## 🧾 Project: Text Recognition from Images using Pytesseract & OpenCV

### 🗂️ Project Structure

ocr\_text\_recognition/

├── images/

│ └── sample1.png # Add some scanned text/image files here

├── ocr\_utils.py # Helper functions

├── ocr\_app.py # Streamlit interface

├── requirements.txt

└── README.md

### 1️⃣ ocr\_utils.py – Core OCR & Preprocessing

import cv2

import pytesseract

def load\_image(image\_bytes):

file\_bytes = np.asarray(bytearray(image\_bytes), dtype=np.uint8)

image = cv2.imdecode(file\_bytes, 1)

return image

def preprocess\_image(image):

gray = cv2.cvtColor(image, cv2.COLOR\_BGR2GRAY)

gray = cv2.medianBlur(gray, 3)

\_, thresh = cv2.threshold(gray, 127, 255, cv2.THRESH\_BINARY)

return thresh

def extract\_text(image):

return pytesseract.image\_to\_string(image)

def extract\_data(image):

return pytesseract.image\_to\_data(image, output\_type=pytesseract.Output.DICT)

### 2️⃣ ocr\_app.py – Streamlit Frontend

import streamlit as st

from ocr\_utils import \*

import numpy as np

import cv2

st.set\_page\_config(page\_title="OCR Text Recognition")

st.title("📝 OCR Text Recognition with Pytesseract")

uploaded\_file = st.file\_uploader("Upload an image with text", type=["jpg", "jpeg", "png"])

if uploaded\_file:

image = load\_image(uploaded\_file.read())

st.image(image, caption="Original Image", use\_column\_width=True)

thresh = preprocess\_image(image)

st.image(thresh, caption="Preprocessed Image", use\_column\_width=True)

st.subheader("📄 Extracted Text")

text = extract\_text(thresh)

st.code(text)

st.subheader("🔍 Word Detection with Bounding Boxes")

data = extract\_data(thresh)

n\_boxes = len(data['level'])

for i in range(n\_boxes):

(x, y, w, h) = (data['left'][i], data['top'][i], data['width'][i], data['height'][i])

cv2.rectangle(image, (x, y), (x + w, y + h), (0, 255, 0), 2)

st.image(image, caption="Detected Words", use\_column\_width=True)

### 3️⃣ requirements.txt

streamlit

opencv-python

pytesseract

numpy

### 4️⃣ Download Tesseract-OCR

* **Windows:** [UB Mannheim Build (recommended)](https://github.com/UB-Mannheim/tesseract/wiki)
* **Linux:** sudo apt install tesseract-ocr
* **macOS:** brew install tesseract

🛠️ **Don't forget** to set the path on Windows:

# Add this in ocr\_utils.py (if needed)

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

## ✅ Run the App:

streamlit run ocr\_app.py

## 📌 Bonus Features to Add

* Language selection (lang='eng+hin')
* Handwriting OCR (limited)
* Highlight low-confidence words
* Download extracted text

Would you like me to zip all this as a download package or help you deploy it to Hugging Face or Streamlit Cloud? <https://youtu.be/HNCypVfeTdw?feature=shared> <https://github.com/tesseract-ocr/tesseract>

The error indicates that you've installed Pytesseract via pip install pytesseract, but the Tesseract OCR engine itself isn't installed via pip. Tesseract OCR is a separate executable, and pip install pytesseract only installs the Python wrapper for Tesseract, not the engine itself.