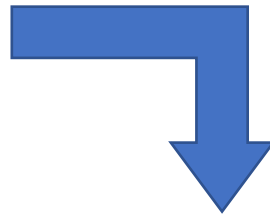


Documentation

Objective of the OCR code



| | food | Cost | Receipt_ID | Supermarket | date | Store | Processed | UserID | StoreID |
|----|-----------------------------|------|------------|-------------|------------|--------------------------|-----------|--------|---------|
| 0 | DAIRY FULL CREAM SLITRE | 8.25 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 1 | KELLOGGS CORN FLAKES OBGRAM | 6.30 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 2 | COBS CHEDDAR CHEESE GRAM | 2.85 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 3 | COBS GLUTTEN ZGRAM | 2.85 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 4 | OBS POPCORN FOR | 0.70 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 5 | COLES MWAVE POPCORN GRAM | 0.95 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 6 | KELLOGGS COCO POPS BGRAM | 8.50 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 7 | BLACKBER TES GRAM | 9.00 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 8 | STRAWBERRIES FOR | 7.50 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 9 | BLACK GRAPES PERKG | 7.80 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |
| 10 | WILLTAM BARTLT PEARS PERKG | 0.80 | 7155/in/in | Coles | 2022-04-03 | 965 - CS WESTMEAD -in/in | 0 | | 10 |

The objective of the system is by means of computer vision techniques, automatically extract the text from a receipt, the text of interest is a list of groceries with their price, quantity, etc... and have it digitally so it can be displayed in a nicely formatted table or store it in a database eventually (SQL for example). The system will use traditional computer vision techniques such as canny edge, thresholding, perspective transforms, high pass filtering to essentially make the task easy for the OCR system to properly extract the text.

The whole system can be broken into these steps:

- Loading of the image in grayscale to start processing it
- An initial contour detection to crop the receipt only and leave out the background
- An adaptive thresholding – this will sharpen the text so that the OCR system has an easier time reading small letter
- Apply the OCR, this uses an external library (tesseract OCR by Google) to get all the text in the image

For a human, it's easy to see the receipt contour but the machine always needs to be told what exactly needs to be done. For this purpose, the canny edge detector comes in handy. It is generated technique which find edges of objects. The issue is that, it cannot be applied directly. As it can be seen, there are actually a lot of edges because of the text, which will generate a lot of noise. So the text has to be erased somehow. But first it's always good to apply a small gaussian blur to remove any noise. This will make the processing afterwards easier and less error prone.



Now, the text should be erased. Because the image is in grayscale every pixel has a value between 0 and 255. 0 being black and 255 being white. The black pixels in the receipt which is text should be erased and the white pixel replace them somehow. If all the black pixels are simply replaced, the background will be also white, which is not wanted. So small regions of black should be erased.

This is simply handled by morphological operations. The principle is very simple. A structuring element must be defined. This is usually a square number of pixels. For example, 3x3:

| | | |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |

This 3x3 matrix is usually filled with 1 to tell that it is active. Then, this structuring element will be overlayed on each pixel of the image taking the maximum in the neighbourhood of 3x3. As a simple example, say the input image is:

| | | | | |
|----|---|---|---|---|
| 10 | 0 | 0 | 0 | 0 |
| 0 | 3 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |

The structuring element is first laid on the top left pixel which has a value of 10. It is partially laid outside the image. The outside elements are taken to be 0. So in this case, the maximum in the neighbours is itself, 10. So it will put 10 on every element covered by the structuring element.

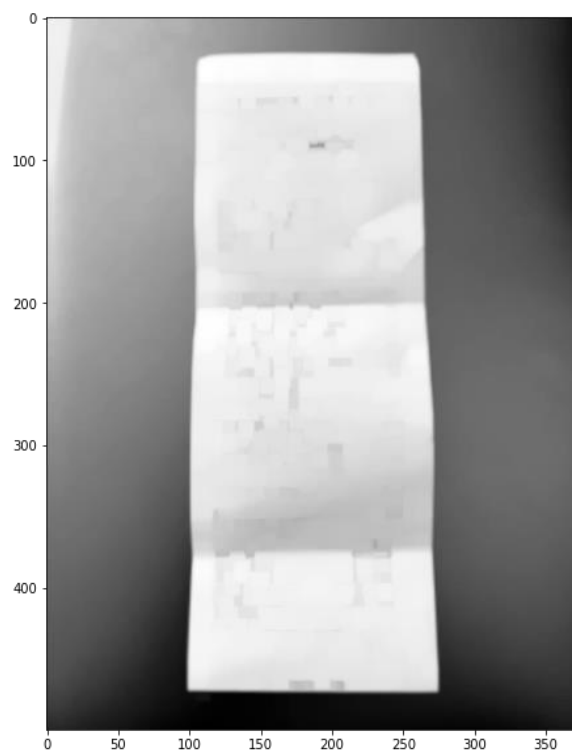
| | | | | |
|----|----|---|---|---|
| 10 | 10 | 0 | 0 | 0 |
| 10 | 10 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |

The same process is repeated for every pixel, and the results is:

| | | | | |
|----|----|---|---|---|
| 10 | 10 | 0 | 0 | 0 |
| 10 | 10 | 3 | 0 | 0 |
| 0 | 3 | 3 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |

Note that the intermediary results is not taken for the next pixel but the initial image is taken each time. Otherwise the image would be filled with the maximum value.

This has to purpose to fill the low value pixels but if a large amount of pixels is black, it stays black. Such as the background. The result on the receipt image is shown below:



In this case, a rectangular structuring element of 9x9 is used. The text has completely vanished to leave a white receipt paper. The canny edge detector can now be used to detect the outline without any difficulty.

The contour is a list of (x,y) points in the image. The contour of the receipt should be rectangular so in the ideal case, it should only be 4 points. But because of the creases and the perspective, it is not perfectly rectangular and will give more points. It is still possible approximate the contour with less points. There exist simple algorithms which will only keep the most important points and leave the points which are not characteristic. After the approximation is done, 4 points should be left.

The process has already come a long way. But this is mostly done for the computer vision preprocessing part. The OCR will take care of the rest.

Once the 4 points have been found. It is used to crop the receipt from the image. Because the 4 points, are not necessarily perfectly aligned with 90 degrees, the image has to be rotated, scaled and stretched in certain directions to get a rectangular image.



The previous steps were actually done on a smaller scale image. It was not important to see the small details. Now that the OCR must be done, it is important to take the high-resolution image again.

OCR

To extract the text, the library pytesseract is used. This is the python binding for the C++ open-source library TesseractOCR. It is the most robust, freely available open-source engine. It uses sophisticated techniques such as neural networks to give the best results.

Before giving the image to tesseract, a last processing step is done. The image is sharpened in order to read the small text. The focus on the image could have been bad, and it is hard for tesseract to read fuzzy text. The sharpening is done with a process like the canny edge detection.

| Description | \$ |
|-------------------------------|----------------|
| A2 DAIRY FULL CREAM 3LITRE | 16.50 |
| 2 @ \$8.25 EACH | |
| % COLES BETTER BAG 1EACH | 0.15 |
| KELLOGGS CORN FLAKES 725GRAM | 9.90 |
| % COBS CHEDDAR CHEESE 100GRAM | 2.85 |
| % COBS GLUTTEN FREE 120GRAM | 2.85 |
| DBS POPCORN 2 FOR \$5 | -\$0.70 |
| COLES M/WAVE POPCORN 100GRAM | 0.95 |
| KELLOGGS CORN POPS-6 650GRAM | 9.50 |
| BLACKBERRIES 125GRAM | 9.00 |
| 2 @ \$4.50 EACH | |
| STRAWBERRIES 250GRAM | 7.80 |
| 2 @ \$3.90 EACH | |
| STRAWBERRIES 250G 2 FOR \$7 | -\$0.80 |
| BLACK GRAPES PERKG | 4.48 |
| 0.895 kg NET @ \$5.00/kg | |
| WILLIAM BARTLT PEARS PERKG | 3.36 |
| 1.402 kg NET @ \$2.40/kg | |
| Total for 14 items: | \$61.24 |
| EFT | \$61.24 |
| GST INCLUDED IN TOTAL | \$0.47 |
| Coles | NSW AU |
| 03/04/22 18:05 | 20519500 99501 |
| **** 9324 | MASTERCARD |
| CREDIT ACCOUNT | CBA Credit |
| APSN 0001 ATC 0003 | A0000000041010 |
| PURCHASE | AUD\$ 61.24 |
| RRN 000010715500 | (00)APPROVED |
| AUTH 033383 | |
| NO PIN OR SIGNATURE REQUIRED | |

% = Taxable items

The text is read by tesseract and every block of text is aggregated together. Tesseract will automatically see that certain parts are lines, etc..

