PAYMENT RECEIPTS VALIDATION THROUGH DUPLICATE ELIMINATION USING OPTICAL CHARACTER RECOGNITION

Vinoch Selvarathinam

(189355L)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

July 2020

PAYMENT RECEIPTS VALIDATION THROUGH DUPLICATE ELIMINATION USING OPTICAL CHARACTER RECOGNITION

Vinoch Selvarathinam

(189355L)

Dissertation submitted in partial fulfilment of the requirements for the degree

Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

July 2020

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another per-son except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: ……………… Date: ……………….

Name: S.Vinoch

The supervisor/s should certify the thesis/dissertation with the following declaration.

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for MSc Thesis.

Signature of the supervisor: …………………………. Date: ………………..

Name: Dr. Indika Perera

ACKNOWLEDGEMENTS

My sincere appreciation goes to my family for the continuous support and motivation given to make this thesis a success. I also express my heartfelt gratitude to Dr. Indika Perera, my supervisor, for the supervision and advice given throughout to make this research a success. I also thank my parents, brothers for their heartfelt support. Last but not least I also thank my friends who supported me in this whole effort.

ABSTRACT

Optical Character Recognition (OCR) is the method of digital image retrieval of the characters. The idea behind OCR is to obtain an image or pdf format document and extract the characters from that image and present it in an editable format to the user.

This thesis focused on research related to extract the information such as vendor name, category of the receipt (food related, travel related etc.) and amount from the receipt which can be printed and hand written. Further to identifying the mentioned information, expanded the research on identifying the duplicate receipts as well.

Petty Cash is an accessible store of money kept by organizations for expenditure on small items. When an employee wants to reimburse the amount that he/she spent, they need to fill a voucher with the date of the expense, amount, vendor, reason of the expense and attach the supporting documents (receipts) which will consume papers. In this digital world, easily we can automate this process using digital platforms and tools.

Mobile phones are significantly playing major roles in our day-to-day life more than ever and the usage of mobile phones are increasing drastically compare to desktop computers. In order to reduce the carbon foot print, we can take necessary steps to reduce the paper usage. Building a mobile application which can automate the petty cash process which includes OCR capability on receipts would engage the users to use it in their organizations.

This is the first time, OCR on receipts and duplicate identifier is researched and done. There are no researches conducted on this.

Keywords: Image Processing, Optical Character Recognition, Neural Network

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
OCR	Optical Character Recognition
CAGR	Compound Annual Growth Rate
CAPTCHA	Completely Automatic Public Turing Test to Tell Computers and Humans Apart
ANPR	Automatic Number Plate Recognition
GISMO	Geographic Information Systems and Mapping Operations
ANSI	American National Standards Institute
DIA	Document Image Analysis
DPI	Dots Per Inch
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
GRU	Gated Recurrent Unit
REGEX	Regular Expression
CNN	Convolutional Neural Network
ETL	Extract, Transform and Load
CV	Custom Vision

1. INTRODUCTION

Optical Character Recognition (OCR) is a software which converts the printed text and images into digitized form so that they can be comprehensible by machine. Human brain has the ability to understand and identify a picture's text or characters, but computers are not that smart enough to discern the details in the picture. Work has also been carried out to turn a document image into a computer comprehensible format. [1]

1.1 Use of petty cash in the organizations

Petty cash represents small transactions both in materiality and quantity. The word "petty" is used because cash or cash-like transactions are small (small) amounts. "Petty" is no longer so small, so think of it as small quantities that you might not be worried about monitoring.

For example:

1. Buying stationaries for the office and office staffs.
2. Travelling to a location for a meeting via cabs.
3. Treating employees to have dinner to celebrate an accomplishment.
4. Late night stay at office for an important work and pay for dinner.

A small cash fund will be available on most of the office, especially retail businesses which may have customers paying in cash. The cash envelope is used for making change, and in some cases for small business transactions. The box usually begins with a cash balance, in appropriate amounts. When the cash box gets tiny this is then refilled. When using cash box funds to make small transactions, keep a log in to track these transactions.

1.2 Cash frauds in petty cash

There are a variety of ways an person can commit fraud by robbing a business of cash. Since cash is effectively untraceable once stolen, anybody planning to steal assets would concentrate particularly on this form of asset. Below are some ways of committing cash theft in petty cash statements.

1. Over billing – Over billing is a method where an employee submits an inflated or altered invoice for payment. By submitting such receipts, employees can claim more than what they have spent actually. E.g. Travelled to a location for a business meeting and it costed about Rs.300 for taxi. Since there are no invoice presents for taxi rides, employee can claim more than Rs.300 and organization has to pay that.

2. Duplicate receipts submissions – Employees in the organization can use their relationship to do few frauds using the inner trust. E.g. Employee X will submit for a petty cash claim by attaching a receipt and which will be approved by his/her supervisor and cashier will provide the money. Meanwhile, Employee X will provide a copy of the same receipt to Employee Y where Employee Y will also submit the same for claim. In such situation, Employee X and Y's supervisors will not be aware of this and they will approve the requests and these frauds will go undetected until auditors find out by manual check.

3. Falsification of identity – In this case if the cashier is new to the organization, he/she might not know all of the employees. An employee will pretend like he/she is the requester of a particular claim and they do the sign and collect the cash from the cashier. This fraud is really rare since this can be detected easily.

4. Petty cash theft – One of the best ways to get rid of cash is to take cash out of the tiny cash box unguarded. Another choice is to rob the entire box making sure all cash and coins are removed.

1.3 Research Problem

Organizations are focused on their business expansion where they do not see the actual value of saving money by reducing unwanted expenses in small items. Petty cash is an important part of the process in many organizations and each employee will reimburse at least once in their working period in that particular organization. Many employees face hassle with this petty cash such they should always remember to get a receipt for their expense, always keep the petty cash money from their personal money and make copies of the receipt incase if they lose the original receipts. When employees want to reimburse their money back, they need to fill vouchers, get approval from their supervisors and collect money from cashier and if the cashier does not have money to provide, they have to wait. These time-consuming tasks which reduce the productivity of the employees. Organizations also face issues when employees submit duplicate receipts which slip through finance and money will be given twice for same receipt.

1.4 Main Challenges in Optical Character Recognition

There are no researches conducted on identification of duplicate receipts till now. This is a main challenge on comparing the work done previously by others and the improvements to be done in this research.

OCR is also used to retrieve text from image-only files for categorization purposes. There are, however, many OCR limitations that can lead to imprecise or incomplete text that makes text-based classification difficult or impossible:

1. Font Size - OCR can not shift over characters of extremely large or unusually small sizes in text type. This is inaccessible the key imperative characters and terms for text-based structures.
2. Unidimensional - For OCR, single words have one metric, often they are before or after other words. OCR does not catalog page coordinate information for characters while page coordinates may be useful for classification and attribute extraction.
3. Sequential Editing - Classically, OCR errors have to be corrected consecutively, with frequent editing of the same errors. Certain errors that may be ignored might be created by global spell checks.
4. Case Sensitivity for Editing - Using spell checking to correct OCR text does usually not allow for consideration in the case in letters, e.g. dog and DOG will be handled equally.
5. Languages - Many languages have special characters, and if the right OCR program is loaded those characters may be lost or incorrectly recognized.
6. Non-Symmetrical DPI for Faxes - Faxes are sometimes stored in files where the number of dots per inch horizontally is not the same as the Density Per Pixel (DPI) vertically, and OCR engines may be struggling with this non-symmetrical DPI.

7. Partial Text - Writers of the studies also add illustrations with clear text. However, OCR software can detect some text, assume that OCR is not necessary, and skip processing the document, making the text in the picture invisible for text search or analysis only. A similar thing will happen when PDFs are added to document headers, footers, or legends. OCR systems may detect the presence of a text layer and do not attempt to convert the image layer, even though it may have the most significant content.
8. Non-Textual Glyphs - There are important non-textual characters which are not translated by OCR to characters, leaving them unnoticed for text analysis or text-based retrieval. Some examples include logos and map symbols.
9. Inferring the Obvious - Graphic elements also provide the most obvious clues about how to define a file or document, such as the position and size of logos or text blocks. Since these graphical elements may not be specifically accessible in text-restricted systems, it is up to them to try to decide what is most visible to anyone who is looking at the details.
10. Incorrect Document Boundaries - Image-only files often contain several documents per file, and OCR provides no way to rectify document boundaries. This creates problems with programs which recognize files based on similarities between the words used in documents. Losing secret documents and misclassifying which ones are classified is easy.

Other than the above, there are issues when the paper quality is less, distortion form binding, bleed through / shine through, poor inking, obsolete fonts, annotations by users and lack of dictionaries / bad spelling.

1.5 Motivation for the research

With the mentioned challenges with using OCR, it is more difficult when it comes to scanning the receipt and identifying the relevant information including identifying the duplicate receipts.

In organizations there are lot of pain arears which could be addressed nowadays by digital platforms and tools. There are applications which addresses the issues related to petty cash process in the organization and which is costly but improves the process automation. None of the products in the current market addresses the main problem with petty cash, is preventing fraudulent activities / transactions before it occurs.

1.6 Application of OCR and duplicate identifier process

Optical Character Recognition is widely used nowadays for different purposes and in different applications. [2]

1. Banking – OCR is used without a human interaction to treat cheques. A check can be embedded in a system where the mechanism filters the amount to be issued and exchanges the appropriate sum of the cash. This invention has been idealized for written checks, and is fairly reliable to write by hand checks, as well as to minimize the holding time in banks.

2. Legal Industry – Additionally, Legal Industry [3] is one of the beneficiaries of OCR innovation. OCR is used to digitize documents and is stored in a computer database explicitly. Legitimate experts can support the look records needed from colossal databases by writing some watchwords in essence.

3. Invoice Imaging – For other company applications, Receipt Imaging is commonly used to keep track of financial documents and to predict accumulation of increments from heaping up.

4. Captcha – A CAPTCHA [4] is a program that can build and check checks that can be passed by humans but are not feasible by current computer software engineers. Hacking may pose a real danger to the use of the network. Today, most of the human activities such as financial transactions, instruction confirmation, enlistments, travel bookings etc. are conducted on the web and all this includes a watchword that programmers misuse. In CAPTCHA, a picture is created consisting of arrangement of number letters, which is obscured by techniques of image distortion, variety of estimation and text design, diverting foundations, random parts, highlights and commotion within the picture.

5. Handwriting Recognition – Handwriting recognition [5] is a computer's ability to get and decode manually written data from sources such as paper records, images, touch-screens, and other gadgets consistently. The picture of the written material can be sensed "off track" by optical screening or clever word recognition from a chunk of paper. Then again, the write tip innovations can be observed "on paper," for a pen-based computer screen surface to highlight them.

6. Automatic Number Recognition – Automatic number plate recognition [6] is used as a strategy for mass observation which allows use of optical character recognition on images to differentiate vehicle registration plates. In addition, ANPR was generated to store the images recorded by the cameras, counting the numbers recorded from the permit plate.

ANPR innovation possesses plate diversity from place to place as it may be an innovation that is unique to the region. They are used by different police forces and as a method of collecting electronic tolls on pay-per-use roads and cataloging crime or people innovations.

7. Finance, Petty Cash – In this case, where the receipts are recognized using OCR and identifying the duplicate receipts to prevent unnecessary petty cash claims.

1.7 Objective of the research

This research objective is to identify such issues mentioned in section 1.2 and find an automated solution to improve employee's productivity, reducing paper consumption and saving money to organizations by reducing duplicate receipts submissions by employees. The solution is a mobile application which includes built-in OCR reader capability on receipts and identify the vendor, amount and what type of expense (Food, Travel and etc.). This mobile solution automates the entire petty cash and cash advance processes of an organization and integrates to their finance system as well. By identifying duplicate submission of receipts, Supervisors/Approver will be notified/alerted with the actual receipt submitted by the employee and the receipt which the system compared and flagged as duplicate. With this visibility Supervisors/Approvers can take decision-based approval from anywhere anytime.

1.8 Contribution of the research

This research required a large dataset which will be given out for other researchers to re-use it to accomplish the same and more. Since this is the first-time research which is conducted on identifying duplicate receipts, this research will be focused on that scenario only. However, in future this research can be expanded more in identifying the sudden deviation of petty cash claims to detect the anomalies and identifying the category of receipts by extracting the line items in the receipt.

2. LITERATURE REVIEW

2.1 Necessity of Petty cash in organizations

Petty cash is a small sum of cash on hand used to pay the reported small sums, rather than to write a check. Petty cash is also classified as a petty cash reserve. The person responsible for the little cash is known as the custodian of the little cash. Petty cash operates on the basis where an initial amount of cash is placed into an account that is drawn on for a specific purpose. If the account goes below a given number, the machine will be recharged.

In the petty cash process, there are couple of guidelines shall govern for all petty cash accounts such as;

1. The petty cash account will be constrained to sums authorized by Finance Department.
2. The department supervisor shall defend or safeguard over the petty cash fund.
3. It is entirely disallowed to cash checks from the petty cash fund.
4. Requests for reimbursements from trivial cash must incorporate the first receipts or solicitations.
5. Petty cash may be utilized for little buys, reimbursements, postage, grants, and expenses. Any exemptions must be pre-approved by the Back Office.
6. Petty cash vouchers may be requested from Back Office.

2.2 History of Optical Character Recognition

Recognition of character is not a new problem but its origin can be traced back to machine developments some time ago. The most punctual OCR systems were not computers, but mechanical devices that could identify characters, but outstandingly good speed and moo accuracy [1]. M. 1951. Sheppard designed a GISMO per user and robot which can be considered the timeliest work on cutting edge OCR [7]. GISMO will study both melodic documentations and words one by one on a printed sheet. It will recognize 23 characters in any situation. The computer also has the ability to appear to repeat a written page of some kind. K. In 1954, Rainbow developed a computer that could research capitalized type, one per diminutive, composed English characters. The early OCR implementations have been criticized for errors and slow level of recognition. Therefore, not much research focus was placed on the subject in the mid-60s and 70s. The only advances on government organizations and massive organizations such as banks, daily papers and carriers etc. were made.

