**Project: Aadhaar/PAN Card Data Extraction**

**Objective:** Extract details from Aadhaar or PAN card images using OCR techniques.

**Requirements:**

●   Use **easyocr** or **pytesseract** for text extraction

●   Use OpenCV to load .png images

●   Use Python’s re library to extract name, ID number, DOB, etc.

*(Optional)* Create a simple UI using **Streamlit** or **Gradio** to upload image and show results

**OCR - optical character recognition**

* Electronic/mechanical conversion of images of typed handwritten or printed text into machine encoded text, whether in the form of a scanned doc, a photo of a doc, and scene photo or subtitle text on an image.
* Singles out letters, puts to words, which are put into sentences
* Eliminates wasted effort of manual data entry
* Hardware: optical scanners, specialized circuit boards
* Software: deals with advanced processing.
* OCR software can use AI to implement more advanced methods of ICR (Intelligent Character Recognition) to identify languages or handwriting.
* Can be used to convert printed legal/historical docs into pdf docs so users can modify as required.
* Internet connected mobile devices that do not have built in OCR functionality can use OCR API.

**TYPES**

* Simple OCR: This method relies on pattern matching, comparing scanned characters to a database of known glyphs. It's less sophisticated and may have limitations with diverse fonts and languages.
* Optical Mark Recognition (OMR): OMR identifies specific marks on forms, such as bubbles or checkboxes, and logos or watermarks. It relies on matching these marks to stored images.
* Intelligent Character Recognition (ICR): ICR uses AI and machine learning to analyse text, recognizing individual characters based on features like curves, loops, lines, and intersections. This approach enables the recognition of handwriting and diverse fonts.
* Intelligent Word Recognition (IWR): IWR builds upon ICR by recognizing entire words instead of individual characters. This can be more efficient for languages with complex word structures.
* Other OCR Variations: Beyond these primary types, there are other variations, including scene text recognition, which extracts text from natural scenes like signs and billboards, and optical music recognition, which converts musical notation into a digital format.

**EASY OCR**

* Python OCR module that is easy and flexible to use.
* Multiple language support, pre - trained text detection and identification models, focus on accuracy, speed and efficiency in word recognition inside images.
* Components :

               . Feature extraction (Resnet and VGG)

               . Sequence labelling (uses LSTM - long short-term memory)

               . Decoding (uses CTC algorithm - Connectionist Temporal Classification)

**OpenCV - Open Source Computer Vision Library**

* Open source library for comp vision, image processing and ml software. One can process images and videos to identify objects, faces or even handwriting of a human.
* Free for both academic and commercial use
* Opencv allows to perform - image reading, image enhancement, object detection, image filtering, draw in the image, save changes
* Applications: face recognition, automated inspection and surveillance, object count etc.
* Image processing is a method to perform some operations on an image, in order to get an enhanced image and or to extract some useful information from it.

**RegEx - regular expression**

* Sequence of characters that forms a search pattern.
* Used to check if a string contains the specified search pattern.
* Built in package re, used to work with regular expressions
* Re module as functions that allow to search string for a match: findall, search, split, sub

**SUMMARY**

Made an aadhaar/pan card data extraction in python, for core logic and gradio for easy ui layout to upload the card images, and display extracted details.

Used different imported libraries like cv2, easyocr and re (regex). Used different functions named:

* pre-process image(simplifies the image by removing colour and applying binary thresholding)
* Extract text from image (get image file path, initialize reader from OCR and Returns the recognized text as a single string, with each line separated by a newline (\n).)
* 2 functions to extract details from pan and aadhaar
* Process document(used to extract text from image, check if aadhaar or pan by identifying format).