| Description | \$ |
|-------------------------------|----------------|
| A2 DAIRY FULL CREAM 3LITRE | 16.50 |
| 2 @ \$8.25 EACH | |
| % COLES BETTER BAG 1EACH | 0.15 |
| KELLOGGS CORN FLAKES 725GRAM | 9.90 |
| % COBS CHEDDAR CHEESE 100GRAM | 2.85 |
| % COBS GLUTTEN FREE 120GRAM | 2.85 |
| DBS POPCORN 2 FOR \$5 | -\$0.70 |
| COLES M/WAVE POPCORN 100GRAM | 0.95 |
| KELLOGGS CORN POPS-6 650GRAM | 9.50 |
| BLACKBERRIES 125GRAM | 9.00 |
| 2 @ \$4.50 EACH | |
| STRAWBERRIES 250GRAM | 7.80 |
| 2 @ \$3.90 EACH | |
| STRAWBERRIES 250G 2 FOR \$7 | -\$0.80 |
| BLACK GRAPES PERKG | 4.48 |
| 0.895 kg NET @ \$5.00/kg | |
| WILLIAM BARTLT PEARS PERKG | 3.36 |
| 1.402 kg NET @ \$2.40/kg | |
| Total for 14 items: | \$61.24 |
| EFT | \$61.24 |
| GST INCLUDED IN TOTAL | \$0.47 |
| Coles | NSW AU |
| 03/04/22 18:05 | 20519500 99501 |
| **** 9324 | MASTERCARD |
| CREDIT ACCOUNT | CBA Credit |
| APSN 0001 ATC 0003 | A0000000041010 |
| PURCHASE | AUD\$ 61.24 |
| RRN 000010715500 | (00)APPROVED |
| AUTH 033383 | |
| NO PIN OR SIGNATURE REQUIRED | |

% = Taxable items

The complete extracted is shown below:

```
RCH
~ RRN 000010718800

"Coles Supermarkets Australia Pty
Ltd
Tax Invoice ABN: 45 004 189 708

Value the Australian way

Store: 965 - CS WESTMEAD -

; Store Manager: Reza | a
}' Phone: 02 8837 7700
Served By: Hagin tne

Register: Receipt: 7155

Date: - 03/04/2022 Times 18:05
description _ eer 2 ;
A2 DAIRY. FULL CREAM SLITRE 16. 50
| 8,25 EACH a -
oe coves. BETTER BAG JEACH=0 ~~ 01
5
KELLOGGS CORN FLAKES: ST OBGRAM 6.
30
% COBS CHEDDAR CHEESE 100GRAM 2.85
|
% COBS GLUTTEN hs 120GRAM a 2.85 |
OBS POPCORN 2 FOR $5 ~$0.70

COLES M/WAVE POPCORN 100GRAM 0.95
|

KELLOGGS COCO POPS:6 B50GRAM 8.50

BLACKBER TES 125GRAM 9.00
2 @ $4.50 EACH |
| aa 250GRAM 7.80
STRAWBERRIES. 2506 2 FOR $7 | -$0.
80
BLACK GRAPES PERKG 4.48.
: 0,895 kg NET @ $5. 00/kg 7m
+ WILLTAM BARTLT PEARS PERKG ~ 3.3
6 |
1.402 kg NET @ $2.40/kg :
Total For 14 items: - $61.24
EFT $61.24
GST INCLUDED IN TOTAL $0.47

| | Coles NSW AU
| 03/04/22 18:05 20519500, /NOB501
-.

fe RR OBZ WS TERCARD
CREDIT ACCOUNT A Credit
pr eh OE ATC. 0003 soot

UD$ 61.24
(00) APPROVED
AUTH 035383
NO PIN "OR SIGNATURE REQUIRED

!

= heshig items
```


The text has to be cleaned up. Any informations which is not relevant is removed, this is usually very hard to do because it requires an understanding of written text but in the case of receipt, the text has usually the same format, and using pattern matching, it's possible to see which part represent the grocery list.

The following informations are extracted:

- The supermarket name (Coles, Woolworth, ...)
- The date
- The store (for ex. 965 – cs westhead)
- The receipt ID
- The list of products and prices
- The amounts

The final list is shown below:

| | food | Cost | Receipt_ID | Supermarket | date | Store | Processed | UserID | StoreID |
|----|-----------------------------|------|------------|-------------|------------|-------------------------|-----------|--------|---------|
| 0 | DAIRY FULL CREAM SLITRE | 8.25 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 1 | KELLOGGS CORN FLAKES OBGRAM | 6.30 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 2 | COBS CHEDDAR CHEESE GRAM | 2.85 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 3 | COBS GLUTTEN ZGRAM | 2.85 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 4 | OBS POPCORN FOR | 0.70 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 5 | COLES MWAVE POPCORN GRAM | 0.95 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 6 | KELLOGGS COCO POPS BGRAM | 8.50 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 7 | BLACKBER TES GRAM | 9.00 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 8 | STRAWBERRIES FOR | 4.50 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 9 | BLACK GRAPES PERKG | 7.80 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |
| 10 | WILLTAM BARTLT PEARS PERKG | 0.80 | 7155\n\n | Coles | 2022-04-03 | 965 - CS WESTMEAD -\n\n | 0 | | 10 |