Because of the uncertainties associated with recognition, three oughts were felt to standardize OCR textual types to promote OCR's acknowledgment errands. Subsequently, ANSI and EMCA produced OCRA and OCRB in 1970, which was granted comparatively worthy levels of recognition [8]. Major enquiries have been made on OCR over the past thirty years. This has powered the emergence of multi-lingual, transcribed, and omni-font OCRs for document image analysis (DIA) [8]. Despite these wide-ranging inquiries regarding projects, the capacity of the computer to accurately analyze material is still well below the individual. The current OCR analysis is then carried out on going forward OCR precision and pace for various fashion records printed / written in unconstrained circumstances. No opensource or commercial software accessible for complex dialects such as Urdu or Sindhi etc. has been available.

2.3 Growth of Optical Character Recognition

In a profoundly divided and competitive worldwide showcase for optical character recognition (OCR), adroit players are leveraging a multi-pronged technique to surge ahead of their competitors. One technique well known with them is competitive estimating which makes a difference to draw more customers. Other than, they are moreover managing an account upon mergers and acquisitions and carefully-considered associations to grow their outreach. Expansion of unused administrations is another way in which they are looking to advance improve their showcase offers.

Optical character recognition is an innovation which has truly revolutionized the report administration prepare in different divisions, specifically legitimate, managing an account, instruction, healthcare, finance, and government. It has made a difference to convert workplaces into a paperless one by making a difference to digitize archives and communications that have remained safe to digitization with speed and exactness. Those incorporate filtered paper reports and PDF records, or pictures captured by an advanced camera.

Further, the headway of optical character recognition from a uncommon reason to a multi-purpose intelligently framework has brought down the fetched of information capturing and has cleared the way for the improvement of more dependable systems. This will too have a positive effect on the advertise.

As per Straightforwardness Advertise Inquire about, the worldwide optical character recognition showcase will likely be ended up worth US$25.182 billion by 2025-end by clocking a strong 14.8% CAGR from 2017 to 2025.

the worldwide optical character recognition advertise can be classified into two based on sort – computer program and benefit. The computer program fragment assist incorporates desktop-based OCR, portable based OCR, cloud-based OCR, and others such as clump OCR, server-based OCR, etc. The benefit section, on the other hand, incorporates counseling, outsourcing, and usage and integration.

Between computer program and administrations, the two sections within the worldwide optical character recognition advertise, computer program accounted for greatest share. Within a long time ahead as well, program section will hold its prevailing share buoyed by the keeping money segment in specific. In terms of development rate, however, the benefit segment is anticipated to witness a better CAGR within the figure period.

From a topographical point of view, North America held most extreme share within the worldwide showcase for optical character recognition in 2016 since of the fast framework advancement and changing arrangements and controls framed by the government. Within the up and coming a long time, the locale is slated to hold on to its driving position. Endeavors within the locale are moreover selecting for cloud based optical character recognition program since of its ease of utilize and fetched adequacy.

Vis-à-vis development rate, Asia-Pacific optical character recognition showcase is slated to outpace all others by enrolling a CAGR of 15.6% from 2017 to 2025. The development is basically driven by little and medium undertakings (SME) mushrooming within the locale. Other than, the developing utilize of optical character recognition program by IT and telecom industry for report administration is serving to act as a catalyst for the Asia Pacific showcase.

Fast-expanding creating countries of China and India are at the forefront of receiving OCR within the locale on account of the gigantic speculations in innovation.

A few of the key players working within the worldwide advertise for optical character recognition are Anyline GmbH, ABBY Computer program Ltd., Adobe Frameworks Joined, ATAPY Computer program, CCi Insights Co. Ltd., Creaceed S.P.R.L., CVSION Innovations Inc., Exper-OCR Inc., Google Inc., LEAD Advances Inc., I.R.I.S.S.A. (Rule), IBM Enterprise, Microsoft Enterprise, Subtlety Communications Inc., NTT Information Enterprise, Paradatec, Inc., Prime Acknowledgment Enterprise, Ripcord Inc., Transym Computer Administrations Ltd., Dark Ice Computer program LLC, SEAL Frameworks, Ricoh Bunch, and Accusoft Organization. [9]

2.4 Types of Optical Character Recognitions System

There have been several bearings in which OCR inquiries have been performed in the midst of a long time ago. As a consequence of these inquiries this section talks about distinctive kinds of OCR structures have grown. These structures can be classified based on the mode of picture procurement, character network, font constraints, etc. The OCR frameworks can be classified as handwritten recognition and machine-printed character recognition based on the sort of data. The previous problem is somewhat less complex because characters are, as a rule, uniform dimensions and it is possible to predict character positions on the page. [8]

Handwriting character identification can be an extremely intense task due to the different client's composing fashion as well as the client's distinctive writing innovations for the same character. These frameworks can be split into two sub-categories, i.e. frameworks online and off-line.

The preceding is performed in real time while the character is created by the clients. These are less complex as they can capture the transient or time-based data i.e. distance, distance, number of strokes produced, stroke composition heading etc. There is no need for diminishing methods in expansion, as the write follows are few pixels long. The systems of offline recognition operate on static information, i.e. the data can be a bitmap. Subsequently acknowledgement is particularly problematic. [10]

2.5 Text identification and extraction from Image using Optical Character Recognition

2.5.1 Histogram based approach

The histogram-based approach was used to extricate bimodal images from records. On the histogram of the bimodal image, hereditary calculation is linked to extricate useful data from the basis. The histogram can be a plot or chart of the recurrence of each gray level occurrence in image over gray scale values. Let an arbitrary N measure populace be initialized and the part get an estimate of between 255. The operations are performed over two guards who are haphazardly selected. The appropriate values of the hybrid probability and the potential to shift are resolved. Each contest winner (the one with the finest wellness) is chosen. After measuring off-spring wellness values, methodology of competition choice is used to allow off-springs to compete with the guardians. A competition is held among the haphazardly selected people and the fittest is selected between the two. Hereditary calculations make use of this technique for population preference. Two guardians participate, and two off-springs. [11]

2.6 Techniques in Optical Character Recognition

The most important stages, and optical character recognition technologies. These steps include pre-processing, separation, standardization, future extraction, classification, and post planning. In order to prepare a convincing OCR-related application, we must consider the obstacles that will arise in each stage in order to achieve a high recognition rate for characters. [12]

2.6.1 Pre-processing Phase

The purpose of pre-processing is to dispense with undesired features or commotion in a picture without losing any noteworthy detail. Preprocessing procedures are required on images containing content and/or design in color, gray-level or double archive. Since color pictures are computationally more costly to prepare, most applications in character recognition frameworks use double or dim images. Preprocessing diminishes the clamor and contradictory knowledge. It enhances the image and planes it in OCR stages for further stages. Ready to improve the efficiency and ease of preparing a picture in other stages by translating the picture into the appropriate structure within the primary stage of preprocessing. In this way, the most important issue in the preprocessing stage is to diminish the commotion that triggers the lessening within the levels of character recognition.

Processes	Description
Binarization	Separates image pixels as text or background.
Noise Reduction	Better improvements of image acquisition devices produced by the advancements in technology.
Skew Correction	Because of the possibility of rotation of the input image through captured image device, document skew should be corrected.
Morphological Operations	Adding or removing pixels to the characters that have holes or surplus pixels.
Thresholding	For an image, separating information from its background.
Thinning and Skeletonisation	thinning process is the Skeletonisation, which regularize the map of the text until reaches most medial one pixel width

Table 2.1: Some important pre-processing operations

Thus, because preprocessing regulates the suitability of the input for the consecutive stages, the pre-processing step is a main stage prior to the extraction of the element. Many of the problems listed in the section of OCR Problems need to be centered in the preprocessing stage. The following may be listed: binarization, noise reduction, skew correction, morphological operations, slant elimination, filtering, thresholding, smoothing, compression, and thinning.

2.6.2 Segmentation Phase

The basic and principal component of an optical character recognition (OCR) device is segmentation of the text line from images. Text segmentation from a text image usually fuses row segmentation, word segmentation and then character segmentation. Segmentation is the method of isolating a part of the text from the image context inside a document. To properly reorganize the editable text lines from the recognized characters, first segment the text line, then segment the words from the segmented line, then segment the characters. Data segmentation is an important pre-processing phase when introducing an OCR program. This is the method of classifying a document image into homogeneous zones, i.e. each zone contains only one form of material, such as text, a map, a table or a halftone image. In some instances, the precision rate of the OCR-related systems is highly dependent on the accuracy of the algorithm used for page segmentation.

Below are 3 classes of record segmentation algorithms [13]:

1. Top-down methods
2. Bottom-up methods
3. Hybrid methods

In a portion of a text the top-down approach repetitively divides wide area into smaller sub-regions. The document segmentation process should end when conditions are met, and the ranges obtained at that stage reflect the results of final segmentation. But, approaches from the bottom up start by searching for interesting pixels and then group pixels. Instead we handle certain pixels of interest in linked components that make up characters that are then incorporated into words, and lines or text blocks. Hybrid methods are called both the top-down and bottom-up phase synthesis.

Throughout the last decades, several segmentation approaches have already been proposed through different facets of the OCR process. Satadal Saha, Subhadip Basu, Mita Nasipuri and Dipak Kr Basu have suggested a novel Text Segmentation technique based on A Hough Transform. [14]. S. Basu et al researched extracting Text line from many distorted handwritten text images. [15].

The suggested technique for eliminating text lines incorporates a water flow technique with high rate of performance. A. Khandelwal et al [16] proposed a technique by examining connected components of the neighborhood from unconstrained handwritten document images on text line segmentation. Shinde, D and Archana A. In G. Chougule had also proposed a segmentation technique [17]. We presented that the modern method of projection of vertical and horizontal profiles renders text easily segmented into lines and phrases. They announced experimental findings with a 98 per cent accuracy in line and word segmentation.

2.6.3 Normalization Phase

Due to the segmentation method isolated characters that are ready to move through the application extraction stage are collected, the isolated characters are then minimized to a particular size depending on the algorithms used. The segmentation approach is important, because it converts the image into the format of the m*n matrix. These matrices are then usually standardized by minimizing the size and removing redundant information about the image without losing any significant information [18].

2.6.4 Feature Extraction Phase

Extraction of the feature is the process of extracting unique characteristics from objects or alphabets to create vectors of the function. Then, classifiers use these vectors of features to describe the input unit with objective output. By looking at these features, classification between dissimilar classes is effortless for the classifier, as deciding [19] is fairly straightforward. This indicates other methods to derive characteristics from the segmented characters of the literature.

U. Pal et al [20] proposed characteristics and zoning of the directional chain code and a feature vector of length 100 for handwritten numeral recognition, and provided a high degree of recognition accuracy. But the procedure for extracting the function is time-consuming and complex [19]. Dinesh et al [21] suggested end points as potential characteristics for horizontal / vertical stroke recognition and use and for handwritten Kannada numerals obtained 90.50 percent accuracy in recognition. But this approach uses the process of dilution which leads to a certain loss of features [19].

E. Srinivasan et al [19] suggested extraction of diagonal-based features using neural network for the method of handwritten alphabet recognition. Sharma, Om Prakash et al [22] proposed improved zone-based hybrid extraction model using the Euler number for handwritten recognition of alphabets. After Suen [23], there are two major character groupings: statistical characteristics and structural features. In a character matrix, statistical features from the statistical distribution of each point, such as zoning, moments, crossings, Fourier transformations, and projection histograms are obtained [24].

Statistical features are also noteworthy as global features because they are usually combined and extracted in sub-images such as meshes. Statistical features are initially provided for recognizing machine-printed characters. The structural or topological features, on the other hand, concern the geometry of the group of characters to be considered. Convexities and concavities in the characters are all of these features, number of character gaps, number of endpoints etc.

2.6.5 Classification Phase

OCR systems typically use methodologies for pattern recognition which assign each example a predefined class. Classification is the process of distributing inputs in relation to detected information to their comparing class, in order to create groups of homogeneous attributes, thus separating different inputs into different classes. Classification is expressed on the basis of placing features in the feature space, such as structural features, global features, and so forth.

Classification can be said to divide the space of functions into several parts, taking into account the decision law. Choosing a classifier depends on several agents, including number of free parameters, collection of available training, and so on. The scientists are discussing various OCR procedures. The OCR classification techniques [25] can be classified as Statistical Techniques, Neural Networks, Template Matching, Support Vector Machine (SVM) algorithms, and Combination Classifier.

2.6.6 Postprocessing Phase

It has been shown that up to 60 per cent of people can read handwriting by meaning. While preprocessing attempts to clean the record in a specific sense, critical data may be evacuated because the background data is not available at this stage.

It will add a substantial measure on the off chance of the textual data not accessible to a particular degree to the accuracy of the OCR phases. On the other hand, OCR's entire problem is for determining the context of the saved image. Therefore, the incorporation of meaning and type data in all phases of OCR frameworks is essential for substantial improvements in recognition rates.

This is done in the Postprocessing stage with a reference to the early stages of OCR. Using a dictionary to modify the minor errors in OCR frameworks is the least complicated way to compile context data. The basic thinking is to spell check the OCR yield and to provide the recognizer yields that occur in the dictionary with some distinct choices.

2.7 OCR Applications

Optical character recognition has been implemented in a number of applications. We discussed some of those areas of application in this section.

2.7.1 Handwriting Recognition

Recognition of handwriting is the ability of a Computer to acquire and interpret intelligible handwritten data from sources like paper papers, photographs, touch-screens and various gadgets. The image of the written text could be detected "off track" from a bit of paper by optical scanning (optical character recognition), or clever word recognition. At the other hand, the advances in pen tips can be seen "at paper," for instance from a pen-based PC screen sheet.

