# Automating Receipt Data Extraction And Classification for Generating Expense Reports

Vedant Ghavate
Student,AISSMS's Institute of Information Technology
Pune, India
vedantghavate259@gmail.com

Kshipra Dhame
Student,AISSMS's Institute of Information Technology
Pune, India
kshipradhame.kd@gmail.com

Prithvi Chaudhari
Student,AISSMS's Institute of Information Technology
Pune, India
prithvichaudhari85@gmail.com

Garima Ghatge
Student,AISSMS's Institute of Information Technology
Pune, India
garimaghatge21@gmail.com

Prashant Wakhare
Assistant Professor,AISSMS's Institute of Information Technology
Pune, India
pbwakhare@gmail.com

***Abstract*— Image-based documents are converted into digital and processable forms by the use of OCR. However, present technology is unable to convert the image of a table to tabular text form like an invoice or receipt. A receipt contains important data points like product description, product value etc which can be used for many data science applications. This paper provides assessment and analysis of a system planned to extract relevant data from a receipt by enhancing OCR, a neural network is used to build a database by the process of extraction and categorisation. The database can further be used to present an expense report, it is also used in market analysis for product manufactures and to generate buyer patterns. The system implements data extraction on the generated data by utilising an already well devised OCR library.**

**The neural network for classification is trained to split data into five broad categories. The analysis of data extracted from receipt is done effectively by combining above mentioned techniques.**

***Keywords— Receipt, Optical Character Recognition, Neural Network, invoice***

## I. Introduction

When it comes to feeding tabular data, manual data entry is an exhaustive, tiring and laborious process. Each and every purchase receipt that is transacted is still manually typed worldwide. This process misses out on useful patterns and information that could be extracted and utilizes significant amounts of time and effort. This system was devised to complete the process electronically rather than manual data entry of receipts, bills or purchase orders. The system categories data and generates expense reports from databases. Instead of keeping and maintaining records of the receipt the user simply has to upload an image of the receipt to the system. The user receives an expense report at the output and they can upload multiple images at one time. OCR technologies are implemented on the image file to generate plain text files once the receipt has been uploaded to the system. [7]

On the basis of accuracy of reading Dot Matrix printed based receipts the existing OCR Technologies were compared for e.g. Tesseract, ABZZY etc. and the best one is chosen for implementation. The plain text data of the receipt is break-up and important key points such as product name and product price are extracted. The data is extracted and cleaned in the process, as different receipts will have different format and structure and they are later stored on a noSQL database. The neural network is then trained by the generated database to categorise the receipt based on item name. For example in the category of groceries, cheese is assigned as a product from the receipt.

The neural network is trained to assign five such broad categories. [2]

This system provides automated expense reports with powerful data visualisation tools while the purpose of manually adding receipts is to generate an expense report. Thus, without any human intervention solves the problem of categorising receipts also the system provides a robust solution to avoid manual data entry. Throughout the development, we have created two unconventional techniques, the categorisation of receipts items using a neural network and extraction of tabular data.

## II. Existing System

The form of receipts has changed drastically from hand written to printed receipts over the years. There are a number of systems which work on detection of text from an image, the method in [4] is one such system that extracts the date and price of items from a receipt. It has been designed for an array of bills from ranging from grocery store receipts to travel tickets. The recognition of tabular data from scanned images is an important aspect for understanding the receipts as most invoices are in tabular format. The system understands the layout of a document and labels the entities as table header, table trailer, table cells and non table regions. [6]

Receipts2go [9], a system which focuses on one page documents which does not have a fixed layout. such as receipts, invoices, tickets, price tags, etc. It aims at extracting data from images taken from a cell phone of varied camera specification. The document undergoes three stages as follows : Image Normalisation, Entity Extraction, Information Extraction and User Rule Application to extract information such as date, price, telephone and total.

