A.I. Based Document Digitization

Sujoy Dev
Department of Information Technology
A. P. Shah Institute of Technology
Thane, India
sujoydev99@gmail.com

Rashmi Shetty
Department of Information Technology
A. P. Shah Institute of Technology
Thane, India
shettyrashmi085@gmail.com

Priya Naik
Department of Information Technology
A. P. Shah Institute of Technology
Thane, India
naikpriya1999@gmail.com

Kiran B. Deshpande
Department of Information Technology
A. P. Shah Institute of Technology
Thane, India
kbdeshpande@apsit.edu.in

Sameer Nanivadekar
Department of Information Technology
A. P. Shah Institute of Technology
Thane, India
deanadmin@apsit.edu.in

***Abstract*— One of the crucial challenges faced by every industrial organization is the maintenance of records, mainly non-digitized type, i.e. hard copies and prints. While a few documents can be obtained in a pre-digitized format, majority of the documents which are crucial in nature remain non-digitized. Such documents are majorly billing invoices, which contain all the information about the seller, consumer, products, prices/taxes etc. Extraction of such data requires manual labour which is prone to human error and expensive. Hence data extraction systems are becoming increasingly important for cost cutting, efficient data processing and analysis. This paper presents a application that can be built using modern software technologies that can be used to automate the process of data extraction from invoices (single/bulk) by performing image processing,character/pattern recognition. This application will provide notifications to it's users about the metadata extracted from the documents such as the Invoice ID, amount, etc. as a reminder with the help of date parameters, if any. This application will also allow training of a centralized model to extract newer fields which were previously unidentified.**

Keywords*— *Named Entity Recognition (NER), Image Processing, Assistive Technology, Natural Language Processing (NLP), Optical Character Recognition (OCR), Data Extraction

I. Introduction

In today's fast paced world every thing has moved to the "Cloud". Everything from reports to credit cards are available online, yet a majority of the documents are still processed via manual paperwork. There are various methods of storing these documents, but none provide an efficient knowledge-base. Organizations both, small and large, process a large number of documents on a daily basis with an lot of incorrect input. Some choose to work with similar softwares that cost a fortune to run, thus pulling small scale organizations out of the race. Yet these softwares lack important features such as insight collection and reporting. Also they fail to recognize document which are of an unknown type.

Businesses, both large and small, receive a large number of document invoice and bills in printed format. Manual data extraction is a slow and cumbersome process which still revolves around the copy-paste methodology. It requires a moderately large workforce and is not error free. Frequent errors such as incorrect data and spelling errors lead to data inconsistency. Furthermore it is a time consuming process, this is one of the leading cause of business failures. A manufacturing unit with slow document processing accompanied by incorrect data evaluation will give poor results, which may drive the business away, hence incurring loss.

By automating data entry, OCR and data capture tools can lower costs, eliminate keying errors, and streamline the approval and payment cycle [6]. According to Wikipedia, ' Intelligent Data Capture (IDC) or learning systems enable end users to extract content from invoices without the system having to learn the layout of the invoice.Some intelligent engines are able to correctly sort batches on the fly, locate data fields such as invoice and PO number, as well as line item information, and then extract the desired content from those data fields. Intelligent solutions do not require the coding of rules or design form templates. Rather the system learns by reviewing a relatively small number of invoice samples. This helps the system scale to large invoice volumes and widely varying document layouts without requiring a human operator' [7].

With the advent of machine learning, image and language processing it is now possible to scan any image and extract just about everything. We can leverage the existing libraries, developed over time, to implement an automated invoice scanning and extraction pipeline. Computer vision can scan, eliminate noise, and extract localized text areas. Natural Language Processing (NLP) can extract meaningful data from arbitrary sentences, i.e., it mimics the human lexical capabilities. The combination of both the said technologies in an application with notification features will can prove to be a cost effective solution for document digitization.

II. Literature Review

Lu, H. et al. [1], have conceived a method to for better preprocessing of shadowed text images, for which the character recognition performance of Tesseract drops significantly. In this paper, we propose a new method to process the shadowed text images for the Tesseract's optical character recognition engine. First, they performed a local adaptive thresholding to transform the document to gray-scale image into a binary image to capture the contours of texts. Now in order to get rid of the salt-and-pepper noise in the shadow areas they applied a double-filtering algorithm, in which a vertical and horizontal projection method is method is used to remove the noise between texts and after that median filter removes the noise within characters.This type of preprocessed data when provided to Tesseract OCR produces much better result.

