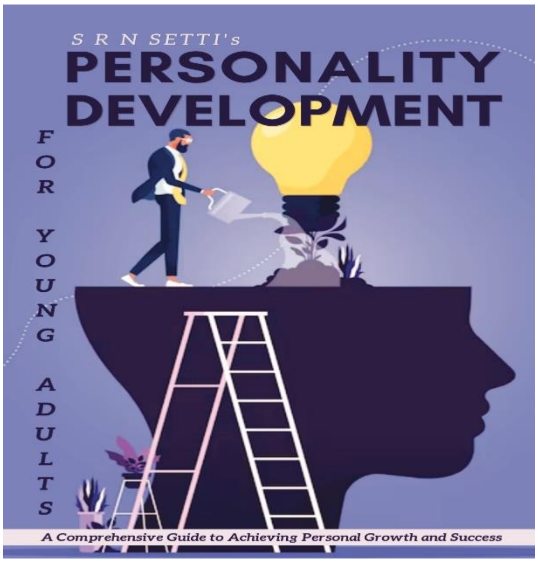
- ✗ Diagrams themselves cannot be processed by OCR.
- ✗ Complex layouts can confuse OCR engines, causing text from diagrams to be mixed with body text.
- ✓ Labels inside diagrams may be extracted if clear.

Sample Text Extraction using pytesseract

Program :

```
import pytesseract

from PIL import Image

pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tesseract.exe'

img_path = r"C:\Users\pavit\OneDrive\Pictures\Screenshots\Screenshot 2025-05-12 220101.png"

# Load the image

img = Image.open(img_path)

# Run OCR

text = pytesseract.image_to_string(img)

# Print extracted text

print("----- Extracted Text -----")

print(text)
```

Image given :


S R N SETTI's
PERSONALITY DEVELOPMENT
FOR YOUNG ADULTS
A Comprehensive Guide to Achieving Personal Growth and Success

Output :

```
----- Extracted Text -----

PERSONALITY

DEVELOPMENT

QOQ2aGQ0* aw O Tl

qyrcqop,p

A Comprehensive Guide to Achieving Personal Growth and Success

| ay 1 |
```