## III. System Architecture Diagram

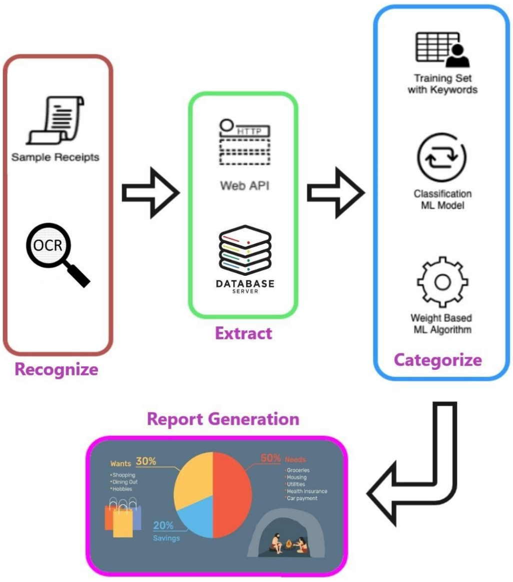


Fig. 1. System Architecture Diagram

## IV. Modules

### A. Image Pre-processing and OCR Technology Deployment

The uploaded images of receipts may be irregular with respect to brightness, contrast and other camera specific parameters. The variation in the receipt images needs to be processed. Hence, image processing techniques are applied to crop, binarize, skew the receipt images.

Understanding the limit of the receipt in an image and cropping is most important of all the image processing techniques applied to the receipt image. For adjusting the brightness of an image to a standardized limit, binarization technique is used. Skewing is used for appropriate orientation of an image. Other image processing techniques are applied to adjust the contrast, clean noise from the image, etc.[5]

After processing the image, the OCR algorithm is applied to the image. In this system, the existing OCR engines are outsourced. Four different OCR engines are studied properly in order to use the one which gives appropriate results with highest accuracy for receipt images. We have studied Tesseract, PyOCR, Transym OCR systems. Image processing and optical character recognition is provided by these OCR systems as a bundled package.

The OCR systems and algorithms were tested on 3 different parameters:

1. Reading the tabular data

2. Reading the special characters e.g. currency

3. Reading the dot matrix printer receipt

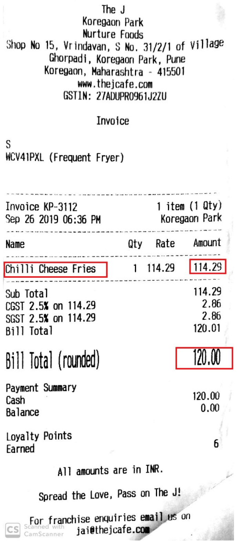
After a detailed study, we concluded that ABZZY, PyOCR, and Tesseract could be used in our system. All the three OCR systems gave nearly the same output and hence they can be used in the system. However, for this project we have used ABZZY OCR engine.

### B. Data Extraction and Pattern Recognition

The OCR converts the image to plain text. It converts the received contend into a text file. It analyses and extracts the required key value pairs like item name, amount. The OCR does not remove or add extra spaces. [2]

The information of a single product will lie on the same line. Hence, the system traverses the text file line by line. The extra unwanted information about the product such as the GST, address,etc. are removed. Then from that line the key value pairs are extracted. There are some white spaces between the key value pairs such as item name and amount. This helps us to separate the key value pairs efficiently. The algorithm works effectively on all structures of receipts. If the receipt contains items with multiple lines of product description the algorithm keeps traversing the next line until it receives entire product information. [3]

The information is not uniform at structure and fields are different from each other. Hence, a structure less database is used. This system uses the MongoDB database which is a noSQL database management system. The key value pairs that are extracted from the text file are stored in MongoDB database.



```
                   The J
               Koregaon Park
               Nurture Foods
Shop No 15, Vrindavan, S No. 31/2/1 of Village
      Ghorpadi, Koregaon Park, Pune
      Koregaon, Maharashtra - 415501
             www.thejcafe.com
          GSTIN: 27ADUPR0961J2ZU

                  Invoice

S
WCV41PXL (Frequent Fryer)

-----------------------------------------------
Invoice KP-3112                 1 item (1 Qty)
Sep 26 2019 06:36 PM             Koregaon Park
-----------------------------------------------
Name                     Qty    Rate    Amount
-----------------------------------------------
Chilli Cheese Fries        1  114.29    114.29
-----------------------------------------------
Sub Total                               114.29
CGST 2.5% on 114.29                       2.86
SGST 2.5% on 114.29                       2.86
Bill Total                              120.01

Bill Total (rounded)                    120.00

Payment Summary
Cash                                    120.00
Balance                                   0.00

Loyalty Points
Earned                                       6

            All amounts are in INR.

        Spread the Love, Pass on The J!

      For franchise enquiries email us on
              jai@thejcafe.com
```

Scanned with CamScanner

Fig. 2. Data extraction from Receipt

The data extraction part takes place in PHP where the insert queries of MongoDB are run through PHP script. After uploading the receipt image on the web server, the processes of data extraction and database creation takes place. This stored data is later used for classification of receipts and report generation, and can also be used by a third party for research.