Sidhwa et al. [2], are using OpenCV to extract the bill or invoice document from an image buy substracting the background after finding the largest rectangular contour. Next they perform line segmentation in which they scan the document horizontally, then they perform line segmentation. However due to the absence of any NLP algorithm their process of data extraction only returns the test but no lexical definition.

K.M. Yindumathi et al. [3], have studied and compared the various methods,metrics and algorithms for text extraction from bills and invoices. According to their findings OpenCV, Tessseract are able to extract text from most of the images but when it comes to handwritten documents they fall short.

Zhang, J et al. [4], have extracted textual data from VAT invoices by preforming image preprocessing, OCR and localizing the data of a particual region of interest (ROI). In the preprocessing stage they performed skew-correction by applying local adaptive thresholding and Hough transformations. The ROI was segmented and extracted with the help of the projection method. They remove stamps from the VAT invoice with by elimination of the red color channel, i.e., the color of the stamps.

Wei Ruan and Won-sook Lee [5], built a Named Entity Recognition medical imaging procedure recognition system based on conditional random fields (CRF) model with word-based, part-of-speech. The NER model has been trained on a custom annotated dataset of medical notes from I2B2 with the F1 score up to 0.923 for recognizing medical imaging procedure entities. This system can be used to add and recognize new entities by simply creating new medical entities to extract data from the medical notes/documents.

III. Problem Statement

Document maintenance is a difficult and cumbersome task. Everyday huge chunks of data are misplaced due to improper handling of documents during and after their need. This must be very disruptive since there may be loss of confidential or important data which my potentially be harmful to any organization.

In this research paper we propose to create automated system for secure digitization and maintenance of documents. Institutions/users must be able to upload documents(expense bills/ invoices etc.) and the important data from these documents must be extracted and stored for later use. If the data is not extracted from the documents the users must be able to select the data themselves. Send the users insights about their document/invoices ,i.e., expiration, validity, etc. The application will store the data extraction model in a centralized repository. This centralized model will retrain at regular intervals to keep the models updated. This will increase the model accuracy over time as well as the number the entities that can be recognized. A web based user interface will be provided to upload the invoices(single or bulk) and view the extracted information in unstructured format.

In order to extract data from invoices the application will make use of Named entity Recognition(NER) which is a subset of natural language processing and openCV image processing. With the help of openCV we will perform optical character recognition on the document/invoice, furthermore we will extract tabular data using custom algorithms. The OCR data will be passed to the NER service for extraction of important Invoice metadata, which may be used later for analysis.

IV. proposed system

The proposed software application will automate the process off invoice data extraction. The application will be running on the browser connected to a server hosted on the cloud. Users will be able to register themselves and start using its extraction features immediately. In case the invoice has entities such as 'due dates' and 'due amount' the user will be notified via an email or a text SMS. In the first step of the extraction process the user will upload the documents onto the application interface. These documents will be saved on a secure cloud-based storage, in *.pbm* format, without any public access to maintain data integrity and confidentiality.

A. Image Preprocessing

Using the OpenCV library we can perform Preprocessing on the invoice image. Preprocessing in essential to optimize the image for the best possible extraction [8].

Binarization: We process the image by using Gaussian blur to reduce noise. Binarization converts a colored image into an image which account consists only of black and white pixels. This can be done by fixing a threshold (normally threshold=127, as it is exactly half of the pixel range 0–255). If the pixel value is greater than the threshold, it is considered as a white pixel, else considered as a black pixel.But since

