## **Overview**

OCR technology is useful for processing scanned documents and images, but it is extremely challenging to process engineering documents. A dedicated area needs to be designated to extract the required text, and counts such as regularization need to be used after OCR to adjust the accuracy and format of the data. Extracting accurate text and numbers from tables, infoboxes and technical information is important to automate processes and get a highly accurate result. Unfortunately, there is no high-quality open source solution that implements the entire process.

## **Methodology**

图示, 工程绘图

描述已自动生成

The area of the information part in all PDF documents is fixed, which means that OCR can be performed by specifying the area to avoid interference from information from other parts.

### **1.3 OCR Library Test and Evaluation**

In our OCR library evaluation, we compared three popular OCR tools: PytesseractOCR, EasyOCR, and KerasOCR, highlighting each tool's unique approach to extracting text from images.

PytesseractOCR: PytesseractOCR is a widely used OCR tool that excels at recognizing text in images. When extracting text from a rectangular object using PytesseractOCR, we first identify the rectangular box surrounding the text area in the image. During the subsequent text extraction process, we take appropriate segmentation measures to maintain the structure of the text lines. Text extraction is performed by using the pytesseract.image\_to\_string(roi) function, where roi represents the region of interest, which is the detected text area in the rectangle. We can leverage its ability to recognize text to infer the structure of the table. By analyzing the layout and position of the text, we can infer the rows and columns of the table and further predict the position and boundaries of the cells. PytesseractOCR can effectively identify the boundaries and positions of text lines and extract the text content.

EasyOCR: EasyOCR is another powerful OCR tool known for its simplicity and efficiency in extracting text from images. EasyOCR uses an end-to-end deep learning model and is able to process text in a variety of languages and fonts. By simply calling the function easyocr.Reader(), we can load the EasyOCR model and extract text from the image. EasyOCR's flexibility and accuracy make it ideal for handling a variety of OCR tasks.

KerasOCR: KerasOCR is an OCR tool built on the Keras framework and has simple and powerful features. It uses advanced deep learning technology to accurately recognize text in images. By using KerasOCR’s pre-trained model, we can easily extract text from images. KerasOCR's model flexibility and customizability make it ideal for specific OCR needs.

PaddleOCR:

**Result Analysis**