### C. Classification Using Neural Network

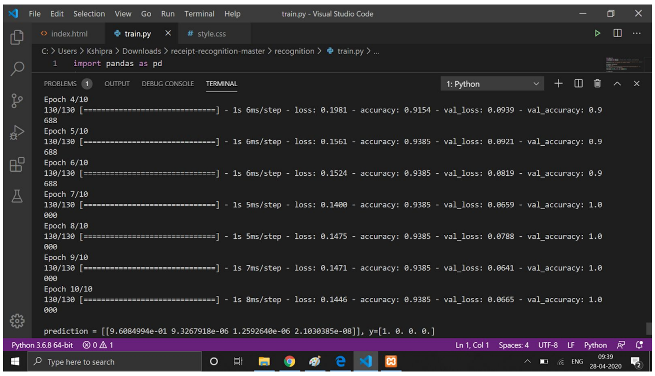
The extracted data from receipts is used to build a model to categorise the products using the product name into different categories namely travel, groceries, miscellaneous. A neural network is trained using a dictionary which consists of words and their meaning.

The neural network contains ten layers that include four convolution layers, one dropout layer, two pooling layers, one embedding layer, one sigmoid layer to convolve the output tensors. The convolution layers use ReLU function for activation. [3]

The model is compiled with the categorical cross entropy as parameter. Around 20 epochs are trained per node for a test dataset of nearly 150 values.

A training dictionary of 150 values of item name and category was created. Generalised item names like chilli, cheese, potato were used. These words can thereafter be used for classifying other products such as chilli fries, cheese fries, potato cheese fries, etc. As the layers in the neural network get trained, this dictionary can be used to classify various other collections. The layers in the model are:

1. Input image
2. Convolution layer
3. Non-linearity
4. Pooling layer


File Edit Selection View Go Run Terminal Help
train.py - Visual Studio Code
index.html
train.py
style.css
C: > Users > Kshipra > Downloads > receipt-recognition-master > recognition > train.py > ...
PROBLEMS 1
OUTPUT
DEBUG CONSOLE
TERMINAL
1: Python
Python 3.6.8 64-bit ⊗ 0 ⚠ 1
Ln 1, Col 1 Spaces: 4 UTF-8 LF Python
Type here to search
ENG
09:39
28-04-2020

```
1   import pandas as pd
```

```
Epoch 4/10
130/130 [==============================] - 1s 6ms/step - loss: 0.1981 - accuracy: 0.9154 - val_loss: 0.0939 - val_accuracy: 0.9
688
Epoch 5/10
130/130 [==============================] - 1s 6ms/step - loss: 0.1561 - accuracy: 0.9385 - val_loss: 0.0921 - val_accuracy: 0.9
688
Epoch 6/10
130/130 [==============================] - 1s 6ms/step - loss: 0.1524 - accuracy: 0.9385 - val_loss: 0.0819 - val_accuracy: 0.9
688
Epoch 7/10
130/130 [==============================] - 1s 5ms/step - loss: 0.1400 - accuracy: 0.9385 - val_loss: 0.0659 - val_accuracy: 1.0
000
Epoch 8/10
130/130 [==============================] - 1s 5ms/step - loss: 0.1475 - accuracy: 0.9385 - val_loss: 0.0788 - val_accuracy: 1.0
000
Epoch 9/10
130/130 [==============================] - 1s 7ms/step - loss: 0.1471 - accuracy: 0.9385 - val_loss: 0.0641 - val_accuracy: 1.0
000
Epoch 10/10
130/130 [==============================] - 1s 8ms/step - loss: 0.1446 - accuracy: 0.9385 - val_loss: 0.0665 - val_accuracy: 1.0
000

prediction = [[9.6084994e-01 9.3267918e-06 1.2592640e-06 2.1030385e-08]], y=[1. 0. 0. 0.]
```

Fig. 3. Training neural network

Convolutional layers summarize(convolve) along one or n dimensions. After a non-linearity e.g. ReLU is applied to feature maps that is the output of the convolution layer, a pooling layer is added. For ordering the layers in a convolutional neural network, the pooling layer is used. Two types of pooling layers are used in the system:[6]

1. Max pooling layer: The largest or maximum value in every patch of each feature map is calculated by the max pooling layer.

