4. OCR Tesseract

4.1 Software and technique:

ImageMagic

ImageMagic is free and open source software that can create, edit, compose, or convert bitmap images. It supports PNG, JPEG, GIF, HEIC, TIFF, DPX, EXR, WebP, Postscript, PDF, and SVG files. It can be utilized from the command line or program. It supports various types of programming languages such as c, c++, Java, Python, and many more.

OCR

Optical character recognition(OCR) is a technique of converting images to machine encoded text. Generally, most of the OCR has two part which are pre-processing, and character recognition. On pre-processing, it edits the given images to raise the accuracy of the character recognition rate. To do this, OCR binarize the image which convert images from color to black and white. This process will make easier for program to separate the text from the background. After binarization, OCR takes segmentation process. First, it separates the page into different sections such as photo, table, and text box. After segmenting the page, it locates where the text is and segment it to line, word, then character. When pre-processing is done, OCR recognize the characters and converts to machine encoded text. To do this, OCR look for pattern and features of given image. For example, when image of the character given, OCR finds similar character from library, and extracts features such as number of lines and angle between lines. After evaluating given image, it gives the closes machine encoded character.

OCR Tesseract

OCR Tesseract is an open source free character recognition engine. It can be used by entering command in terminal or codding program with built in API. It supports 10 different programming language interfaces, and more than 100 different languages including top-to-bottom languages (such as Chinese) and right-to-left languages (such as Hebrew). It also provides different types of page segmentation methods. OCR Tesseract is good at recognizing printed text, but not good at recognizing hand-written text. However, OCR Tesseract is trainable. It can train new language, and it can train different font up to 64 fonts.

4.2 Why we are using it

Our group decided to use both ImageMagic and OCR Tesseract to build signature locating program. The basic algorithm of the program is like this. When pdf document is given, it converts the given pdf document to machine encoded text. Then, it searches for certain words such as “Signature” to locate the signature box. Lastly, program save the signature box and pass it to our convolutional neural network to detect signature.

The main reason of using ImageMagic and OCR Tesseract is that our group have limited time. We could build software that can convert pdf to image file and recognize the characters, but it will take a lot of time. Since our client wanted us to focus on making accurate signature detection AI, our group wanted to focus on building decent AI. To do this, we decided to use free open source software for locating signature.

The other reason that ImageMagic and OCR Tesseract is suits to build signature locating program. ImageMagic is good at dealing with bitmap images, so it can be used to convert pdf to image file. OCR Tesseract converts printed text to machine encoded text well. It also provides different types of page segmentation method. Since all the documents in the workplace have different format, having decent page segmentation is critical for locating where the text is. In addition, OCR Tesseract can recognize more than 100 different languages. For this project, we only focus on dealing with English document, but for later update, it is good to support different languages.

4.3 What it has been able to do

The alpha version of our signature locating program can convert pdf to png, do page analysis, convert image of text to machine encoded text, search “CUSTOMER SIGNATURE”, and save the signature box as png image file. Since this is alpha version, there are two major issue. First problem is the search method. For locating the signature box, we search word “CUSTOMER SIGNATURE” from converted document by tesseract OCR. However, since many documents have different signature box format, we need to come up with the search word list to work on different types of documents. Also, when signature is signed over the search word, program have difficulties locating the signature box.

The second limitation for our alpha version program is that it provides rough location of signature box. Since we locate the signature box by searching word, we do not know exact location and size of the signature box. Therefor, we manually set variable for calculating location and size of the signature box. It locates the signature box with many noises.

4.4 Possible solution

Search Method

It is difficult to locate the signature by searching certain word in the document. The performance of the program can be improved by making decent search list. Also, when program searches the word, rather than finding exact word, find the similar word in the document. For example, when we search for “CUSTOMER SIGNATURE,” we can take word such as “CUST@MER S!GNATURE” as valid input.

Finding Accurate Location

The possible solution for this problem is searching location of the horizontal line. Signature box is consisted with two parts which are text and line. If we are able to search word and locate it, we can search for closest horizontal lines to location of searched word. After locating the signature sign line, software can cut the slightly below and above the sign line to locate more accurate signature box.