1. Definition
   1. OCR is a process of recognizing text inside images and covering it into an electronic form.

A common pipeline of OCR system is:

1. Image preprocessing(alignment, cleaning, binarization)
2. Text detection
3. Text segmentation
4. OCR
5. Restructuring
6. NLP

* There are 2 type of OCR: Structure(document) and unstructure data (street sign, text on bottle…) -> We are focusing on document type
* OCR system is filled with problems like different fonts in images, poor contrast, multiple objects in an image

2. Application:

* Reducing manual human effort.
* OCR is increasingly being used for digitization by various industries to cut down manual workload. It makes very easy and efficient to extract and store information from business documents, receipts, invoices, passports, etc

**SOTA Analysics:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Main Method** | **Weakness** |  | **Dataset** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

<https://github.com/chongyangtao/Awesome-Scene-Text-Recognition>

Tesseract OCR (originally developed as proprietary software by Hewlett-Packard).

As of today(May 16, 2020), Tesseract can detect over 100 languages and can provess even right-to-left text such as Arabic or Hebrew! It is uses by Google for text detection on mobile devices. From version 4, Google has given a significant boost to this OCR engine. OCR engine that uses a neural network system based on LSTM (Long Short term Memory), one of the most effective solutions for sequence prediction problems. Tesseract OCR engine 4 has both LSTM and Legacy OCR engines.

Tesseract OCR can work with Vietnamese

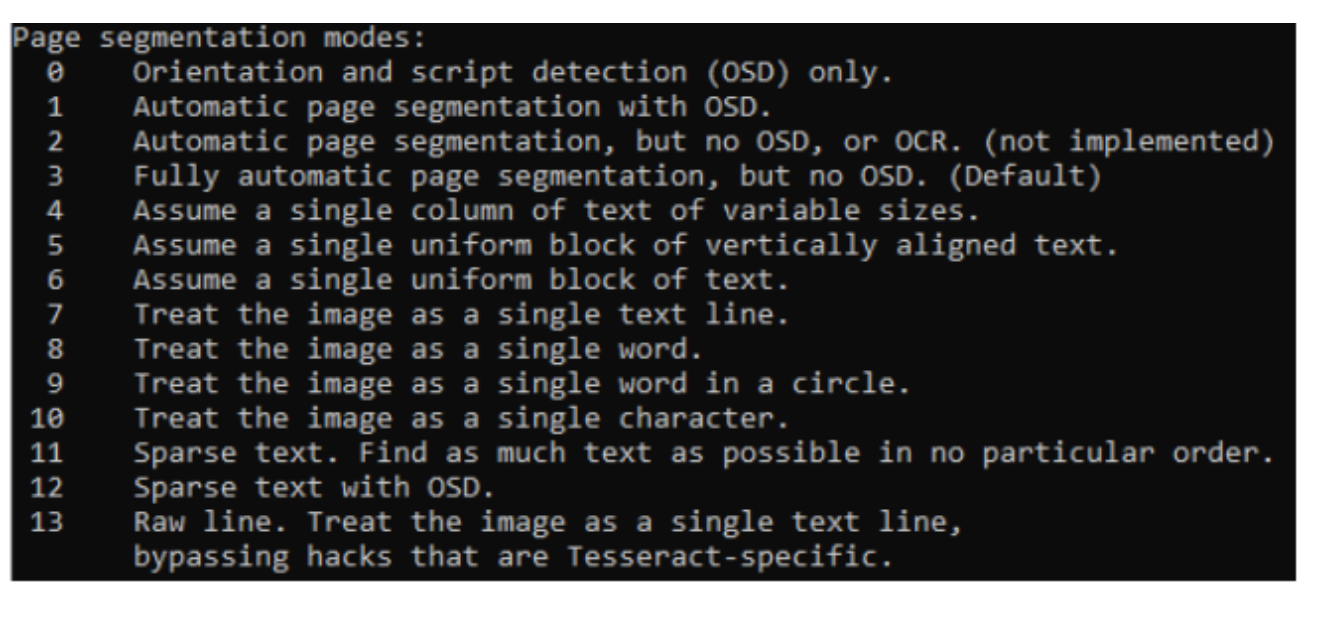
([https://medium.com/@langtv/viết-chương-trình-một-cách-đơn-giản-bằng-python-dùng-tesseract-để-nhận-dạng-chữ-tiếng-việt-9d9b6ee2e49f](https://medium.com/@langtv/vi%E1%BA%BFt-ch%C6%B0%C6%A1ng-tr%C3%ACnh-m%E1%BB%99t-c%C3%A1ch-%C4%91%C6%A1n-gi%E1%BA%A3n-b%E1%BA%B1ng-python-d%C3%B9ng-tesseract-%C4%91%E1%BB%83-nh%E1%BA%ADn-d%E1%BA%A1ng-ch%E1%BB%AF-ti%E1%BA%BFng-vi%E1%BB%87t-9d9b6ee2e49f))

Challenges of Tesseract:

* Weakness: Noise images, or font of the language is one on which Tesseract OCR is not trained. Brightness or skewness of text may be affect the performance of Tesseract. Tesseract need the image is fairly clean. So we need a good text detection system that can detect text which can then be easily extracted.
* There’re some ways for text detection:
  + Opencv
  + Deep Learning Model: **Contemporary Deep Learning Model – EAST**

https://www.analyticsvidhya.com/blog/2020/05/build-your-own-ocr-google-tesseract-opencv/

* + Own Custom Model



You can config modes for specific images type that you use.

SOTA:

The existing OCR engines provide high-quality results for modern printed text, when the output is often above 99% correct. There is, however, a large room for improvement for their efficacy of OCR on historical text or manuscripts. The current technology, when applied for example to 16th century books, often leads to text where most of the words are recognized wrong. The reasons for this low perfermance are multiple:

* Old fonts which are not adequately interpreted by modern OCR devices.
* SOTA research(lasted update: 8 months ago): <https://github.com/chullhwan-song/OCR>

**SOME SOLUTIONS:**

**Base case:**

A basic systems have to cover at least the following steps:

* Image preprocessing(alignment, cleaning, binarization)
* Character recognition
* Text post processing(for example, error handling based on language models)

Some famous models:

* CTPN(text detection) and CRNN(text regconizition): <https://github.com/chrishzhao/ocr_dev> (printed paper OCR)
* CNN(text detection), RNN + some methods (text regconizition) 84.2% on FSNS(French Street Name Signs) dataset.(<https://paperswithcode.com/task/optical-character-recognition>)
* <https://medium.com/synapse-medicine/recognizing-drugs-on-medical-prescriptions-557b4e8103b2> (specific case: prescriptions in French)

Medical prescription ocr

References: <https://www.analyticsvidhya.com/blog/2020/05/build-your-own-ocr-google-tesseract-opencv/>

Useful resources: <https://sites.google.com/site/textdigitisation/home/creation>

NOTES: Dùng tool để generate hình ảnh từ text: [Ứng dụng Deep Learning cho OCR | Bài toán nhận diện chữ viết OCR](https://topdev.vn/blog/ung-dung-deep-learning-cho-ocr/) (CTC loss func ENGLISH)

<https://dropbox.tech/machine-learning/creating-a-modern-ocr-pipeline-using-computer-vision-and-deep-learning> (generate data)