2.Global pooling layer: To summarize the presence of a feature in an image, a global pooling layer is used.

The model is trained by a dataset having two categories namely number and item name which consists of 150 such values. This model is used for testing the newly created database, and hence the results for every collection are saved.[7]

### D. Report Generation and Data Visualization

Report generation from the extracted data is the only and important output of the project. The expense report generation is directly shown after the set of receipt images are uploaded. The other processes of data extraction from receipts, database creation, classification of receipts are done in the background. [2]

Various reports along with their purposes were studied thoroughly. The project's aim is to help the user manage the overall expenses. Hence, representation of the expenses and the category of items were concluded to be most significant to the user. The MongoDB database aggregates the categorized information and the information stored in the existing collection.

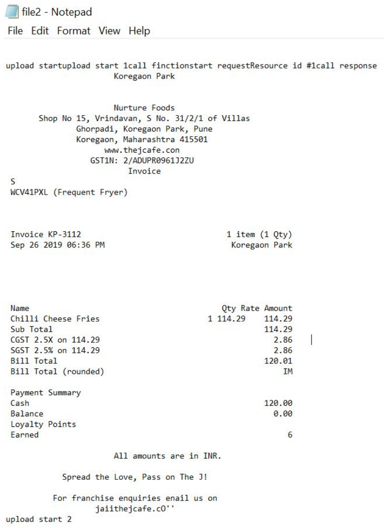
To help the user in budget management and to provide the user with an overview of his expenses, the expense report is represented category wise. The canvas Javascript API creates effective data visualization. It uses the data stored after classification of receipts and the data extracted from the receipts stored in the MongoDB database. [1]

Data visualization is the key part of our system. In our system, the user is able to view his expenses category wise by the means of Pie Chart. When a set of receipt images are uploaded, the extracted data of those receipts is aggregated and classified, and the result is represented in the form of an expense report.

## V. Result

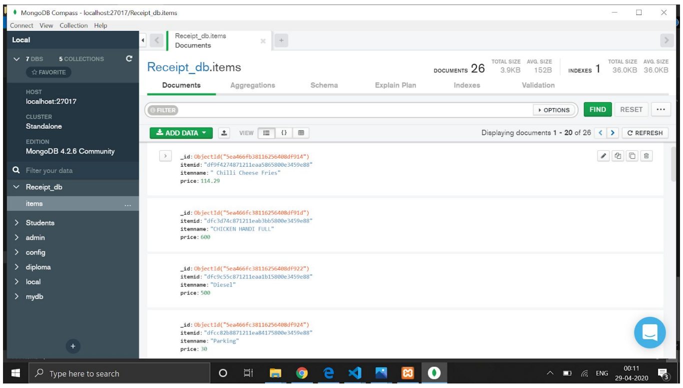
The system gives appropriate results for data extraction and classification of receipts. The system achieves the desired output provided that the receipts have a proper and visible print. However, we encountered some issues for receipts having multiple lines of product information.[4]

A dataset of around 60 receipts was used. The system successfully detected the product name and amount for 38 receipts.



```
upload startupload start 1call finctionstart requestResource id #1call response
                      Koregaon Park


                      Nurture Foods
       Shop No 15, Vrindavan, S No. 31/2/1 of Villas
               Ghorpadi, Koregaon Park, Pune
               Koregaon, Maharashtra 415501
                     www.thejcafe.con
                  GST1N: 2/ADUPR0961J2ZU
                         Invoice
 S
 WCV41PXL (Frequent Fryer)


 Invoice KP-3112                                 1 item (1 Qty)
 Sep 26 2019 06:36 PM                             Koregaon Park



 Name                                          Qty Rate Amount
 Chilli Cheese Fries                          1 114.29    114.29
 Sub Total                                                114.29
 CGST 2.5X on 114.29                                        2.86
 SGST 2.5% on 114.29                                        2.86
 Bill Total                                               120.01
 Bill Total (rounded)                                         IM

 Payment Summary
 Cash                                                     120.00
 Balance                                                    0.00
 Loyalty Points
 Earned                                                        6

                      All amounts are in INR.

            Spread the Love, Pass on The J!

          For franchise enquiries enail us on
                  jaiithejcafe.cO''
upload start 2
```