Storing in an SQL database

The list is now digitally available and can be stored in any database. This is useful to automatically save and do statistics on the database with queries. For example, a list of shops can be queried with:

SELECT * FROM shops

| | id | name | address | postcode |
|---|----|--------------|------------------|----------|
| 0 | 1 | Woolworths | 1234 Fake street | 3000 |
| 1 | 2 | Coles | 123 Fake street | 3011 |
| 2 | 3 | Aldi | 123 Fake street | 3050 |
| 3 | 4 | Seven eleven | 123 Fake street | 3012 |
| 4 | 9 | Walmart | 1234 fake street | 3020 |

Accuracy of the code.

| Receipt ID | store | Text extraction accuracy |
|--|------------|--------------------------|
| eReceipt_3349_Dandenong_04Mar2022__ztpfx | woolworths | 95% |
| eReceipt_3395_Pakenham Market Place_07May2022__gnjwy | woolworths | 95% |
| eReceipt_3806_Clayton_06May2022__xgtgz | woolworths | 95% |
| eReceipt_3807_Dandenong South_03May2022__mrkhe | woolworths | 95% |
| eReceipt_3807_Dandenong South_05May2022__ayqyv | woolworths | 98% |
| eReceipt_3807_Dandenong South_06Apr2022__xpwwv | woolworths | 95% |
| eReceipt_3807_Dandenong South_07May2022__zgpoa | woolworths | 95% |
| eReceipt_3807_Dandenong South_09Apr2022__aofvy | woolworths | 98% |
| eReceipt_3807_Dandenong South_18Apr2022__dwlpn | woolworths | 95% |
| eReceipt_3807_Dandenong South_22Apr2022__mbvjd | woolworths | 98% |
| eReceipt_3807_Dandenong South_22Mar2022__bcecw | woolworths | 95% |
| eReceipt_3807_Dandenong South_22Mar2022__bcecw | woolworths | 98% |
| eReceipt_3807_Dandenong South_23Apr2022__kbxcw | woolworths | 90% |
| eReceipt_3807_Dandenong South_23Apr2022__wkkxy | woolworths | 95% |
| eReceipt_3807_Dandenong South_26Apr2022__btmpe | woolworths | 95% |
| eReceipt_3807_Dandenong South_28Apr2022__phzid | woolworths | 98% |
| eReceipt_3807_Dandenong South_28Apr2022__qmyau | woolworths | 98% |
| eReceipt_3807_Dandenong South_30Apr2022__gxdkn | woolworths | 98% |
| eReceipt_3807_Dandenong South_31Mar2022__cbaqb | woolworths | 98% |

The eReceipts are working perfectly with the code and has good accuracies of text extraction. The only remark is character "@" is taking as "0".

Completed Deliverables

Computer Vision and Data Engineering Team

The following is the tasks/deliverables that had been completed by the Computer Vision and Data Engineering team in this trimester.

| Task | Completed by |
|---|--|
| Research and upskilling for existing technology stack for OCR. | Sandesh, Mahalaxmi, Preet, Victor, Chiru |
| Apply OCR on some test data and extracted the output into Excel format. | Sandesh, Mahalaxmi, Preet, |
| Analysing the receipt layout, fields and the differences for Woolworths and Coles | Victor, Chiru |
| Construct a table contain all the required attributes for the extracted data from the receipt. | Victor, Chiru |
| Defines rules for extracting required data out of the receipts and stored in a pandas data frame. | Sandesh, Mahalaxmi, Victor |
| Test the OCR code using different set of receipt image in local machine | Preet, Chiru |
| Test the accuracy of the OCR code (Check the extracted text and compare with the actual | Sandesh, Victor |

| | |
|---|--|
| receipt) and record the accuracy of each receipt in Excel file. | |
|---|--|

Suggested Deliverables for Next Semester Team

- Improve the OCR model and increase the accuracy of extracted text data.
- Improve the OCR model so that it can scan every type of receipt in different angle or background.
- Improve the efficiency of extracting required data out of the receipts and stored in the database.