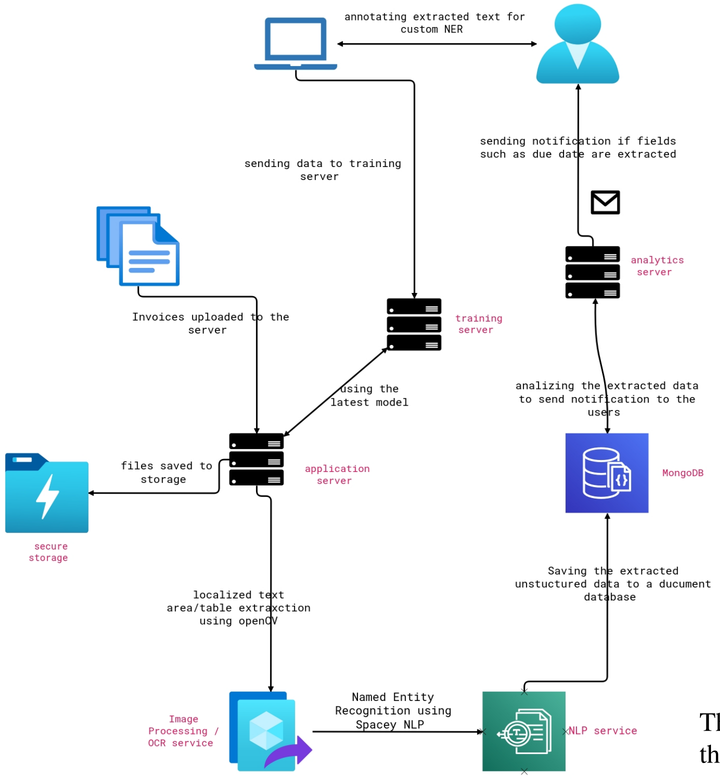

annotating extracted text for custom NER
sending data to training server
sending notification if fields such as due date are extracted
analytics server
Invoices uploaded to the server
training server
using the latest model
analizing the extracted data to send notification to the users
files saved to storage
application server
MongoDB
secure storage
localized text area/table extraxction using openCV
Saving the extracted unstuctured data to a ducument database
Image Processing / OCR service
Named Entity Recognition using Spacey NLP
NLP service

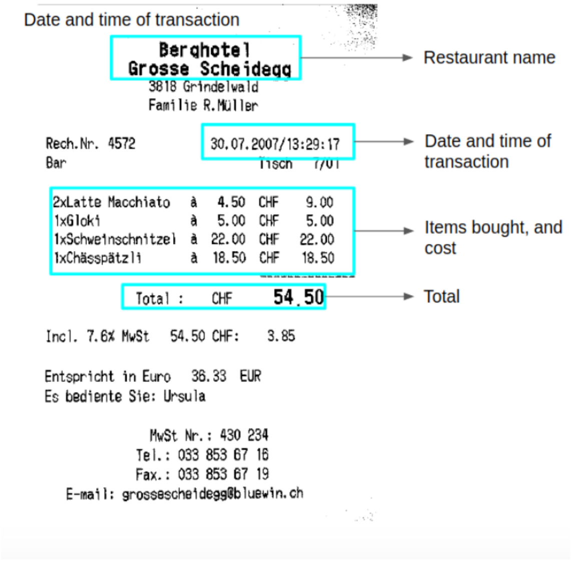
Fig. 1. Block Diagram of Proposed System Architecture

image may vary in pixel values throughout the image we use Adaptive Thresholding. It gives the threshold value for a small part of the image depending upon the characteristics of its locality and neighbors. There will not be a single fixed threshold for the whole image but every part of the image will have different thresholds depending upon the current region of Interest (ROI).

Skew Correction: Scanned documents may be slightly skewed or misaligned This will affect the performance of the OCR tool. Hence we apply skew correction. The image is rotated with the projection profile method, until it's alignment is corrected .

Thinning and Skeletonization: in case of handwritten invoices, thinning and skeletonization may be required to normalize the uneven brush strokes.

Segmentation: Next we find the largest Contours(tables) in the images and save them temporary arrays. With this method openCV will be able to extract multiple tables from the invoice. upon extraction these tables will be replaced by white spaces. By implementing Hough transforms, we extract the grid lines from the tables. using the grid lines as coordinates we create a ROI in each cell and extract the cell information, via premature OCR, save it in JSON format along with table headers. Next we dilate the all the Contours so that we can extract localized text areas from the image. These localized text areas may depict the addresses, we save them as well.


Date and time of transaction
Restaurant name
Date and time of transaction
Items bought, and cost
Total

Berghotel
Grosse Scheidegg
3818 Grindelwald
Familie R.Müller

Rech.Nr. 4572 30.07.2007/13:29:17
Bar Tisch 7/01

2xLatte Macchiato	à	4.50	CHF	9.00
1xGloki	à	5.00	CHF	5.00
1xSchweinschnitzel	à	22.00	CHF	22.00
1xChässpätzli	à	18.50	CHF	18.50