Fig. 4. Text file generated using OCR



```
_id: ObjectId("5ea466fb38116256408df914")
itemid: "df9f4274871211eaa5865800e3459e88"
itemname: " Chilli Cheese Fries"
price: 114.29

_id: ObjectId("5ea466fc38116256408df91d")
itemid: "dfc3d74c871211eab3bb5800e3459e88"
itemname: "CHICKEN HANDI FULL"
price: 600

_id: ObjectId("5ea466fc38116256408df922")
itemid: "dfc9c55c871211eaa1b15800e3459e88"
itemname: "Diesel"
price: 500

_id: ObjectId("5ea466fc38116256408df924")
itemid: "dfcc82b8871211ea84175800e3459e88"
itemname: "Parking"
price: 30
```

Fig. 5. Database with noSQL (MongoDB)

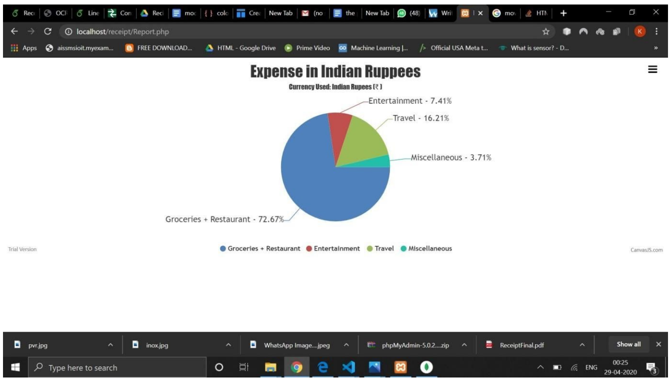


Fig. 6. Categorised Report

## VI. Conclusion

This system is capable of automatically extracting information from the receipt images irrespective of its structure and converting it to a categorized expense report. We have implemented two independent technologies, data extraction from receipt images and classification using a neural network. [6]

Plain optical character recognition doesn't work on tabular data and hence this system is specifically designed to deal with the tabular data on receipts. A dictionary is used to train the neural network to classify and categorise the receipt information. [8]

This system reduces the manual feeding of receipt data into a computer system and automatically processes the image to build a database.

Other applications can use the techniques implemented in this system with optimized functionality and accuracy.

## REFERENCES

[1] A. Coates et al., ”Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning,” 2011 International Conference on Document Analysis and Recognition, Beijing, 2011, pp. 440445.

[2] Pise, A., & Ruikar, S. D. (2014, April). Text detection and recognition in natural scene images. In Communications and Signal Processing (ICCSP), 2014 International Conference on (pp. 1068-1072). IEEE.

[3] Rohit Verma and Dr.Jahid Ali, “A-Survey of Feature Extraction and Classification techniques in OCR Systems.” Proceeding of the international journal of Computer Application and Information Technology, Volume 1, Issue 3, November 2012.

[4] Alex Yue, ”Automated Receipt Image, Identification, Cropping and Parsing”

[5] Miguel E. Ruiz, Padmini Srinivasan, ”Automatic Text Categorization Using Neural Networks”School of Library and Information Science, The University of Iowa

[6] Anukriti Bansal, Gaurav Harit, Sumantra Dutta Roy, ”Table Extraction from Document Images using Fixed Point Model”

[7] Wang, G., Hoiem, D., Forsyth, D. (2009). Building text features for object image classification. 2009 IEEE Conference on Computer Vision and Pattern Recognition.

[8] Gao, S., Wang, Z., Chia, L.-T., & Tsang, I. W.-H. (2010). Automatic image tagging via category label and web data. Proceedings of the International Conference on Multimedia - MM ’10

[9] Bill Janssen, Eric Saund, Eric Bier, Patricia Wal,l Mary Ann Sprague,”Receipts2Go: The Big World of Small Document