2.7.2 Receipt Imaging

Receipt imagery is widely used by companies as part of various applications to monitor financial transactions and to avoid accumulations of payments from heaping up. OCR simplifies information collecting and sharing between the diverse systems of government agencies and autonomous organizations.

2.7.4 Legal Industry

The legal sector is likewise one of the beneficiaries of OCR growth. OCR is used for the digitization of documents and direct access to the PC database. Legitimate experts can further search required documents from huge repositories by writing a few keywords in essence.

2.7.4 Banking

The further imperative use of OCR is in banking, where it is used to carry out tests without involvement of man. A check can be embedded with a device where the frame filters the sums to be issued and the cash measure is exchanged accordingly. This innovation has been idealized for typed cheque, and is genuinely useful for handwritten checks that decrease bank hold-up time.

2.7.5 Healthcare

Medicinal facilities have saw an increase in the use of OCR innovation to generate printed material. Experts in the medicinal services need to maintain ongoing large quantities of paperwork for each patient, including protection frames and, moreover, general health forms. It is necessary to enter relevant information into an electronic database, in order to remain aware of each of these records. With OCR retrieval tools, we can extract data from systems and put it into databases, so that any patient's information can be easily recorded and retrieved as needed in the future.

2.7.6 Captcha

A CAPTCHA is a system that creates and tests assessments that can be passed by humans but not through modern software technology. Malicious programmer can make software for the misuse of personal information on websites. Dictionary attack is an assault on recorded secret word structures in which a programmer composes a scheme to over and over attempt distinctive passwords like most regular dictionaries of passwords. In CAPTCHA, a picture is produced that contains an arrangement of letters and numbers with a variety of sizes and types of text, distracting backgrounds, random lines, highlights and noise, so that text can not be read through

OCR. Current OCR systems should be used to evacuate the noise and portion of the image to make tractable to such malicious users.

2.8 Optical Character Recognition reading by Tesseract open source tool

Various range of devices include the optical character recognition and one of them is tesseract. This approach was used to translate the written text into editable text. The method is also capable of extracting text from handwritten texts, including some cursive writing. This tesseract was developed at HP between 1984 and 1994 [26] and is an open source. Then, in 1995, it was updated and improved with greater precision. HP published this in late 2005 as an open source for others to also contribute [27].

These are the architecture steps of tesseract OCR,

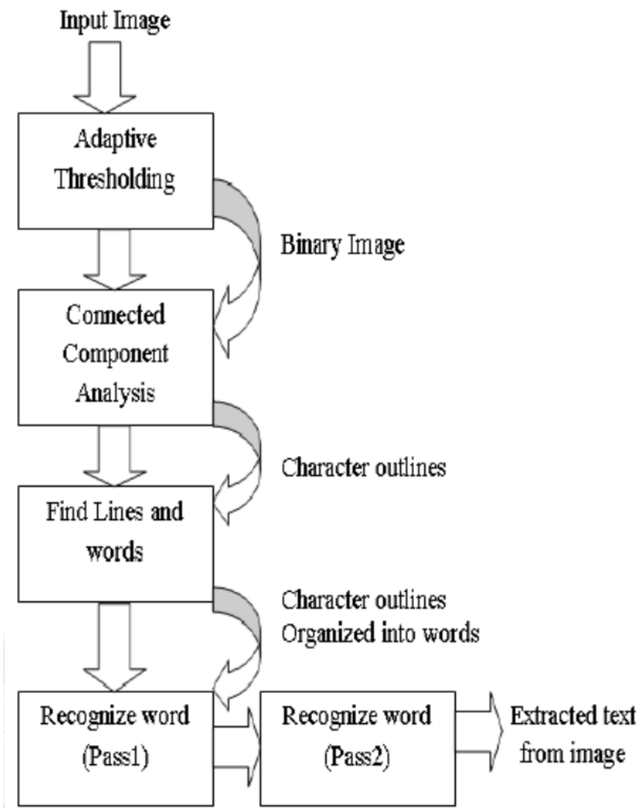

Figure 2.1: Architecture steps of Tesseract OCR

There is a research conducted comparing the tesseract and Transym tools to extract the texts from image. The authors have used set of images and scanned them through those tools to get the texts and compared them manually to check which read the texts accurately.[28]

The conclusion from the authors have resulted that the tesseract has better accuracy. Tesseract is often better or quicker than Transym by extracting text from the images is more accurate. Authors have used Vehicle license plate to conduct this experiment.

Image No	Image Type	Number of character in image	No of characters extracted	Accuracy of OCR of color images (in Percentages)	Time taken for OCR (in Seconds)	Image Type	No of characters extracted after converting color to gray scale image	Accuracy of OCR of gray scale images (in Percentages)	Time taken for OCR (in Seconds)	Change in Accuracy (In percentages)
1	color	**12**	5	42	0.4	gray scale	5	42	0.397	
2	color	**12**	12	100	0.202	color	12	100	0.202	
3	color	**12**	8	67	0.301	gray scale	8	67	0.601	
4	color	**9**	9	100	0.5	color	9	100	0.5	
5	color	**8**	8	100	0.505	color	8	100	0.505	
6	color	**9**	7	78	0.909	gray scale	7	78	0.909	
7	color	**8**	8	100	0.805	color	8	100	0.805	
8	color	**9**	7	78	1.01	gray scale	7	78	1.01	
9	color	**10**	7	70	0.85	gray scale	7	70	0.798	
10	color	**9**	4	44	0.907	gray scale	5	56	0.402	20
11	color	**10**	1	10	1.007	gray scale	4	40	0.548	75
12	color	**10**	4	40	0.699	gray scale	7	70	0.402	42.86
13	color	**10**	3	30	1.51	gray scale	4	40	0.701	25
14	color	**9**	0	0	1.008	gray scale	4	44	0.705	100
15	color	**9**	0	0	1.815	gray scale	2	22	0.7	100
16	color	**11**	6	55	1.619	gray scale	8	73	1.717	25
17	color	**9**	5	56	0.99	gray scale	6	67	0.806	16.67
18	color	**11**	5	45	0.907	gray scale	6	55	0.596	16.67
19	color	**9**	9	100	3.048	color	9	100	3.048	
20	color	**9**	9	100	1.007	color	9	100	1.007	
			Average Accuracy	**61**			**Average Accuracy**	**70**		

Figure 2.2: Tesseract Optical Character Recognition Result Analysis

2.9 OCR Engine to extract food items from receipts

A research has been conducted already on extracting the food items and amount from the receipts using Tesseract OCR open source by Google with novel image processing technique to get better results of parsing the grocery receipt images. In this research, the authors have taken below actions in converting the captured image to a readable one.[29]

1. Image stitching.
2. Image background removal
3. Image Binarization
4. Text DE skewing
5. Image Resizing
6. Text extraction
7. Spell correction

Figure 2.3: Walmart receipts before and after image background removal

After above processes, captured image will be now ready for text extraction by tesseract OCR. After extracting the texts, the already written regex model will discard any unwanted text by simply using ‘constant words dictionary’. Then using the heuristic item name and prices are extracted from each line in OCR result using regular expressions.

This approach seems too difficult to maintain since we have to supply all grocery items into the system to match the wordings. Authors had limitations where the heuristic used there can be misguide in many situations. However, this approach can be combined with my research to give better results in duplicate identification

```
FAS wc ==== 7.97
 HAND TOWEL ==== 2.97
 GATORADE ==== 2.00
 GATORADE ==== 2.00
 OXICLEAN VSR ==== 7.52
 MTC CAR FLAG ==== 9.97
 SHIRT ==== 16.88====
 PUSH PINS ==== 1.24
 ULTRATECH ==== 5.97
 REESE MINI ==== 2.88
 oz CUP ==== 0.87
 COPY PAPER ==== 4.22
 SEAGRAMS LQ ==== 23.47
 ANEX TEND ==== 95.75
```

Figure 2.4: Final text retrieved from Walmart receipt

2.10 Text extraction on bills and Invoices

A further work on extracting texts from bills and invoices has been carried out. Authors also concentrated on the performance efficiency. Authors used OpenCV, a series of Image Processing Algorithms developed by Intel primarily for real-time image and video analysis. Under the open source BSD license, they used the open source library. [30]

Tesseract OCR is used for optical character recognition. Authors have followed below steps to prepare their data before sending to Tesseract.

1. Edge Detection – Used Canny Edge Detection Algorithm.
2. Contour Detection – The edges in the image were used to find the contour (outline representing the piece of paper being scanned.
3. A four point perspective transform was applied to obtain the document's top-down view.

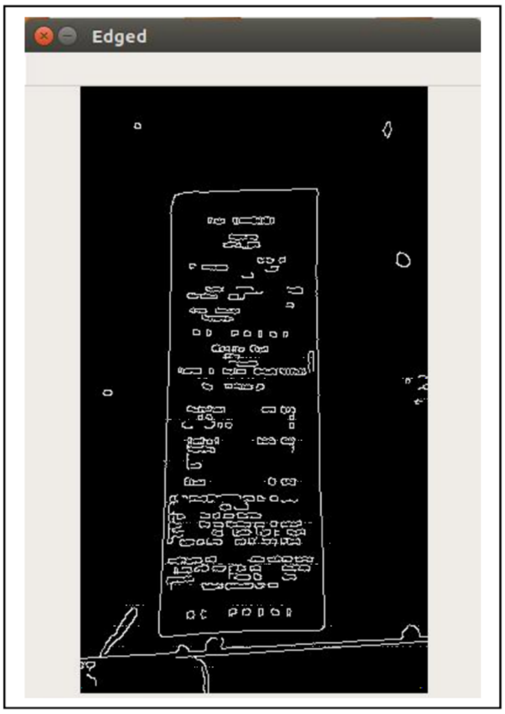

Figure 2.5: Example output image after the canny edge detection

Authors used OpenCV technology to get the document from the user's bill image which could have background noise to provide a transformed bill or invoice image just to increase the accuracy of text extraction. The Canny Edge Detector is an effective edge detection algorithm which was developed in 1986 by John F. Canny [31]. It can find an image's edges. It uses a multi-state algorithm to detect a great range of edges of images.

If the edges are found, the image is transformed to a grayscale. They later apply the four-point perspective transform to obtain the document's top-down view. They have used the Tesseract OCR engine for the text recognition as described above. For text-based images, they use segmentation, aimed at extracting detailed details from the entire image.

1. Line Segmentation – It is the first move for segmentation of the images. The engine does a horizontal scan of the file.

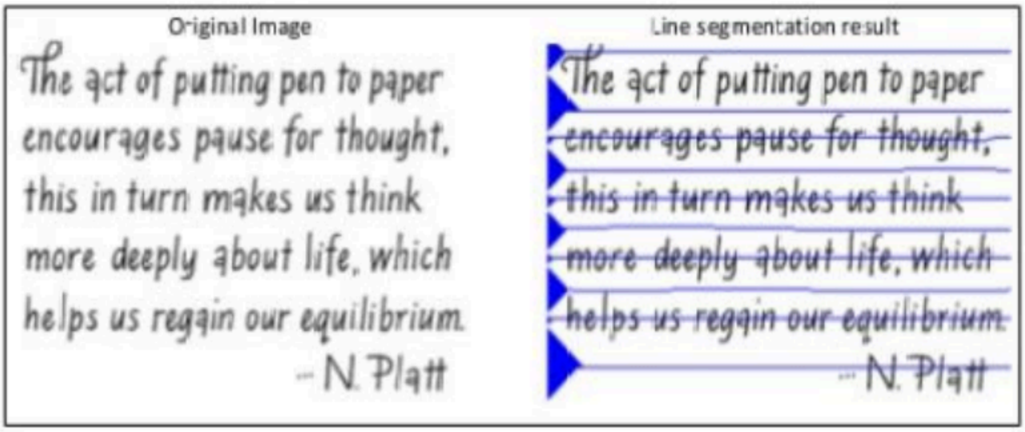

Figure 2.6: Line segmentation process

2. Word Segmentation – After the horizontal scan the second stage is the vertical scan, which is called as word segmentations.

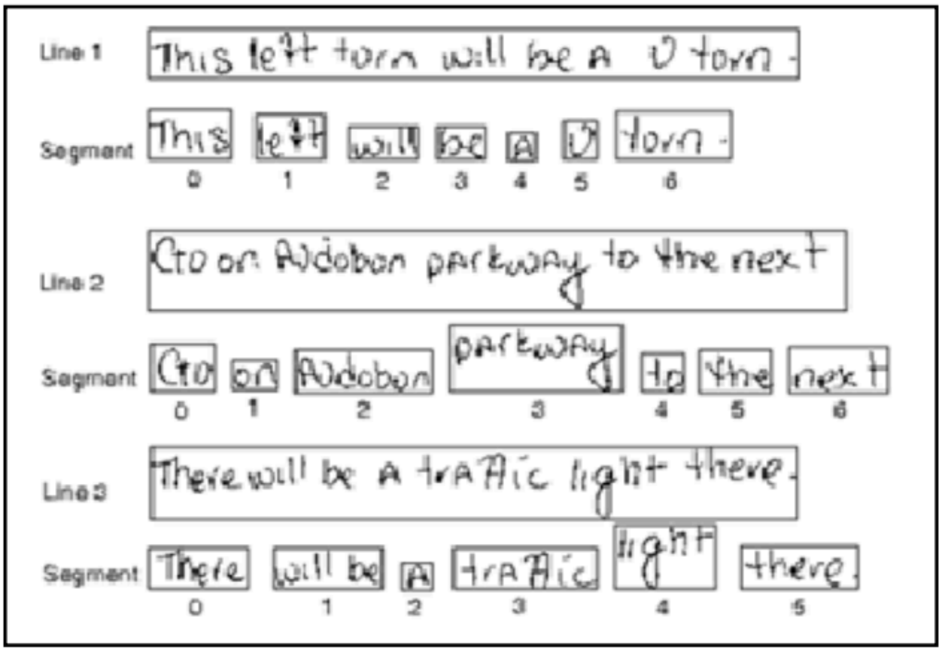

Figure 2.7: Word segmentation process

3. Character Segmentation – Segmentation of characters is the final stage for segmentation of text-based images. It includes vertical picture scanning.