Total : CHF 54.50

Incl. 7.6% MwSt 54.50 CHF: 3.85

Entspricht in Euro 36.33 EUR
Es bediente Sie: Ursula

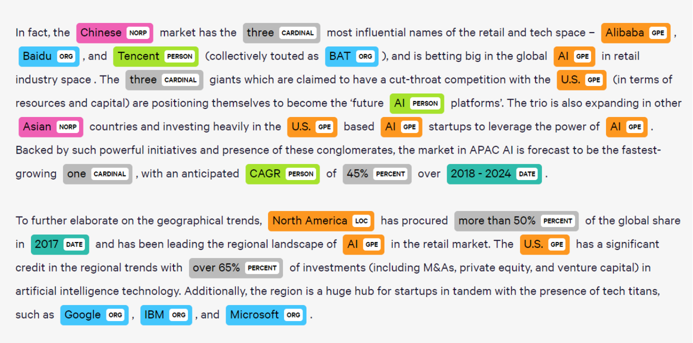
MwSt Nr.: 430 234
Tel.: 033 853 67 16
Fax.: 033 853 67 19
E-mail: grossescheidegg@bluewin.ch

Fig. 2. ROI Extraction

B. Optical Character Recognition

Once the preprocessing is complete Tesseract OCR or Google-Vision API is used for extraction of pure textual data from the segments. We apply OCR to all the extracted contours. This returns with the textual data optionally we may Preserve the white-spaces(optional).

C. Named Entity Recognition



In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

Fig. 3. Named Entity Recognition on text extracted by OCR

A Named Entity Recognition(NER) model pre-trained with spaceyNLP trianed on the ICADR-2019-SROIE dataset of 626 invoices [9], will be used to extract named entities from the segments. Entities may include the 'bill amount' which may be depicted in the document

by =⟩ 'total amount' — 'total' — 'amount' — 'amount';
invoice number=⟩ 'ID' — 'regNo' — 'billnumber';

address=⟩ ’from’ — ’to’ etc.

These are the named entities. This step will extract named entities from the text data from the OCR step. These named entities contain all the document metadata. This data is highly unstructured since invoice may be of varying type. Hence we use a NOSql database such as MongoDB to store the data. Till this stage the application server handles all the processes and is exposed to the other services.

D. Retraining Spacey

Case: unknown entity, custom entity: In case the user feels that new entities must be added do the model to extract even more data from the invoices, they can choose to do so by selecting one of the uploaded documents which will undergo OCR and appear resulting data will appear in an integrated data annotation tool such as Datatux. The annotation data files along with the images will persist at the Training Server. Automatic Cron events will retrain the space NER model at regular intervals. Such a system will make sure that the centralized model doesn’t have to depend on the creator for training. In such an environment it will enter into a self sustainable life cycle.In order to preserve consistency validation rules will have to be set for the training requests by users, on this server itself.

E. Notification System

An analysis engine/server is used to read the data stored on the server, it will send the invoice document owner insight about his/her billing habits, upcoming due date for pending invoices and more. Either a text SMS or an email can be used for this purpose.

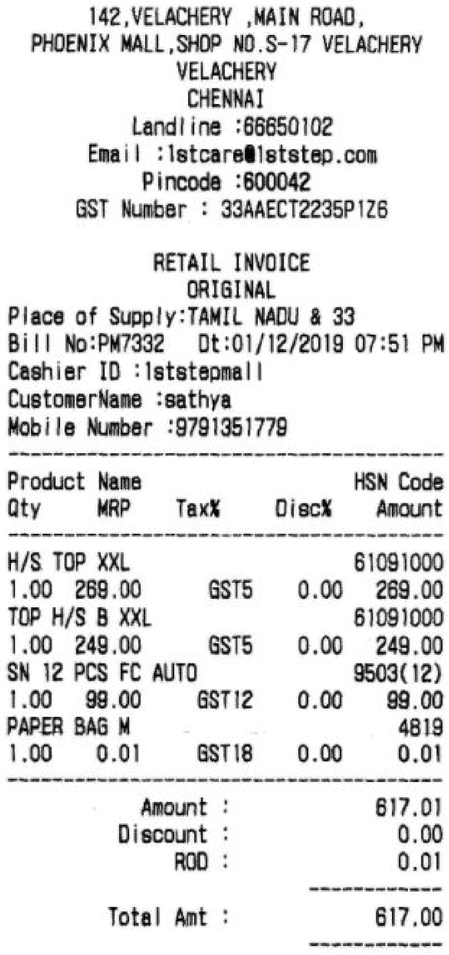
V. Result

The proposed solution was test for an unknown restaurant invoice. We obtained an accuracy of F1 score of 0.663 with just over 600 invoices of massively varying entity classes. This result would improve with the increase in the data-set size. The predictions would be better in case of industrial bull and invoices with some standard layout and jargon.

In the current scenario the result has a rather poor F1 score. With time the the self growing model would increase in size with more annotated entities and more invoice documents. Gradually due to increase in the size of the data-set the predictions would become better giving more accurate results.

