Automating Receipt Digitization with OCR and Deep Learning

Receipt OCR or receipt digitization addresses the challenge of automatically extracting information from a receipt. In this article, I cover the theory behind receipt digitization and implement an end-to-end pipeline using OpenCV and Tesseract. I also review a few important papers that do Receipt D

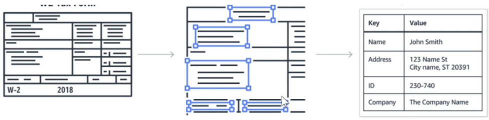
Content

What is Receipt Digitization?

Receipts carry the information needed for trade to occur between companies and much of it is on paper or in semi-structured formats such as PDFs and images of paper/hard copies. In order to manage this information effectively, companies extract and store the relevant information contained in these documents. Traditionally this has been achieved by manually extracting the relevant information and inputting it into a database which is a labor-intensive and expensive process.

Receipt digitization addresses the challenge of automatically extracting information from a receipt.

Extracting key information from receipts and converting them to structured documents can serve many applications and services, such as efficient archiving, fast indexing and document analytics. They play critical roles in streamlining document-intensive processes and office automation in many financial, accounting and taxation areas.



Who will find Receipt Digitization useful?

Here are a few areas where Receipt Digitization can make a huge impact:

Accounts payable and receivables automation

Computing Accounts payable (AP) and Accounts Receivables (ARs) manually is costly, time-consuming and can lead to confusion between managers, customers and vendors. With digitization, companies can eliminate these drawbacks and can have more advantages - Increased Transparency, Data Analytics, Improved working capital and easier tracking.

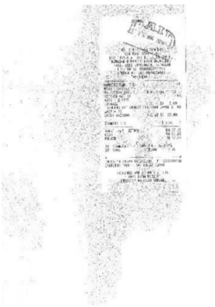
Supply chain optimization

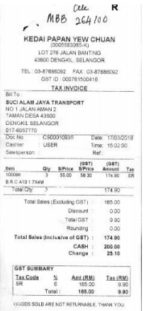
Supply chains are the backbone of many a company’s proper functioning. Managing tasks, information flows, and product flows is the key to ensuring complete control of supply and production. This is essential if organizations are to meet delivery times and control production costs.

The companies that are truly thriving these days have something significant in common: a digitized supply chain. 89% of companies with digital supply chains receive perfect orders from international suppliers, ensuring on-time delivery. One of the key elements of realising the next generation digital Supply Chain 4.0, is automating data capturing and management and a lot of this data is the form of receipts and invoices. Manual entry of receipts acts as a bottleneck across the supply chain and leads to unnecessary delays. If this receipt processing is digitized it can lead to substantial gains in time and efficiency.

Why is it a difficult problem?

Receipt digitization is difficult since receipts have a lot of variations and are sometimes of low quality. Scanning receipts also introduces several artifacts into our digital copy. These artifacts pose many readability challenges.





Cele R
MBB 264/100

KEDAI PAPAN YEW CHUAN
(0005583085-K)
LOT 276 JALAN BANTING
43800 DENGKIL, SELANGOR

TEL : 03-87686092 FAX : 03-87686092
GST ID : 000781500416

TAX INVOICE

Bill To :
SUCI ALAM JAYA TRANSPORT
NO 1 JALAN AMAN 2
TAMAN DESA 43800
DENGKIL SELANGOR
017-6057770

Doc No	: CS00010930	Date:	17/03/2018
Cashier	: USER	Time:	15:02:00
Salesperson	:	Ref.:	

Item	Qty	S/Price	(GST) S/Price	(GST) Amount	Tax
100088 B.R.C A10 1.7X4M	3	55.00	58.30	174.90	SR
Total Qty:	3			174.90	

Total Sales (Excluding GST)	:	165.00
Discount	:	0.00
Total GST	:	9.90
Rounding	:	0.00
Total Sales (Inclusive of GST)	:	**174.90**
CASH	:	**200.00**
Change	:	**25.10**

GST SUMMARY

Tax Code	%	Amt (RM)	Tax (RM)
SR	6	165.00	9.90
	Total :	**165.00**	**9.90**

GOODS SOLD ARE NOT RETURNABLE. THANK YOU.

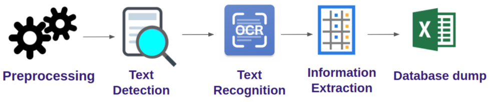
Here’s is a list of the a few things that make it a difficult problem to crack

- Handwritten text
- Small fonts
- Noisy images
- Faded images
- Camera motion and shake
- Watermarkings
- Wrinkles
- Faded text

A Traditional Receipt Digitization Pipeline

A typical pipeline for this kind of an end-to-end approach involves:

- Preprocessing
- Optical Character Recognition
- Information Extraction
- Data dump



Let’s dive deeper into each part of the pipeline. The first step of the process is Preprocessing.

Preprocessing

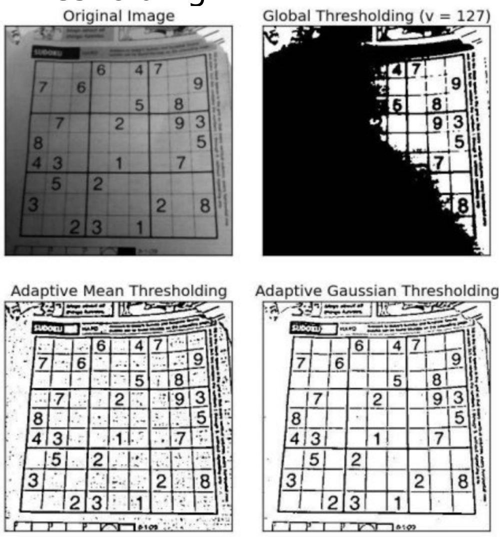
Most scanned receipts are noisy and have artefacts and thus for the OCR and information extraction systems to work well, it is necessary to preprocess the receipts. Common preprocessing methods include - Greyscaling, Thresholding (Binarization) and Noise removal.