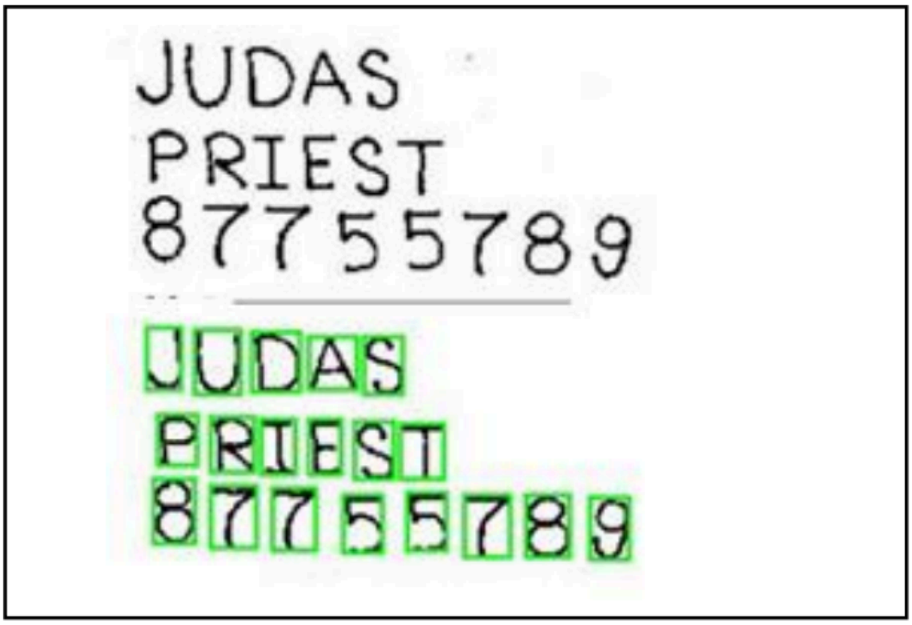

Figure 2.8: Character segmentation process

This above applied methodology has drawbacks which are explicitly addressed by the writers. They should extract the text from the bills and invoices, as tesseract uses text segmentation technique that recognizes first character in the line and then tries to read the entire line according to the location of the first character, and in the case of handwritten text the flow is not the best, because the characters are not well recognized.

2.11 Summary Table

Author(s)	OCR Application	Accuracy %
Shah, Parul, et al [32]	Chassis-number recognition	95.49
Zhai, Xiaojun, et al [33]	Automatic Number Plate Recognition ANPR	97.3
Shamsher, Inam, et al [34]	OCR for printed Urdu script	98.3
Yetirajam, Manas, et al [35]	Classification and Recognition of broken characters	68.33
Ruvan Weerasinghe [36]	Creation of a commercial grade Tamil OCR to recognize independent text for font and scale	81

Table 2.2: OCR Applications and the accuracy

2.12 Deep learning techniques compare with traditional OCR methods

OCR is the method to automatic text extraction from the image with machine vision, letter recognition, and other techniques. The image can be a scan of a printed page, a photo or some other textual data not already readable on your computer.

The OCR method was very easy back in the day. We took a bunch of letters and figured out how they looked like. Maybe the edges of each letter were identified, the angles match, and then we coded it to a system for OCR. If the input was dirty, you might have opted to clear stuff in a threshold function.

This technique was used by early open-source OCR systems such as Tesseract, which originated from Xerox efforts. When we scanned, typed pages, preferably converted to grayscale with the appropriate rates and lighting, they performed well with what we were processing.

Not all actual sources of knowledge are shockingly clean. And so, OCR did not work very well until recently, except in some restricted cases. It was connected into the copier and scanner program, was used in platforms such as Evernote and otherwise written for the highest value where it operated. But precision fell and quickly for anything beyond clean text.

That is achieved by studying letter types and variants on these types in modern, deep learning neural networks. This can be achieved by evaluating the likelihood of the messages, in which sense the scene will occur. It could be magic gnome.

The reality is, we don't even know in many situations. We will see the edge detection algorithm, maybe export the picture of an intermediate or two to test, and have a general understanding of how it works together. The classical OCR vision strategies have become introspective so we can understand it. At the other hand, neural networks are so complex that they are simply a black box. We put images of the text in one end and get the readable text out of the computer in a processing.

3. METHODOLOGY

3.1 Identifying the category of a receipt

3.1.1 Convolutional Neural Network

One of the main classes of neural networks is the Convolutional Neural Network (ConvNets or CNNs) to do image recognition, the classification of images. Face detections, facial recognition, etc. are some of the areas widely associated for CNNs. The processing of CNN images takes, filters and classifies an input image into those categories. Computers view an input image as an array of pixels and that depends on the image resolution. Based on image resolution it will see h x w x d (h = height, w = width, d = dimension). For example, a matrix array of 6 x 6 x 3 RGB (3 refers to RGB values) and a matrix array of 4 x 4 x 1 in grayscale.

Technically, each input image will pass through a series of filter (kernel) convolution layers, pooling, fully connected layers (FC), and use SoftMax to classify an object with probabilistic values from 0 to 1. The following figure is a full CNN flow to process an input image, and classifies objects based on values [37].

Convolution is the first layer in which features from an input image are removed. Convolution preserves the relationship between pixels by the use of tiny squares of input data to learn image properties. It is a mathematical operation involving two inputs such as an image matrix, and a filter or kernel. [38]

3.1.2 Microsoft Azure Custom Vision

Microsoft provides two establish image recognition features. First is Computer Vision, which uses a pre-built model which is created and maintained by Microsoft. The second one is the Custom Vision service, which allows you to build and train your own model, dedicated to a given image domain.

Although there's more work involved in sourcing the images, tagging them individually, training and refining the model, the advantage of a dedicated one is that we will get improved accuracy when using only images from own chose subject. This is because of the model being able to make finer distinctions between images and to avoid distractions from superficially similar but in reality, quite different image subjects.

Microsoft Azure Custom Vision API which uses the CNN (Convolutional Neural Network) can be used to identify the category of a receipt. First, we need to train the CNN model for image classification with the existing receipts. Minimum 50 images are required for each category to give an optimal output.

A machine learning algorithm is used by the Custom Vision service to identify the images. We have to apply group of photos featuring and missing the classification(s) in question. At the time of submission, we determine the appropriate tags for the pictures. The algorithm then trains to this data and measures its own accuracy by checking on the same data itself. After we have learned the model, we can check, retrain and ultimately use it to identify new images according to the app's needs. The platform itself can also be exported for offline use.

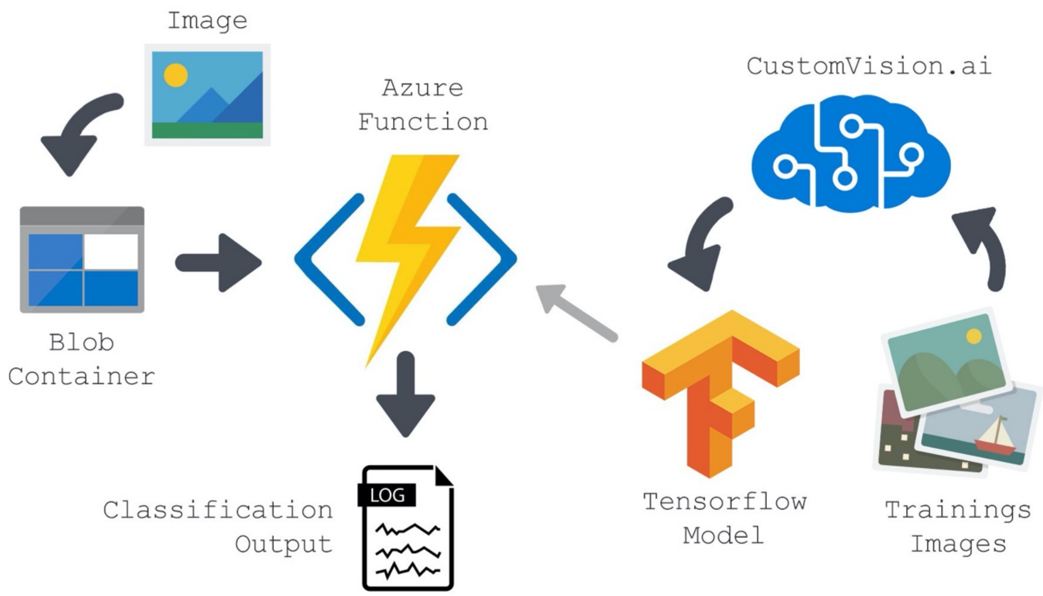

Figure 3.1: Azure Custom Vision Architecture [39]

The training process can take many iterations to train the model by the given set of images. Two options of training available, 'quick training' and 'advance training'. Quick training does not cost, however the computational capacity is low, so the performance will have impact and time taking to train the model will be high. Advance training is charge based on the hour we train the model, but performance will be high since high CPU are used. These two options will not affect the iterations, we can use as many iterations to train the model.

Probability threshold which will be used to adjust the precision and the prediction. We have to upload the images with the tags and train them first to control the probability threshold. Since Custom vision train themselves with the input data, we have to test the precision once the model is been trained with some different set of data.

3.2 Extract the amount from a receipt

3.2.1 Recurrent Neural Networks

Human beings do not start their thinking every second from scratch because thoughts are permanent. This is not possible for conventional neural networks and it seems like a big shortcoming. Imagine, for example, that we want to define what sort of event happens at any point in a film. It's unclear how a conventional neural network could use its reasoning of previous film events to inform later events. [40]

Recurrent neural networks tackle this problem. These are networks with loops inside them that allow for persistent information. Chain-like existence shows that recurring neural networks are closely connected to lists and sequences.

3.2.2 Long Short-Term Networks

Usually named Long Short-Term Memory Networks as LSTMs that are a special type of RNN capable of long-term dependency learning. Hoch Reiter & Schmid Huber (1997) introduced them, and a lot of people developed and popularized them in the research that followed. They work extremely well on a wide variety of problems, and are now commonly used. [41]

The LSTMs had been deliberately designed to avoid the long-term dependency problem. It's simply their natural ability to recall facts for long periods of time, not something they seek to learn. All recurrent neural networks take the form of a series of recurring modules in the neural network. For regular RNNs this repeating module will have a very simple structure, like a single tanh layer. [42]

Recurrent networks are neural networks that take not only the present example of input they see as their input but also what they experienced in the earlier steps of time. The decision hitting recurrent nets at time step t-1 affects the decision that they will make at time step t later. Recurring networks thus have two input sources which are both the current and the recent past. Recurrent networks are often said to have memory because it can be shown that Recurrent networks maintain knowledge about the temporal structure of the data as opposed to Feedforward networks. This sequential knowledge is stored in the secret state of the recurrent network, spanning several steps of time as it cascades forward to affect the processing of each subsequent example. Recurrent networks have the function of precisely classifying or predicting sequential data. To accomplish this, backpropagation and gradient descent are employed. May imagine recurrent networks as shown in Figure 3.2.

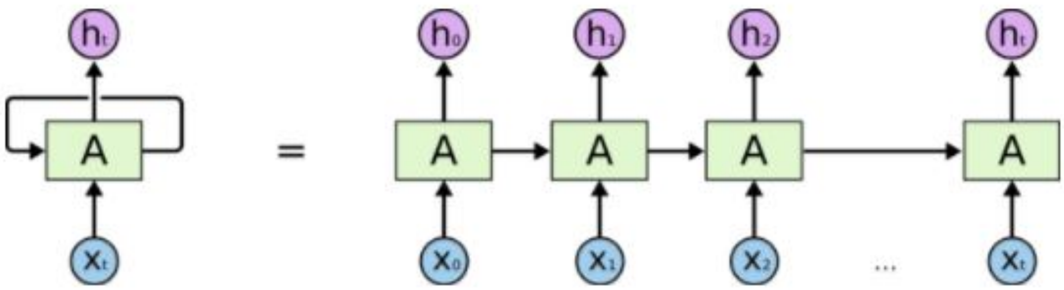

Figure 3.2: An unrolled recurrent neural network

In Feedforward networks, Backpropagation travels backwards from the final error through each layer's outputs, weights and inputs, assigning responsibility for a portion of the error to those weights by measuring their partial derivatives. The gradient descent algorithm will then use these derivatives to change the weights to decrease the error defined by a loss function. Recurrent networks rely on a backpropagation extension called Backpropagation Through Time (BPTT) [43].

Given below in Figure 3.3 illustrates the repeating module in a standard Recurrent Neural Network (RNN) which usually consists of a single tanh layer.

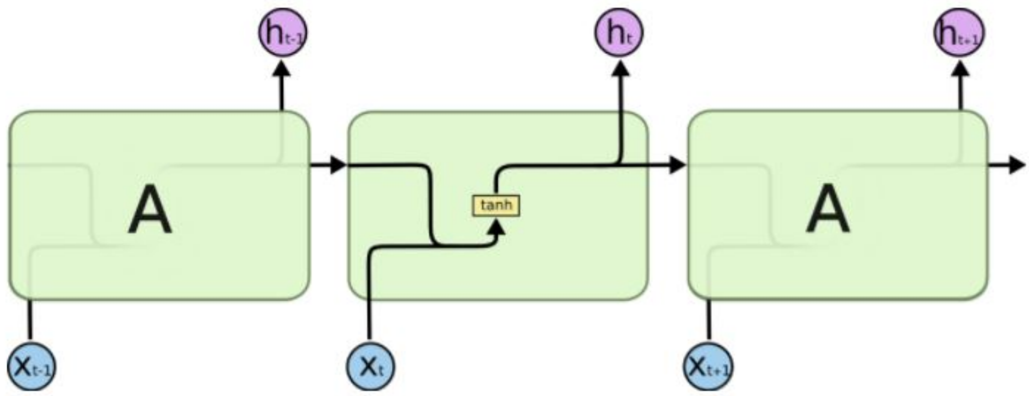

Figure 3.3: The repeating module in a standard RNN

The problem of the vanishing gradient is a major obstacle which has arisen in terms of recurrent net efficiency. Large Short-Term Memory Modules, or LSTMs, were suggested as a solution to the vanishing gradient problem. LSTMs help to preserve the error, which can be propagated backwards over time and layers. They allow recurring networks to continue learning over many time steps by maintaining a more constant error, thus opening a channel to connect causes and impacts remotely.

