Project Title: OCR-Based Text Detection and Extraction Web App

---

1. Introduction

The "OCR-Based Text Detection and Extraction Web App" is a lightweight, user-friendly application built using Python and Flask. It allows users to upload image files and extract the embedded text using the Tesseract OCR engine. This is especially useful in digitizing printed or handwritten documents and images.

---

2. Objective

The primary objective of this project is to:

- Enable text extraction from images (JPG, PNG, etc.)
- Provide a web interface to upload and process images
- Display the extracted text in a readable and usable format
- Help users quickly digitize physical documents or snapshots

---

3. Technologies Used

- **Backend**: Python, Flask
- **Frontend**: HTML, CSS
- **OCR Engine**: Tesseract OCR
- **Image Processing**: Pillow (PIL)

---

4. System Requirements

- Python 3.8+

- Flask
- Tesseract OCR (installed and configured)
- Web browser (for user interaction)

---

5. Installation & Setup

1. Clone the repository:

```
git clone https://github.com/dhanalakshmim-eng/cantilever.git
cd cantilever/project_2_OCR_Text_Extraction
```

2. Create and activate a virtual environment (optional but recommended):

```
python -m venv venv
source venv/bin/activate  # For Windows: venv\Scripts\activate
```

3. Install dependencies:

```
pip install -r requirements.txt
```

4. Install Tesseract OCR:

- Ubuntu: `sudo apt install tesseract-ocr`
- Windows: Download from the official GitHub repository
- Mac: `brew install tesseract`

5. Run the application:

```
python app.py
```

---

6. Working of the Application

- User visits the homepage and uploads an image.
- The Flask backend processes the image.
- Tesseract OCR extracts text from the image.
- The extracted text is displayed below the uploaded image.

---

7. Folder Structure

```
project_2_OCR_Text_Extraction/
├── app.py
├── templates/
│   └── index.html
├── requirements.txt
└── README.md
```

---

8. Sample Output

Input Image:
An image with printed or handwritten text

Extracted Text:

This is a sample OCR output from the uploaded image.

---

9. Applications

- Digitization of documents
- Business card text extraction
- Automating data entry from scanned files
- Assisting visually impaired individuals

10. Limitations

- Accuracy depends on image clarity
- Struggles with skewed, blurred, or handwritten text
- Language support depends on Tesseract model

11. Future Enhancements

- Support for PDF file uploads
- Multi-language text extraction
- Image preprocessing for better accuracy (e.g., grayscale, noise removal)
- Download option for extracted text

12. Demo Video

Watch the demo video here:
https://drive.google.com/file/d/1KbDEVNQXQKUMN1qgZkXyzww81Nr1Cbhp/view?usp=sharing

13. Developed By

Name: Dhana Lakshmi M
B.E. Computer Science and Engineering

GitHub: https://github.com/dhanalakshmim-eng/cantilever.git