Comparative Study of Different OCR Models:

1. **Tesseract**:

* Open sourced developed by HP which is known for the accuracy and versatility.
* 100 languages.
* Offers command line interface.
* 80-90% accuracy with good quality images.
* <https://github.com/tesseract-ocr/tesseract> link to the open-source code
* Best suited for printed text.
* Primary source code is in C++, User interacts with the library which is through Python.

Pros: Languages, Efficiency, the development team fixes bugs and performance increases Cons: Not accurate for handwritten docs (around 50%) if we train it accordingly its efficiency might increase. We can use AWS Textract for handwritten docs.

Increase the efficiency of the model by adding more preprocessing steps.

1. **EasyOCR**:

* Open sourced developed by PyTorch.
* 80 languages.
* Needs python (>3.6)
* Provides better results multilingual and handwritten text.
* <https://github.com/JaidedAI/EasyOCR> link to the open-source code
* <https://www.jaided.ai/easyocr/documentation/>
* 85-95% accuracy with good quality images.
* For multilingual we can use EasyOCR, for English the accuracies are similar.

Pros: Multilingual, Better Efficiency, few steps to install

Cons: Not accurate enough with handwritten (around 75%)

Increase efficiency by adding more preprocessing and use natural language processing tools to correct the spellings.

**Tesseract on CPU but if you have GPU available, use EasyOCR**

1. **PaddleOCR**:

* Open sourced developed by PaddlePaddle.
* 80 languages.
* Gives good accuracy for multilingual docs and complex structured documents
* Lightweight model and supports GPU
* Complex configuration and deployment are difficult
* Limited Documentation and Tutorials
* Real-time OCR.
* 85-95% accuracy with good quality images.
* 70-85% accuracy for handwritten texts.
* <https://github.com/PaddlePaddle/PaddleOCR>
* <https://github.com/Mushroomcat9998/PaddleOCR/blob/main/doc/doc_en/multi_languages_en.md>
* Should convert pdfs to image.

1. **MMOCR**:

* Multi-Modal Optical Character Recognition is developed by OpenMMLab.
* Needs python (>3.7)
* Includes many state-of-the-art OCR algorithms.
* Requires a GPU for efficient processing, especially for large datasets.
* Limited handwriting support and complex setup.
* 85-95% accuracy with good quality images.
* <https://github.com/open-mmlab/mmocr>
* It has integrated cutting edge models
* Customisable and higher efficiency due to the multi model

|  |  |  |  |
| --- | --- | --- | --- |
|  | TESSERACT | EASY OCR | PADDLE OCR |
| Accuracy | 80-90% | 85-95% | 85-95% |  |
| GPU Usage | Optional | Optional | Optional |
| User friendly | Yes | Yes | Moderate |
| Multilingual | Yes | Yes | Yes |
| Handwriting docs | 50% | 75% | 80% |
| PDF | No | No | No |