LSTMs store information that goes beyond the usual recurrent network flow in a gated container. Information can be stored in, written to, or read in a container. By opening and closing gates, the cell determines what to hold, and when to let reads, writes, and erasures. These gates operate on the signals they receive, and similar to the neural network nodes that they block or pass on information based on their strength and import which they filter with their own weight sets. These weights, similar to the weights controling inputs and hidden states, are adjusted via the recurrent network learning process. [44]

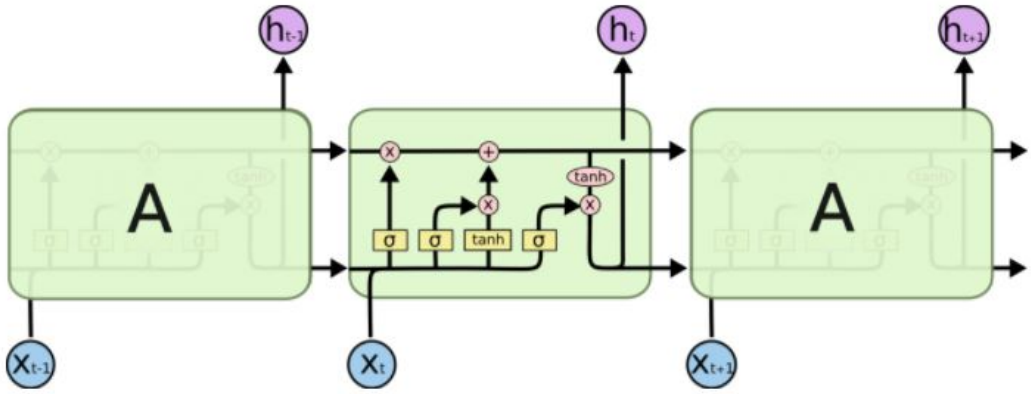

Figure 3.4: For interacting layers in LSTM repeating module

Besides traditional LSTMs, there is a variant known as a Gated Recurrent Unit (GRU), which is essentially an LSTM without an output gate, and therefore writes the contents from its memory cell to the larger net at every stage.

3.2.3 Microsoft Azure Computer Vision – Cognitive Service

Microsoft Azure provides a range of computer vision-based resources. They're packed into as separate APIs. Example: Computer Vision API for general-purpose CV operations, Face API for Face Detection and Recognition, Filtering Content Moderator and a few more which are still in preview status.

The Computer Vision API enables the classification of image content by providing a comprehensive tag list and trying to create a meaningful representation of the scene in the language. The API will recognise the celebrities and landmarks, in addition. Another important feature which was needed for this research was the optical character recognition of printed text.

Use Computer Vision's supported API, we can use this to extract printed and handwritten text from images into a machine-readable character stream. The read API uses the latest models and works on a number of surfaces and backgrounds including receipts, posters, business cards, letters and whiteboards with text. And offline supports with computer vision containers which can be used locally to enable extraction of characters.

Microsoft Azure provides an API which extracts the text from an image and returns the information along with the text coordinates.

Example:


```
{
  "status": "Succeeded",
  "succeeded": true,
  "failed": false,
  "finished": true,
  "recognitionResult": {
    "lines": [
      {
        "boundingBox": [
          117,
          37,
          325,
          40,
          324,
          60,
          116,
          57
        ],
```


Figure 3.5: Sample receipt and the scanned value from Azure computer vision

3.2.4 Identifying the amount using regular expression

The Computer Vision API enables the classification of image content by providing a comprehensive tag list and trying to create a meaningful representation of the scene in the language. The API will recognize the celebrities and landmarks, in addition. Another important feature which was needed for this research was the optical character recognition of printed text.

If the text is extracted from the image, the regex (regular expression) within the text can be used to identify the total sum of a receipt. We define our own search pattern to exactly point to the total amount. In order to get the optimal output, we train our search pattern with different kind/format of receipts.

Example: ‘Total Amount’, ‘Net Amount’, ‘Grand Total’, ‘Total’ those are common key words used to indicate the total amount. Based on that, once the text extracted from the image, Regex model can identify this key word and get the text bound and use that to go to the exact row to get the amount value. By doing this precision control over the key words, we can ignore amounts of the items and the balance amount. However, there are some edge cases where the amount is printed without any keywords indicating, it is the total amount. In those cases, the Regex approach will fail. We could write some more sub regex models to identify those edge cases as well.

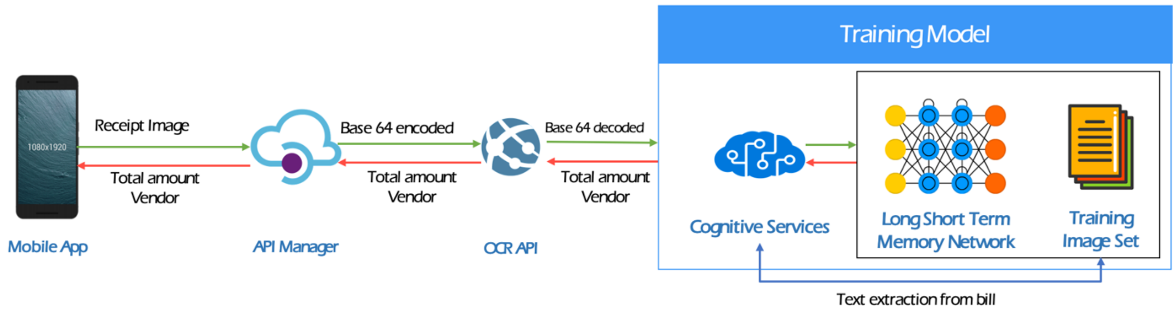

Figure 3.6: LSTM Architecture

3.3 Identifying the duplicate receipts

Duplicate Identifier can be done in two different approach. The approaches I am going to discuss below can be combined together to provide a better solution with better prediction.

3.3.1 Text Similarity

Text similarity can be calculated as to whether two pieces of text are both identical to the surface, which is called lexical similarity, who's also related to semantic similarity. For example, the phrases 'the cat ate the mouse' and 'the dog ate the cat food' are similar. On the surface, if we find just similarity at world level, these two sentences tend to be very similar as 3 of the 4 distinct terms are an exact overlap. Usually, in context, it does not take into account the real meaning behind words or whole sentence. Instead of doing a word for word comparison, in order to catch some of the meaning, we should pay attention to context. In order to understand semantic similarity, we will concentrate on the levels of phrase / paragraph where a piece of text is split into a different category of related terms before computing similarity. We know that although the words overlap greatly, these two terms do have different meanings.

3.3.2 Jaccard Similarity

Jaccard similitude or intersection is defined as the size of the intersection divided by the union size of two sets [45]. Jaccard coefficient will do very well when comparing terms with every letter of the word while calculating the relation. Every letter directly can change positions and count as the same words. This approach can't identify the over-type terms in the data sets, however. Jaccard similarity coefficient is sufficient for the term similarity metric to be used sufficiently. For efficiency calculation the performance of the system should manage high volatility properly when spelling failure and error occurred [46].

3.3.3 Comparing the texts of the receipts

In this approach, once a receipt is scanned using the cognitive services, extract the entire text after removing the text coordinates and store it in the database with a unique identifier. When the image is being scanned again, the model will cross check with the database to see whether it contains any existing data with similar text and will do the comparison of both texts if which percentage is above ~90% similar, then system can suggest this might be a duplicate. We can use ‘jaccard similarity’ model to calculate the similarity between two text sets.

In the above approach, if the receipts are from same restaurant, items are same and only the date and time will be different, in such case this approach might fail.

3.3.4 Comparison between date and time, vendor, amount and category

In this approach, after the identification of category, vendor, amount and the date using the LSTM model, store that information in the database with a unique identifier. When the image is being scanned again, system can cross check for that information of the new receipt in the database and identify the existing record if there are any.

In this case, the prediction will be more accurate. Both approaches will take time to do the scanning since it has to scan the entire database. If we use one of the approaches, we might end up accusing the wrong receipt for a duplicate since the items on the receipt could be same and the purchase might happen at the same shop on the same day but the time will be difference. We should use the jaccard similarity and get the predicted duplicate receipt first and then compare the date and time to reduce the performance and output time.

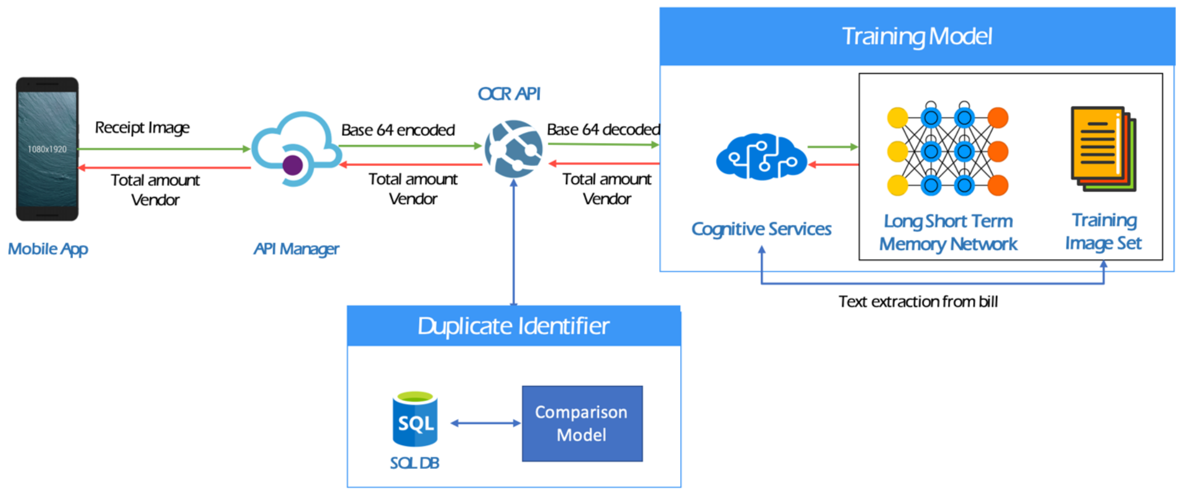

Figure 3.6: Duplicate Receipt Identifier Architecture

4. SOLUTION ARCHITECTURE AND IMPLEMENTATION

As discussed in the methodology, the following steps were followed during implementation of the solution architecture. The details of which will be discussed in detail in this section.

- Load, Extract, Transform (ETL)
- Database preparation
- Mobile App development

4.1 Load, Extract, Transform (ETL)

Under the ETL process, there were two sources integrated. The first source was the receipts collected from the restaurants and taxi services, etc. The second source was the automated script written in ABAP language using SAP system to generate random receipts in different formats. Totally collected of ~ 1,000 different receipts in different formats.

The entire ETL process was carried out using Python and JSON modules being used for accessing the Azure APIs while the csv module was used for outputting the data in flat file format.

Azure custom vision has been trained by those source images by tagging and categorize them individually for the much accurate results. After the completion of training, given new set of receipts, the prediction service has been tested and validated. Azure custom vision does not allow to annotate already predefined data set, if it allows, we could train the model with much more large set of data.

Azure computer vision has been loaded with source images and compared the output results manually to validate that the given bounding boxes and the texts are much accurate for the low resolution and damaged receipts. Given output by the Azure computer vision is 8 out of 10 receipts are correct. There are some limitations when extracting the texts from damaged and hand written receipts.