Grayscaling is simply converting a RGB image to a grayscale image.

Noise removal typically involves removing Salt and Pepper noise or Gaussian noise.

Most OCR engines work well on Black & White images. This can be achieved by thresholding, which is the assignment of pixel values in relation to the threshold value provided. Each pixel value is compared with the threshold value. If the pixel value is smaller than the threshold, it is set to 0, otherwise, it is set to a maximum value (generally 255).

OpenCV provides various thresholding options - Simple Thresholding, Adaptive Thresholding



Original Image

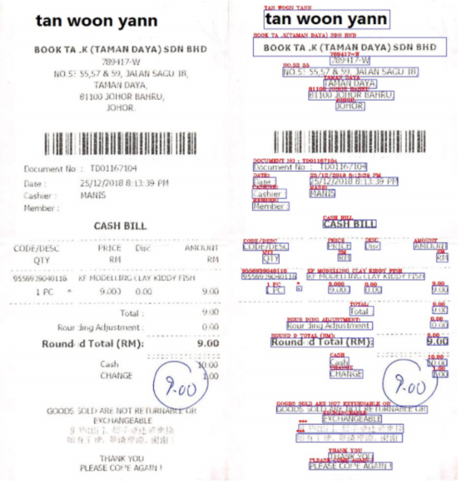
Global Thresholding (v = 127)

Adaptive Mean Thresholding

Adaptive Gaussian Thresholding

Optical Character Recognition

The next step in the pipeline is OCR. It is used to read text from images such as a scanned document or a picture. This technology is used to convert, virtually any kind of images containing written text (typed, handwritten or printed) into machine-readable text data. OCR involves 2 steps - text detection and text recognition.



tan woon yann

BOOK TA .K (TAMAN DAYA) SDN BHD
789417-W
NO.5: 55,57 & 59, JALAN SAGU 18,
TAMAN DAYA,
81100 JOHOR BAHRU,
JOHOR.

Document No : TD01167104
Date : 25/12/2018 8:13:39 PM
Cashier : MANIS
Member :

CASH BILL

CODE/DESC PRICE Disc AMOUNT
QTY RM RM
9556939040116 KF MODELLING CLAY KIDDY FISH
1 PC * 9.000 0.00 9.00

Total : 9.00
Rour :ding Adjustment : 0.00
Round :d Total (RM): 9.00

Cash 10.00
CHANGE 1.00

9.00

GOODS SOLD ARE NOT RETURNABLE OR EXCHANGEABLE
[illegible]
[illegible]

THANK YOU
PLEASE CO!'E AGAIN !

There are a number of approaches to OCR. The conventional computer Vision approach is to

- Using filters to separate the characters from the background
- Apply contour detection to recognize the filtered characters
- Use mage classification to identify the characters

Applying filters and image classification is pretty straightforward, (think MNIST Classification using SVN), but contour matching is a very difficult problem and requires a lot of manual effort and is not generalizable.

Next come the Deep Learning approaches. Deep Learning generalizes very well. One of the most popular approaches for text detection is EAST. EAST (Efficient accurate scene text detector) is a simple yet powerful approach for text detection. The EAST network is actually a version of the well known U-Net, which is good for detecting features of different sizes.

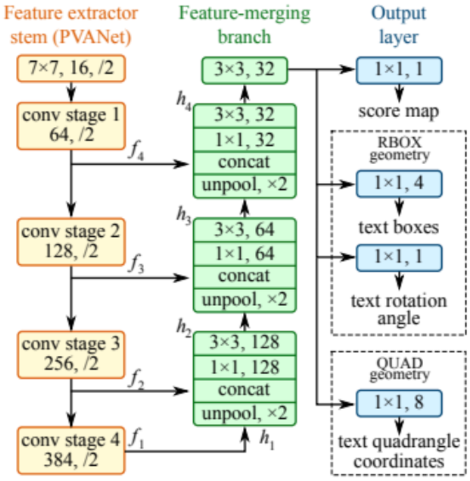


Figure 3. Structure of our text detection FCN.

CRNN and STN-OCR (Spatial Transformer Networks) are other popular papers that perform OCR.

Information Extraction

The most common approach to the problem of Information Extraction is rule-based, where rules are written post OCR to extract the required information. This is a powerful and accurate approach, but it requires you to write new rules or templates for a new type of document.

Several rule-based invoice analysis systems exist in literature.

- Intellix by DocuWare requires a template annotated with relevant fields.
- SmartFix employs specifically designed configuration rules for each template

The rule-based methods rely heavily on the predefined template rules to extract information from specific invoice layouts

One approach that has become very common in the past few years is to use a standard Object Detection framework like YOLO, Faster R-CNN to recognize fields. So instead of pure text detection, field recognition and text detection are performed simultaneously. This makes the pipeline smaller (Text Detection→ Recognition → Extraction to Detection → Recognition). There is no need to write any rules since the object detector’s learn to recognize these fields.

Data Dump

Once you have your information extracted, the data dump can be done as our use case requires. Often a JSON format to store the fields information is convenient. These JSON files can readily be converted into XML files, Excel sheets, CSV files or plaintext files depending on who wants to work with the data and how.