Input =⟩ raw file

Output =⟩ JSON object

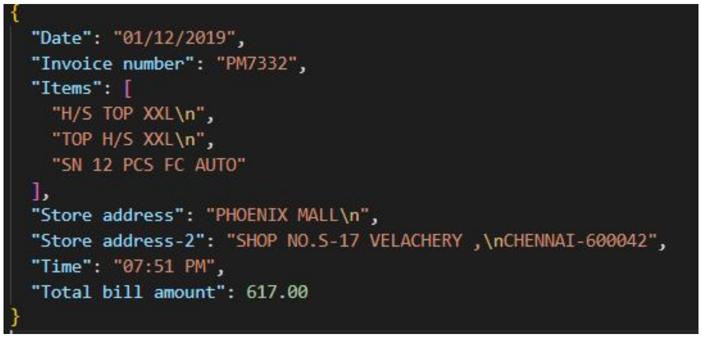


```
     142,VELACHERY ,MAIN ROAD,
PHOENIX MALL,SHOP NO.S-17 VELACHERY
             VELACHERY
              CHENNAI
       Landline :66650102
    Email :1stcare@1ststep.com
         Pincode :600042
  GST Number : 33AAECT2235P1Z6

           RETAIL INVOICE
              ORIGINAL
Place of Supply:TAMIL NADU & 33
Bill No:PM7332   Dt:01/12/2019 07:51 PM
Cashier ID :1ststepmall
CustomerName :sathya
Mobile Number :9791351779
---------------------------------------
Product Name                   HSN Code
Qty      MRP     Tax%     Disc%  Amount
---------------------------------------
H/S TOP XXL                    61091000
1.00  269.00     GST5     0.00   269.00
TOP H/S B XXL                  61091000
1.00  249.00     GST5     0.00   249.00
SN 12 PCS FC AUTO              9503(12)
1.00   99.00     GST12    0.00    99.00
PAPER BAG M                        4819
1.00    0.01     GST18    0.00     0.01
---------------------------------------
            Amount :             617.01
          Discount :               0.00
               ROD :               0.01
                           ------------

         Total Amt :             617.00
                           ------------
```

Fig. 4. Sample Invoice Input



```
{
  "Date": "01/12/2019",
  "Invoice number": "PM7332",
  "Items": [
    "H/S TOP XXL\n",
    "TOP H/S XXL\n",
    "SN 12 PCS FC AUTO"
  ],
  "Store address": "PHOENIX MALL\n",
  "Store address-2": "SHOP NO.S-17 VELACHERY ,\nCHENNAI-600042",
  "Time": "07:51 PM",
  "Total bill amount": 617.00
}
```

Fig. 5. Unstructured JSON output

VI. Conclusion

The main objective of building such an application in to increase work efficiency, speed and to reduce manual labour to a minimum. In today’s world data is of utmost importance and in order to keep up with the speed at which it is being generated we must move towards automation. Digitization comes with it’s own advantages like data persistence, little to no maintenance, everything becomes readily available at our finger tips and most importantly it it has low cos of operation. As the users of the application can generate custom annotations for their own documents they will indirectly help in increasing the data-set size, this will help other users as well. Also the entire application is leveraging Open Source technologies the application can itself remain Open Source, free for any one to use. This will encourage application as well as community growth.

References

[1] Lu, H., Guo, B., Liu, J., Yan, X. (2017). *A shadow removal method for tesseract text recognition. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI).*

[2] Sidhwa, H., Kulshrestha, S., Malhotra, S., Virmani, S. (2018). *Text Extraction from Bills and Invoices. 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN).*

[3] K.M. Yindumathi, Shilpa Shashikant Chaudhari, R. Aparna. (2020). *Analysis of Image Classification for Text Extraction from Bills and Invoices. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT).*

[4] Zhang, J., Ren, F., Ni, H., Zhang, Z., Wang, K. (2019). *Research on Information Recognition of VAT Invoice Based on Computer Vision. 2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS).*

[5] Wei Ruan; Won-sook Lee (2018). *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*

[6] Internet Archive, *https://web.archive.org/web/20081223141143/http://www.paystreamadvisors.com/store/details.cfm?id=278, last accessed 28.7.20*

[7] Wikipedia, *https://en.wikipedia.org/wiki/Invoice_processing last accessed 28.7.20*

[8] towards data science, *https://towardsdatascience.com/pre-processing-in-ocr-fc231c6035a7, last accessed 19.10.20*

[9] Dataset, *https://github.com/zzzDavid/ICDAR-2019-SROIE.git*