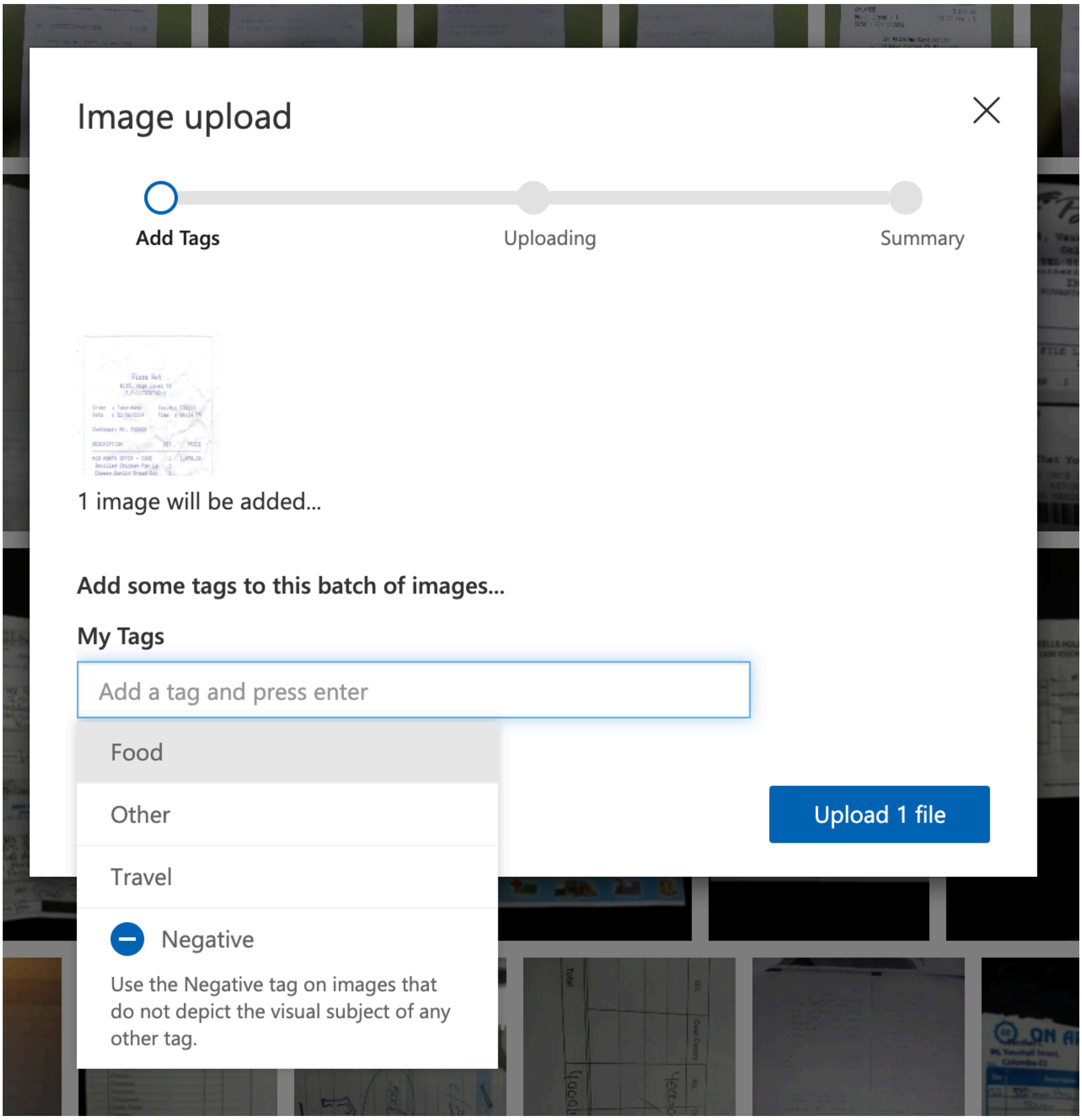

Figure 4.1: Tagging of the receipt in Azure custom vision

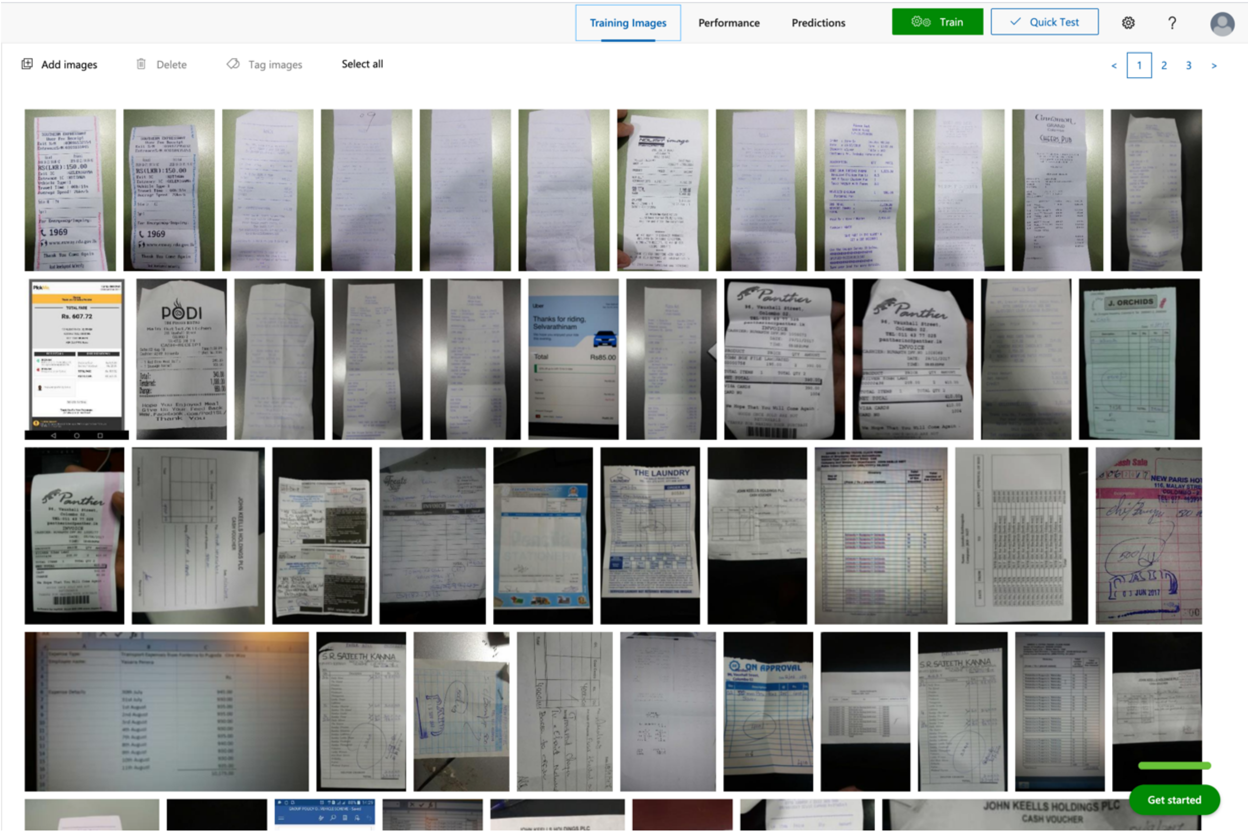

Figure 4.2: Tagged images set in Azure Custom Vision

Adjusting the probability threshold and checking the predictions to get the best precision and performance that the model created will provide. We can continue to tweak the model to look to improve these numbers by providing further images. I aimed for at least 50 images in each category with a balanced number across each, and look to provide the range of viewpoints, backgrounds and other settings in which test images may be provided when the model is in use. Each time I did this, I retrained and got a new iteration of the model with an updated set of statistics.

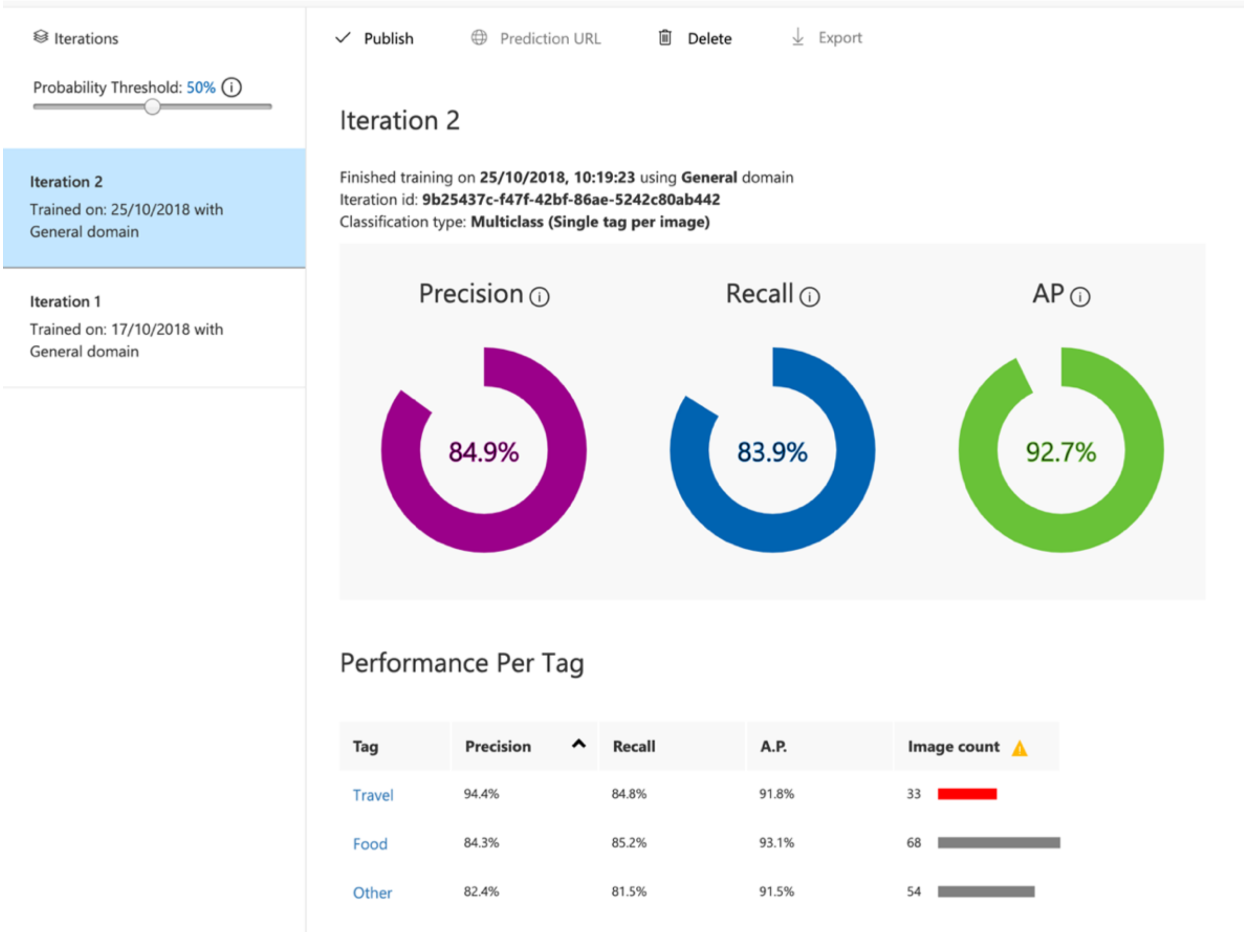

Figure 4.3: Checking the performance after training

For each iteration, the model trains for the given set of images to get the optimal performance which needed for identifying the correct category. Probability threshold is controlled to provide more precision with high performance.

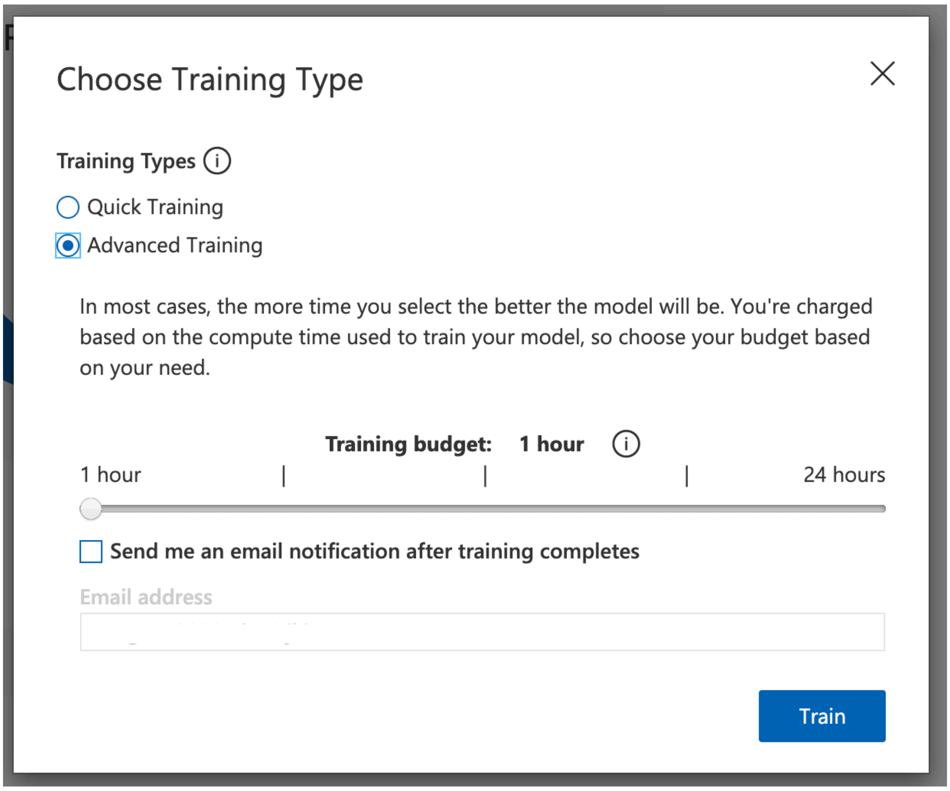

Figure 4.4: Selection interface of the training type

The above figure 4.4 shows the two types of training provided by Microsoft Azure Custom Vision. Quick training gives a free training on the images set however; it takes quite a bit of time for the completion of the training. Advanced training option used high CPU and Memory to train the images set in reasonable amount of time using a better model. We will be charge for the training time if we choose Advanced training.

I used both approaches to see which is better and I could not differentiate any. Both training types provided me the same precision and testing output.

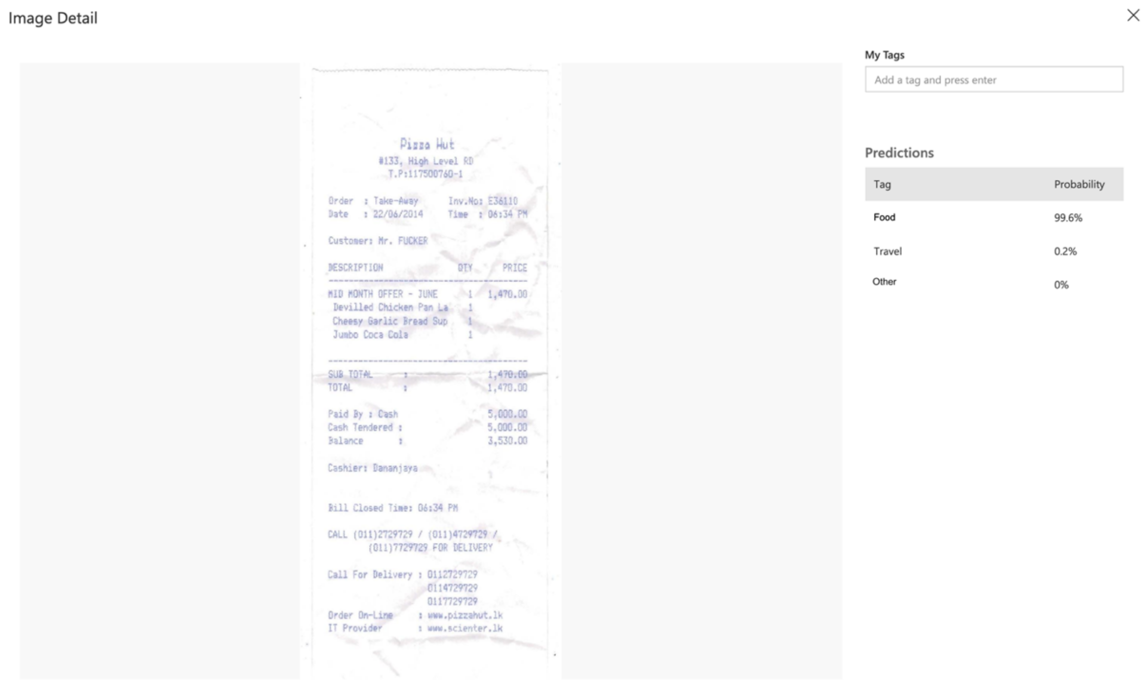

Figure 4.5: Testing the prediction after the training of the model

After the completion of the training with many iterations. I tested the prediction with new set of images to confirm the model which predicts the category of the receipts. In some circumstances, the new receipts show incorrect output, which could be due to the lack images provided to the model to train itself. In such cases, I uploaded the incorrect result shown images to the model and tag them with correct category and train them again with new iteration. After few iterations, I tested the final model with 25 images and got 19 correct prediction.

4.2 Database preparation

Extracted information of the receipts were supposed to store in a persistent database to carry out the duplicate identifier functionality. For this process used the Mongo DB and stored the extracted information as a collection with attaching the unique identifier to them. Whenever the search query comes to get the text information NoSql database approach is faster, hence used the MongoDB.

4.3 Mobile App development

As a front end, Xamarin is used to develop the mobile app. Xamarin provides single language C# and supports multiple platforms like Android, iOS, Mac, etc. It provides native performance which works directly on the hardware, rather than on abstractions like web views. Xamarin provides several ways to share code across platforms because of the high reusable code.

This mobile app interface handles two parts,

1. Capturing of the receipts and send the captured image in Base64 format to the API.
2. When API found that the given receipt is a duplicate, the API will return the duplicate receipts with original receipt in Base64 formats. The mobile app will convert them into images and show to the user for comparison.

4.3.1 Developed mobile app interfaces

Keells

No. 199, Union Place, Colombo 2
0770 108459 / 0112 303 500
05/10/2018 08:17 C:GANE R:C001262

Ln	Item	Price	Qty	Amount
1	CHICKEN SANDWICH BAGUETTE			
	956005	130.00	1	130.00

Gross Amount	130.00
Net Amount	130.00
Cash	200.00
Change	70.00

Thank you Mr. Selvarathinam Vinoch
for your valued patronage!
Nexus Mobile points earned for
this bill: 0.44
Tot redeemable points @04-Oct:195.90
Card No:1121050000643741

Time|Start: 8:17:18 AM End: 8:17:32 AM

IMPORTANT NOTICE: In case of price

Figure 4.6: Sample receipt obtained from Keells

The above figure 4.6 mentioned receipt is scanned using the developed mobile application and the output result are shown in figure 4.7 after extracting the text and identifying the amount, vendor including the category of the receipt.

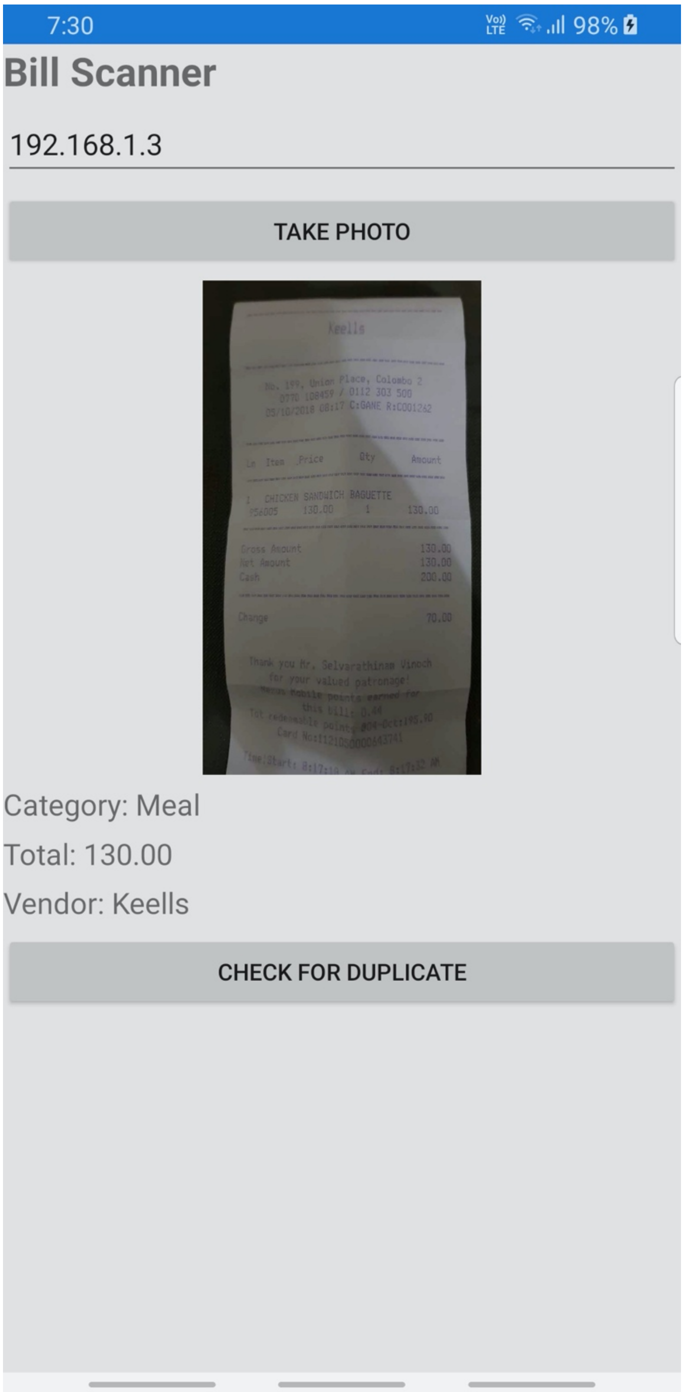

Figure 4.7: Output result after scanning the image

Added a new page to show the duplicate receipts separately because there could be more than one duplicate receipts for the given receipt. If there are no duplicate found for the given image, the app will show "No Duplicates Found" label when user clicks on "Check for Duplicate" button at the bottom of the screen.

The figure 4.8 shows the duplicate receipts of the scanned receipt in figure 4.7. To showcase the duplication, the receipts were scanned three times and the third time, I have checked for duplicates and two images which were scanned previously picked up by the system with prediction 81 percentage.

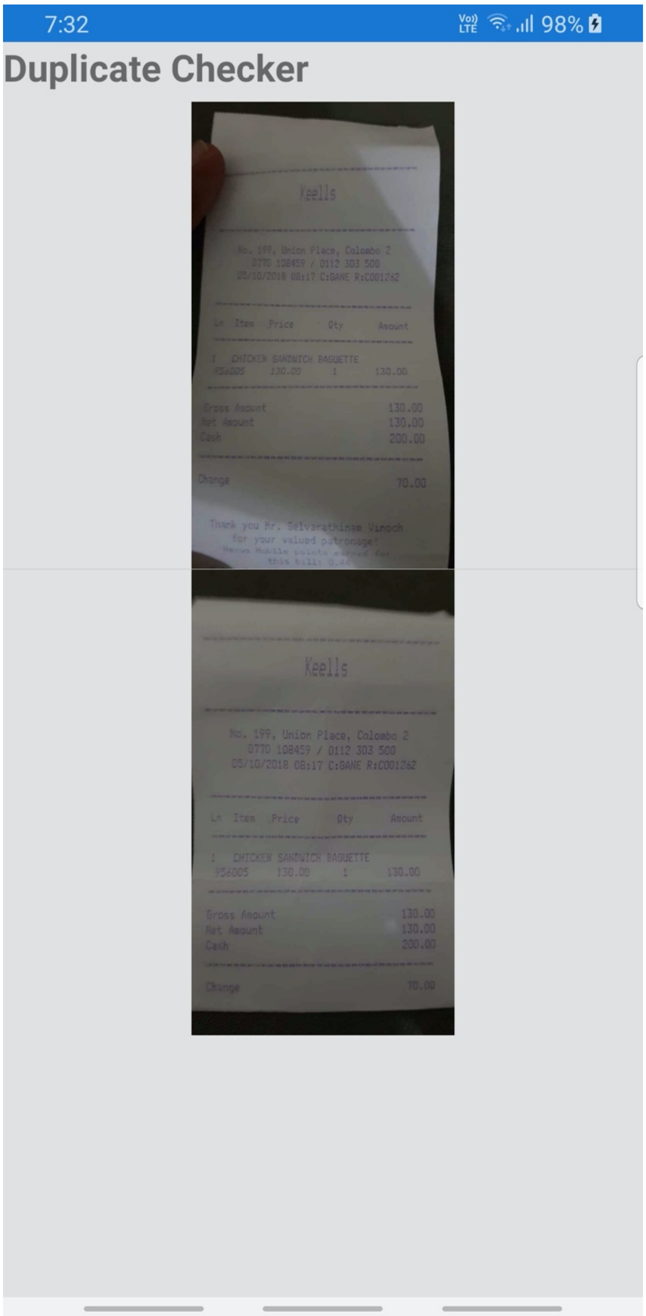

Figure 4.8: Identified duplicate receipts are shown in list

5. DATA & ANALYSIS

5.1 Custom Vision Training

Several iterations of the testing carried out in training the model to capture the images and categorize them precisely. Used the multi class classification type to train the model.

The table below will show,

1. The number of images passed to the model by tagging them with the category of the receipt
2. The probability threshold which is the minimum probability score for a prediction to be valid when calculating precision and recall
3. Precision result which will tell if a tag is predicted by the model, how likely is that to be right
4. Recall result which will tell that the out of the tags which should be predicted correctly, what percentage did the model correctly find.
5. AP result will tell a measure of the model performance, summaries the precision and recall at different threshold.

Number of Images	Probability Threshold (in percentage)	Precision (in percentage)	Recall (in percentage)	AP (in percentage)
250	50	84.9	83.9	92.7
250	92	91.8	72.9	92.7
250	65	87.5	81.9	91.5
600	50	86.6	84.7	90.4
600	93	94.4	67.9	90.3
600	65	90.2	72.9	89.4
800	50	88.7	86.3	88.5
800	92	93.3	59.9	82.8
800	65	90.2	80.1	86.1

Table 5.1: Microsoft Custom Vision training output results

Based on the outcome results, adjusting the threshold to get the most precision result will impact on the performance of the model as well as the recall.

Found that the tagging images with single tag per image and doing a compact training will give more accurate result, however consuming high CPU and training charges are high. To get the best result, train the model for longer will give the more accurate and precision results.

The fast tests were performed after training the model to verify if the model was able to predict the receipt category precisely. Find out that the model could predict low resolution category and damaged receipt. However, this resulted because of the camera that has been used to capture the image is high end mobile phone. This result could vary based on the image environment and lighting reflects on the images that are captured.

5.2 Identifying the amount using Regex

Identifying the amount from the receipt carried out in Python and using the Regex algorithm. The output given by the Azure computer vision which is a JSON results with bounding texts as shown in figure 4.6.

The texts returned by the Azure computer vision will go through set of regex rules to get the amount from the receipt. Sometimes Azure computer vision might miss out some of the characters from the receipt due to the low resolution and damages on the receipts. There are predefined set of texts given to the regex model like ‘grand total’, ‘total amount’, ‘net amount’ to figure out the matching text in the given receipt texts.

These are the used keywords to identify the total amount from the bill,

```
total_names = ['net_bill_val',
               'grand_total',
               'net_total',
               'total_fare',
               'total',
               'total:',
               'gross_amount',
               'gross_amt',
               'amount',
               'net_amount',
               'sub_total',
               'subtot',
               'cash']
```

There are instances where the receipt might not have the text as indicating the total amount, rather will be just the amount printed. In those scenarios, the model will search for any keywords contains 'balance' to ignore the numbers which such as given amount to the cashier and the balance that the cashier given back.

Though the regex approach seems not quite prominent, but this is the best approach that exists till now. If a neural network model used for this to identify the amount, there are lot of edge cases that might be missed out and a lot manual training with precision highlighting on the total amounts required to get the more accurate results.

5.3 Duplicate receipts identification

Duplicate Identifier has been done in with two approaches. First approach was getting the texts returned by the Azure computer vision API stored in the database for future receipt comparison. The second approach was storing the category, date time and vendor separately to identify the duplicate receipt.

Once a receipt or receipt was scanned using the cognitive services, extract the entire text after removing the text coordinates and store it in the database with a unique identifier. When the image was being scanned again, the model will cross check with the database to see whether it contains any existing data with similar text and will do the comparison of both texts if which percentage is above ~90% similarity, then system can suggest this could be a duplicate receipt. This became possible using the 'jaccard similarity' model to calculate the similarity between two text sets.

Piece of code where the jaccard similarity been used,

```
def get_jaccard_sim(str1, str2):
    """input: 2 string values that needs to be compared
       output: returns the similarity score between the 2
strings
    """
    a = set(str1.split())
    b = set(str2.split())
    c = a.intersection(b)
    return float(len(c)) / (len(a) + len(b) - len(c))
```

In some instances, there are chances that the receipts items, the vendor and date are the same. This scenario has been captured by using the ‘time’ variable. When Azure custom vision API returns the category, vendor and the date time, storing that information separately and the next time when a receipt is being scanned, system cross checks with existing information stored in the database and give the prediction based on the output.

Those two approaches are combined to give more accurate results though it is time consuming to process them, using the python written model.

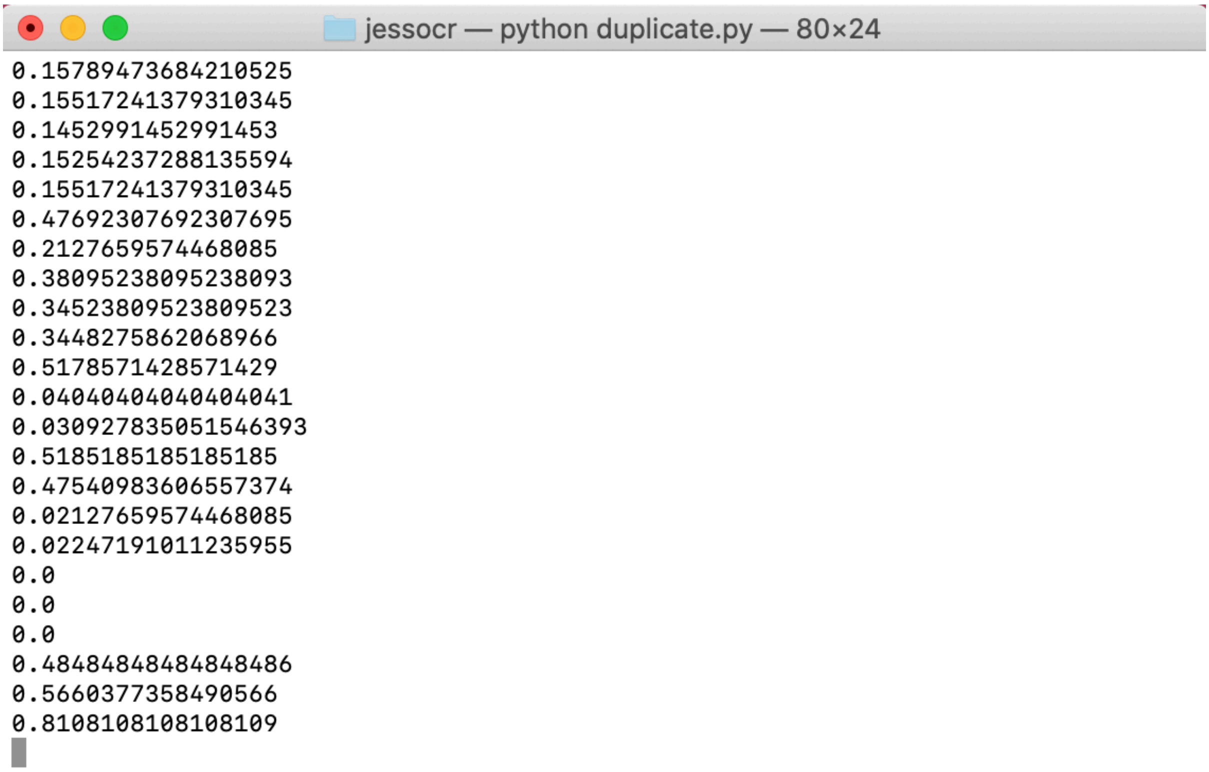

Figure 5.1: The duplicate prediction output by the developed system in percentage

6. GENERAL DISCUSSION & CONCLUSION

6.1 General Discussion on the case study

For the purpose of this study to extract text from the receipts and identify the category, vendor and amount of the receipt along with whether this receipt is duplicated. LSTM, CNN, Recurrent Neural Networks are used with Azure Computer vision and Azure Custom Vision for this study. The study shows that to get the more précised prediction on the correct identification of the category and vendor, the amount of data set and the manual training to the system is required. Tagging and Multi Class classification used to train the LSTM model in the Azure Custom Vision.

It was interesting to note that the probability threshold plays great part in the prediction and the longer time needed to train the model than expected. Though LSTM model was trained with enough amount of data sets, there are edge cases that the model missed out and the expected outcome is incorrect. Online receipts are to give better results due to lack of distortion and damages. Even Azure computer vision could not be able to extract the texts from the handwritten receipts correctly due the way those were written and the words are partially completed.

When considering the duplicate identifier model, the outcome was more than anticipated since the combination of two approaches such as jaccard similarity model and the cross checks gave far better result in predict the receipt which could be a duplicate. Now Azure cognitive services are widely used by developers and data scientists; it will result in the LSTM model to get the text extracted with much precision and that will impact the duplicate identifier to become way better.

6.2 Conclusion

The findings of this study summarized below,

- Azure computer vision which uses the LSTM model to extracts the text from the images much accurate except for below circumstances.
    - Blurry Images.
    - Handwritten or cursive text.
    - Distorted texts or texts with complex backgrounds.
- Azure custom vision using the CNN model was able to predict the receipt category very accurately with ~83 percent accuracy after the manual tagging training given to the model.
- Out of two approaches used to identify the duplicate receipts mentioned in section 5.3, turns out that to get the best results, we must combine both the approaches. This resulted in finding the duplicate receipts with average 73% prediction level.
- The duplicate receipt prediction level is average 73% due to below circumstances.
    - Duplicate identification works by comparing the texts on both receipts. If the captured environment is different and used different mobile devices, there are chances that the quality of the captured image is different. This leads to incorrect reading by the Azure Computer Vision. So, the extracted text does not much correctly.
    - After extracting the texts, we store the date and time separately to compare whether the captured images are different or not after jaccard similarity check. However, extracting the texts from receipts is quite a bit tricky for Azure Computer Vision since the date and time always printed small in all receipts.

6.3 Future work

The following are some areas which could be researched into, based on the knowledge gathered in this study:

- Extend the research to the case where the duplicate identifier model should be able to indicate the duplication based on the items that are bought to ignore the edge case where the two different receipts with same vendor, category, date and time and amount but the items are different.
- Extend the research to identify the receipt category such as food, travel, medical by the items on the receipt rather by tags.
- Bring more datasets from different vendors including the handwritten receipts to train the model to get much accurate results.

REFERENCES

[1] A Survey on Optical Character Recognition System. Journal of Information & Communication Technology. Vol 10. Issue. 2, December 2016. Noman Islam, Zeeshan Islam and Nazia Noor.

[2] A Survey of OCR Applications. International Journal of Machine Learning and Computing.Vol. 2, No. 3, June 2012.

[3] M.D. Ganis, C.L. Wilson, J.L. Blue "Neural network-based systems for handprint OCR applications" in IEEE Transactions on Image processing, 1998, Vol.7 Issue 8.

[4] R. Gosswiler, M. Kamvar, S. Baluja, "What's Up CAPTCHA? A CAPTCHA Based On Image Orientation", in WWW, 2009.

[5] R. Plamondon, S. N. Srihari, "On-line and off-line handwriting recognition: a comprehensive survey" IEEE transaction on pattern analysis and machine intelligence, 2000, 22(1), 63-84.

[6] S.L. Chang, T. Taiwan, L.S. Chen, Y.C. Chung, S.W. Chen, "Automatic license plate recognition" in IEEE transactions on Intelligent Transportation Systems, 2004, Vol. 5, Issue 1, p.p. 42 – 53.

[7] Satti, D.A., 2013, Offline Urdu Nastaliq OCR for printed Text using Analytical Approach. MS thesis report Quaid-i-Azam University: Islamabad, Pakistan.p. 141.

[8] B hansali, M., & Kumar, P, 2013, An Alternative Method for Facilitating Cheque Clearance Using Smart Phones Application. International Journal of Application or Innovation in Engineering & Management 211-217.

[9] Global Optical Character Recognition Market Snapshot https://www.transparencymarketresearch.com/optical-character-recognition-market.html

[10] Qadri, M.T., & Asif, M, 2009, Automatic Number Plate Recognition System for Vehicle Identification Using Optical Character Recognition presented at International Conference on Education Technology and Computer, Singapore, 2009. Singapore: IEEE.

[11] A Survey of OCR Applications. International Journal of Machine Learning and Computing. Vol. 2, No. 3, June 2012.

[12] A Detailed Analysis of Optical Character Recognition Technology, Karez Abdulwahhab Hamad, Mehmet Kaya. 03rd September 2016.

[13] Kaur S, Mann PS, Khurana S. Page Segmentation in OCR System-A Review.

[14] Saha S, Basu S, Nasipuri M, Basu DK. A Hough transform based technique for text segmentation. arXiv preprint arXiv:1002.4048. 2010 Feb 22.

[15] Basu S, Chaudhuri C, Kundu M, Nasipuri M, Basu DK. Text line extraction from multi-skewed handwritten documents. Pattern Recognition. 2007 Jun 30;40(6):1825-39.

[16] Khandelwal A, Choudhury P, Sarkar R, Basu S, Nasipuri M, Das N. Text line segmentation for unconstrained handwritten document images using neighborhood connected component analysis. InInternational Conference on Pattern Recognition and Machine Intelligence 2009 Dec 16 (pp. 369-374). Springer Berlin Heidelberg.

[17] Shinde AA, Chougule DG. Text Pre-processing and Text Segmentation for OCR. International Journal of Computer Science Engineering and Technology. 2012:810-2.

[18] Trier ØD, Jain AK, Taxt T. Feature extraction methods for character recognition-a survey. Pattern recognition. 1996 Apr 30;29(4):641-62.

[19] Pradeep J, Srinivasan E, Himavathi S. Diagonal based feature extraction for handwritten character recognition system using neural network. InElectronics Computer Technology (ICECT), 2011 3rd International Conference on 2011 Apr 8 (Vol. 4, pp. 364-368). IEEE.

[20] Bishnu A, Bhattacharya BB, Kundu MK, Murthy CA, Acharya T. A pipeline architecture for computing the Euler number of a binary image. Journal of Systems Architecture. 2005 Aug 31;51(8):470-87.

[21] Dinesh Acharya U, Subbareddy NV. Krishnamoorthy: Isolated Kannada Numeral Recognition Using Structural Features and K-Means Cluster. Proc. of IISN. 2007:125-9.

[22] Sharma OP, Ghose MK, Shah KB. An improved zone-based hybrid feature extraction model for handwritten alphabets recognition using euler number. International Journal of Soft Computing and Engineering. 2012 May;2(2):504-8.

[23] Suen CY. Character recognition by computer and applications. Handbook of pattern recognition and image processing. 1986:569-86.

[24] Rehman A, Saba T. Neural networks for document image preprocessing: state of the art. Artificial Intelligence Review. 2014 Aug 1;42(2):253-73.

[25] Dongre VJ, Mankar VH. A review of research on Devnagari character recognition. arXiv preprint arXiv:1101.2491. 2011 Jan 13.

[26] SMITH, R. 2007. An Overview of the Tesseract OCR Engine. In proceedings of Document analysis and Recognition.. ICDAR 2007. IEEE Ninth International Conference.

[27] GOOGLE. Google Code. google code. [Online] 2012. http://code.google.com/p/tesseract-ocr/.

[28] Chirag Patel, Atul Patel, PhD, Dharmendra Patel - Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study - International Journal of Computer Applications (0975 – 8887) Volume 55– No.10, October 2012

[29] Rafi Ullah ,Ali Sohani,Faraz Ali, Athaul Rai - OCR Engine to extract Food-items and Prices from Receipt Imagesvia Pattern matching and heuristics approach.

[30] Harshit Sidhwa, Sudhanshu Kulshrestha, Sahil Malhotra, Shivani Virmani Text Extraction from Bills and Invoices - International Conference on Advances in Computing, Communication Control and Networking (ICACCCN2018)

[31] Canny, J F, Finding Edges and Lines in Images, MIT technical report AI-TR-720,1983.

[32] Shah P, Karamchandani S, Nadkar T, Gulechha N, Koli K, Lad K. OCR-based chassis-number recognition using artificial neural networks. InVehicular Electronics and Safety (ICVES), 2009 IEEE International Conference on 2009 Nov 11 (pp. 31-34). IEEE.

[33] Zhai X, Bensaali F, Sotudeh R. OCR-based neural network for ANPR. In2012 IEEE International Conference on Imaging Systems and Techniques Proceedings 2012 Jul 16 (pp. 393-397). IEEE.

[34] Shamsher I, Ahmad Z, Orakzai JK, Adnan A. OCR for printed urdu script using feed forward neural network. InProceedings of World Academy of Science, Engineering and Technology 2007 Aug (Vol. 23, pp. 172-175).

[35] Yetirajam M, Nayak MR, Chattopadhyay S. Recognition and classification of broken characters using feed forward neural network to enhance an OCR solution. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume. 2012 Oct 28;1.

[36] Ruvan Weerasinghe. “Developing a commercial grade Tamil OCR for recognizing font and size independent text”, 130-134.

[37] Prabu. “Understanding of Convolutional Neural Network (CNN) — Deep Learning” https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

[38] Priyanka Patel. “Convolutional Neural Nets” https://medium.com/@priyankapatel2205/convolutional-neural-nets-1813eee0510

[39] henkboelman. “Serverless AI with Custom Vision & Azure Functions” https://www.henkboelman.com/articles/serverless-ai-with-custom-vision-and-azure-functions/

[40] Christopher Olah. “Understanding LSTM Networks” https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[41] Prasoo Singh. “LSTM- Long Short-Term Memory” https://medium.com/analytics-vidhya/lstm-long-short-term-memory-5ac02af47606

[42] apache jspωiki. “Long Short Term Memory networks”
https://ldapwiki.com/wiki/Long%20Short%20Term%20Memory%20networks

[43] Chris Nicholson. “A Beginner’s Guide to Important Topics in AI, Machine Learning, and Deep Learning.”
https://pathmind.com/wiki/lstm

[44] Rajib Rana. “Gated Recurrent Unit (GRU) for Emotion Classification from Noisy Speech”
https://www.semion.io/doc/gated-recurrent-unit-gru-for-emotion-classification-from-noisy-speech

[45] Sanket Gupta. “Overview of Text Similarity Metrics in Python”
https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50

[46] Suphakit Niwattanakul*, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu - Using of Jaccard Coefficient for Keywords Similarity. Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong