JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Data Extraction using Optical Character Recognition and Natural Language Processing

[1]**Shagun Davessar**
[1]Student
[1]Computer Science Engineering specialization in Artificial Intelligence and Machine Learning
[1]Manipal University Jaipur, Jaipur, India

Abstract : Over the recent years, there's been a boost in research on data extraction. In today's digitalized world, everything has become digital like Text files, Invoices, Bills, etc. Optical Character Recognition (OCR) is the technology which converts an image into a machine-readable text format, hence stores as text data. Nowadays, most businesses involve receiving information from print media, like invoices, printed contracts, etc. Therefore, manually this process takes a lot of time and can be tedious and so this project will help them. In this project, our aim is to extract whole data from Invoices along with the texts present inside the table. And classify them like dictionary, Key: Value pair.

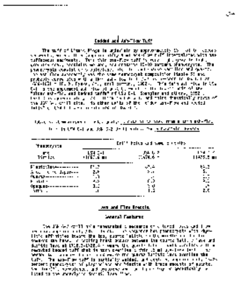
***IndexTerms* - Deep learning, OCR, Named Entity Recognition (NER), Natural Language Processing (NLP), YOLO, Data extraction**

I. INTRODUCTION

An Optical Character Recognition (OCR) is a powerful tool that is used to extract words and lines of text from scans and images. It uses machine learning to perform text extraction and is a deep learning AI-based technology [1][2]. It can identify text inside an image and turn it into an editable digital document [2]. OCR technology has made breakthroughs in text recognition, although it should not be confused with intelligent character recognition (ICR), which only works with handwritten text [3]. This technology can be used to enhance text extraction accuracy, as it employs several machine algorithms for pattern recognition to identify the presence and layout of the text in an image file, complementing data extraction to enhance text extraction accuracy [1][4][3]. This software can also help make digital data editable, like receipts, bills, or invoices [2]. The effectiveness and preciseness of OCR technology depends on the quality of the image being processed, and text extracted using OCR technology should be reviewed to ensure accuracy, which can be further enhanced by using computer vision techniques like box and line detection [3][5]. Beside OCR, few more deep learning technologies are used like, YOLO.

A. You Only Look Once (YOLO)

YOLO is an efficient object detection architecture that requires only one network pass. It does away with the necessity for several runs or a two-step procedure. By splitting the input image into a grid and predicting B bounding boxes with confidence scores for C classes per grid element, the YOLO object recognition technique finds all bounding boxes at once. Each bounding box prediction includes Pc, which indicates the model's confidence and accuracy. The box and by coordinates are the centers of the box with relation to the grid cell and the entire picture, respectively [11].



Figurè 1

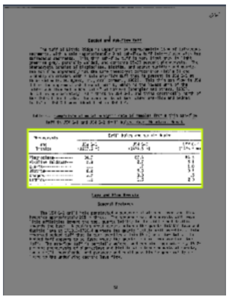


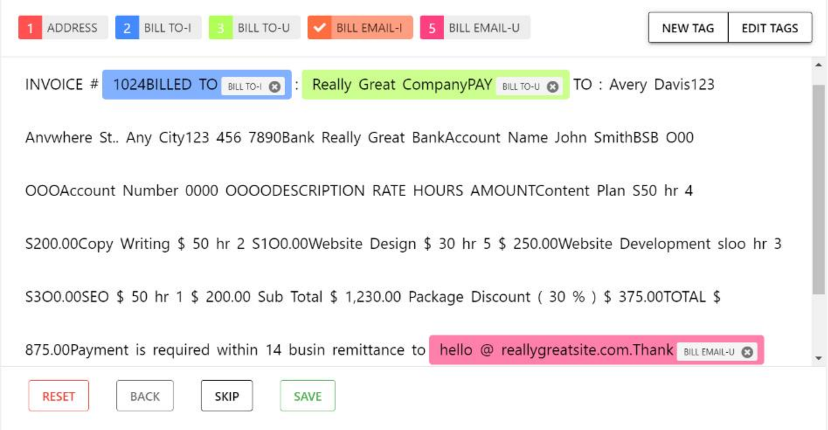
Figure 2

B. Natural Language Processing (NLP)

Natural language processing (NLP) is a type of machine learning that allows computers to interpret, manipulate, and comprehend human language. NLP can be used to interpret and analyze free, unstructured text. To process human language, it blends computational linguistics, machine learning, and deep learning models [12].

C. Named-Entity Recognition (NER)

NER is a subtask of information extraction that aims to identify and categories objects such as persons, locations, organizations, products, and so on. Its purpose is to automatically recognize and classify these named items in text, making the content easier to grasp for machine learning. NER is utilized in a wide Figure 1. Normal Image Figure 2. Image with annotation using YOLO range of applications, including data extraction, question answering, and machine translation.



INVOICE # 1024BILLED TO : Really Great CompanyPAY TO : Avery Davis123

Anvwhere St.. Any City123 456 7890Bank Really Great BankAccount Name John SmithBSB O00

OOOAccount Number 0000 OOOODESCRIPTION RATE HOURS AMOUNTContent Plan S50 hr 4

S200.00Copy Writing $ 50 hr 2 S1O0.00Website Design $ 30 hr 5 $ 250.00Website Development sloo hr 3

S3O0.00SEO $ 50 hr 1 $ 200.00 Sub Total $ 1,230.00 Package Discount (30 %) $ 375.00TOTAL $

875.00Payment is required within 14 busin remittance to hello @ reallygreatsite.com.Thank

Figure 3

II. Literature Survey

OCR technology has its origins in telegraphy. On the eve of World War I, scientist Emanuel Goldberg devised a computer capable of reading characters and converting them into telegraph code - an early version of optical character recognition [6]. Ray Kurzweil, an inventor, futurist, and the founder of Kurzweil Computer Products, created an omni-font OCR system, as well as the CCD flatbed scanner, which marked the commercial start of OCR [7]. The increasing use of personal computers and the internet in the 1990s resulted in a major growth in the use of OCR technology. Books, periodicals, and other printed items were digitised using this technology, making it simpler to search for and retrieve information [8]. In 2000s, Google also famously launched Google Books, code-named Project Ocean, using OCR to digitize tens of millions of books and make their text searchable, which shown improvement in the technology [8]. And in the current scenario, this technology has improved and become more sophisticated than ever before. It has progressed with the introduction of new algorithms and improved hardware [8]. In the recent researches, a few authors experimented text extraction using different algorithms like, OCR, Sentiment Analysis, Bag of Words, Robust algorithm, MSER (Maximum Stable Extremal Regions), SWT (Stroke Width Transformation) [9].

.

S.No.	Author/Paper Title	Method/Algorithm used	Accuracy (%)
1.	Sandhya Arora [11]	Intersection on features with Neural Networks	89.12
2.	R Sanjeev and R D Sudhakar [26]	Two stage Multi-Network (Neural Network)	91

Table 1

III. Methodology

Today's difficulties frequently cross traditional departmental lines, needing an interdisciplinary approach. This study bridges the gap between theoretical concepts and practical data extraction implementation, addressing the critical need for a systematic and effective approach to extracting valuable insights from complex datasets, ensuring a comprehensive view of data extraction from invoices using OCR technology.

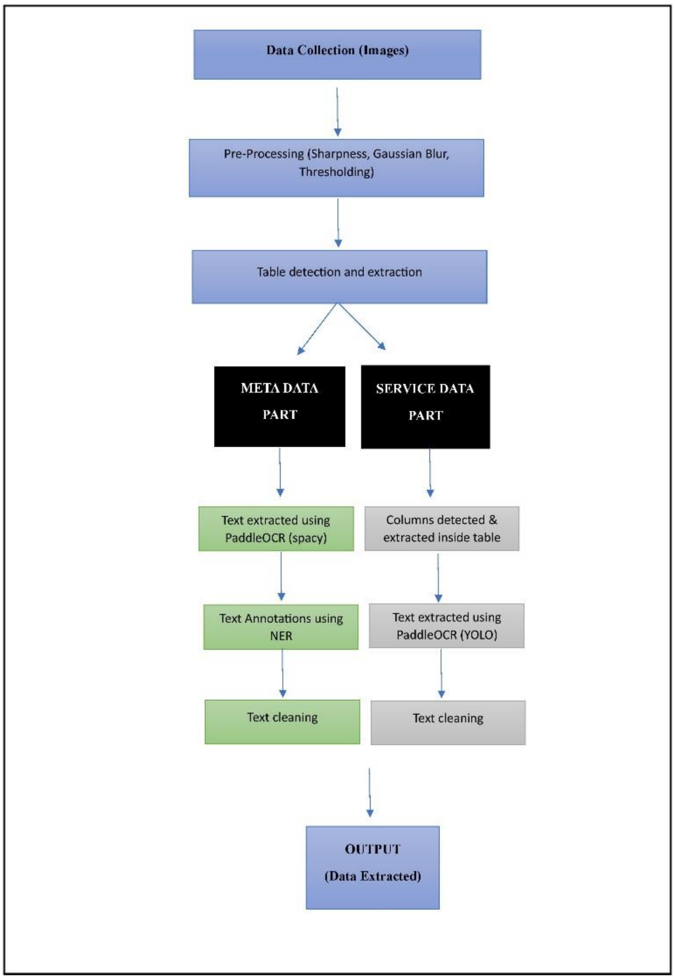
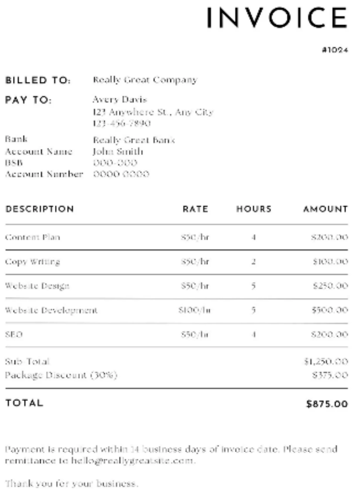

Data Collection (Images)
Pre-Processing (Sharpness, Gaussian Blur, Thresholding)
Table detection and extraction
META DATA PART
SERVICE DATA PART
Text extracted using PaddleOCR (spacy)
Columns detected & extracted inside table
Text Annotations using NER
Text extracted using PaddleOCR (YOLO)
Text cleaning
Text cleaning
OUTPUT
(Data Extracted)

Figure 4

3.1 Data Collection

In this section, dataset containing 300 images is loaded from Kaggle into google collab notebook. Fig. 6 shows an example of an adequate resolution image which would be acceptable.



INVOICE

#1024

BILLED TO: Really Great Company

PAY TO: Avery Davis
123 Anywhere St., Any City
123-456-7890

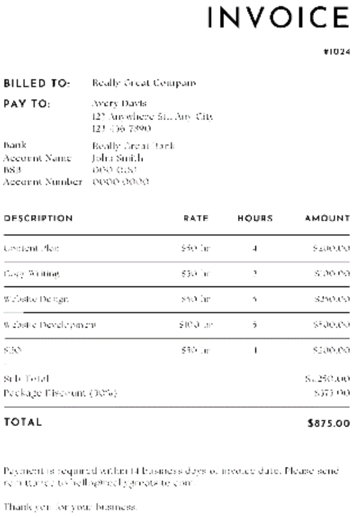
Bank Really Great Bank
Account Name John Smith
BSB 000-000
Account Number 0000 0000

DESCRIPTION	RATE	HOURS	AMOUNT
Content Plan	$50/hr	4	$200.00
Copy Writing	$50/hr	2	$100.00
Website Design	$50/hr	5	$250.00
Website Development	$100/hr	5	$500.00
SEO	$50/hr	4	$200.00
Sub-Total			$1,250.00
Package Discount (30%)			$375.00
TOTAL			$875.00

Payment is required within 14 business days of invoice date. Please send remittance to hello@reallygreatsite.com.

Thank you for your business.

Figure 6



INVOICE

#1024

BILLED TO: Really Great Company

PAY TO: Avery Davis
123 Anywhere St., Any City
123-456-7890

Bank Really Great Bank
Account Name John Smith
BSB 000-000
Account Number 0000 0000

DESCRIPTION	RATE	HOURS	AMOUNT
Content Plan	$50/hr	4	$200.00
Copy Writing	$50/hr	2	$100.00
Website Design	$50/hr	5	$250.00
Website Development	$100/hr	5	$500.00
SEO	$50/hr	4	$200.00
Sub-Total			$1,250.00
Package Discount (30%)			$375.00
TOTAL			$875.00

Payment is required within 14 business days of invoice date. Please send remittance to hello@reallygreatsite.com.

Thank you for your business.

Figure 5

3.2 Pre-Processing

In this section of paper, all the images go through preprocessing to achieve a better output with least loss function. Using OpenCV, enhancing these particular things:

- Sharpness: Applying sharpening filters can enhance the edges and details in an image.

- Blur: Used to reduce noise and unwanted details in an image, making it smoother and potentially improving the quality of subsequent image analysis or computer vision tasks.

- Thresholding: Used in image analysis and computer vision to convert a grayscale image into a binary image, where each pixel is classified as either foreground (object of interest) or background.

- Gaussian Blur: Used in image processing and computer vision to reduce noise and enhance important features in an image by applying a Gaussian filter.

Mathematically, Gaussian Blur G(x,y) is

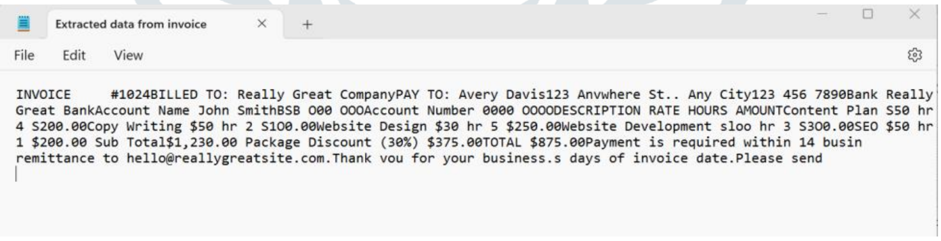
$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2-y^2)}{2\sigma}}$$

3.3 Text Extraction

This project is divided into 2 parts:

- Meta data (All content excluding Table)
- Service data (only content of Table)

Metadata is trained using an NLP library, Spacy whereas Service data using YOLO. For extracting the data, PaddleOCR has been used. To extract the data from inside the table can't be done directly, as it leads to improper formatting in the output which lowers the accuracy of the model. So, we detected columns inside tables. And then, extracted text data from columns directly using PaddleOCR, which shows better and optimized output.



```
INVOICE     #1024BILLED TO: Really Great CompanyPAY TO: Avery Davis123 Anvwhere St.. Any City123 456 7890Bank Really
Great BankAccount Name John SmithBSB 000 OOOAccount Number 0000 OOOODESCRIPTION RATE HOURS AMOUNTContent Plan S50 hr
4 S200.00Copy Writing $50 hr 2 S1O0.00Website Design $30 hr 5 $250.00Website Development sloo hr 3 S3O0.00SEO $50 hr
1 $200.00 Sub Total$1,230.00 Package Discount (30%) $375.00TOTAL $875.00Payment is required within 14 busin
remittance to hello@reallygreatsite.com.Thank vou for your business.s days of invoice date.Please send
```

Figure 7

3.4 Text Annotation using NER

Now, to annotate each and every key to their values, we are using Roboflow and training the set on 320 images. Annotation like, "*abc@xyz.com*" is seller's email address in extracted data. So, to annotate "*abc@xyz.com*" as seller email address, annotation is mandatory.

3.5 Text Cleaning

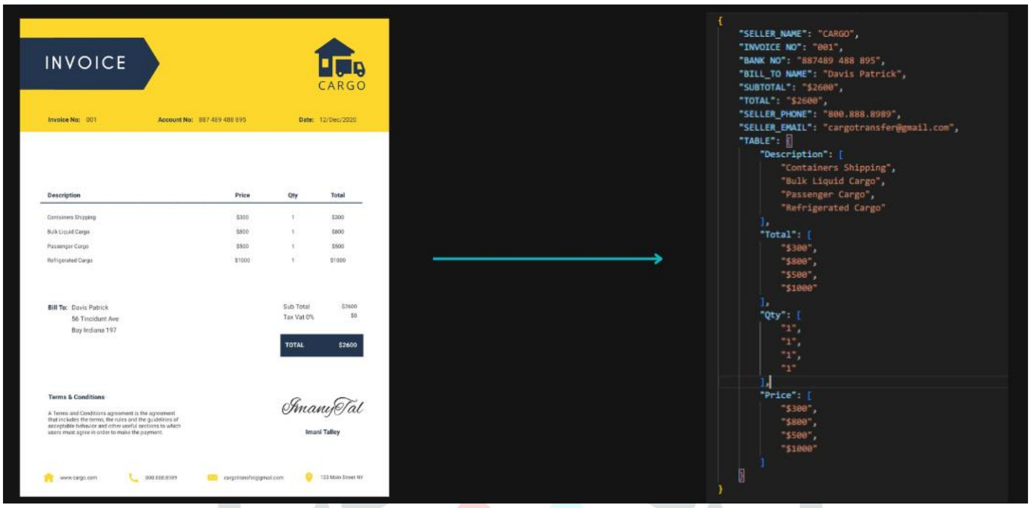
Now, text cleaning is an efficient way of achieving maximum accuracy. It involves various steps like,

- Lower case
- Tokenization (Separating words using comma)
- Removing stop words (I, Am, Are, etc.)

- Removing unnecessary symbols (^ or ~ etc.)
- Removing extra whitespaces.

IV. RESULTS AND DISCUSSION

This research has resulted to some pretty good outcomes, but it also claims that there is still a hope for development in the future. This approach extracts text and information from all printed invoices but fails with handwritten invoices. The overall accuracy achieved is 84% using PaddleOCR technology. Refer to Fig. 8. For the sample output. For more better results, larger dataset is required to train the model so that machine can predict on its own.



INVOICE

CARGO

Invoice No: 001 Account No: 887 489 488 895 Date: 12/Dec/2020

Description	Price	Qty	Total
Containers Shipping	$300	1	$300
Bulk Liquid Cargo	$800	1	$800
Passenger Cargo	$500	1	$500
Refrigerated Cargo	$1000	1	$1000

Bill To: Davis Patrick
56 Tincidunt Ave
Bay Indiana 197

Sub Total	$2600
Tax Vat 0%	$0
TOTAL	$2600

Terms & Conditions

A Terms and Conditions agreement is the agreement that includes the terms, the rules and the guidelines of acceptable behavior and other useful sections to which users must agree in order to make the payment.

ImanyTal

Imani Talley

www.cargo.com 800.888.8989 cargotransfer@gmail.com 123 Main Street NY

```
{
    "SELLER_NAME": "CARGO",
    "INVOICE NO": "001",
    "BANK NO": "887489 488 895",
    "BILL_TO NAME": "Davis Patrick",
    "SUBTOTAL": "$2600",
    "TOTAL": "$2600",
    "SELLER_PHONE": "800.888.8989",
    "SELLER_EMAIL": "cargotransfer@gmail.com",
    "TABLE": {
        "Description": [
            "Containers Shipping",
            "Bulk Liquid Cargo",
            "Passenger Cargo",
            "Refrigerated Cargo"
        ],
        "Total": [
            "$300",
            "$800",
            "$500",
            "$1000"
        ],
        "Qty": [
            "1",
            "1",
            "1",
            "1"
        ],
        "Price": [
            "$300",
            "$800",
            "$500",
            "$1000"
        ]
    }
}
```

Figure 8

V. ACKNOWLEDGMENT

I would like to express my gratitude to Mr. Sunil Patel (Faculty of CSE), Mr. Anil Kumar (Faculty of CSE) and Mr. Santosh Vishwakarma (Head of Dept. CSE AI ML) for their invaluable guidance and support throughout this research project. Their expertise and insights were instrumental in the successful completion of this work. I also want to thanks all my friends, and my family who generously shared their time and kept me motivated. Their contributions were essential to the success of this research.

REFERENCES

[1] Choosing the Right OCR Service for Extracting Text Data. (n.d.) Retrieved August 12, 2023, from urbaninstitute.medium.com
[2] 7 Ways to Convert Images to Text Using OCR. (n.d.) Retrieved August 12, 2023, from geekflare.com/convert-image-to-text/
[3] OCR Data Extraction: How to Use Computer Vision for Intelligent Data Extraction. (n.d.) Retrieved August 12, 2023, from labelyourdata.com/articles/ocr-data-extractionmethods
[4] Extract Text from Images Quickly Using Keras-OCR Pipeline. (n.d.) Retrieved August 12, 2023, from www.analyticsvidhya.com
[5] Copy text from pictures and file printouts using OCR in OneNote. (n.d.) Retrieved August 12, 2023, from support.microsoft.com
[6] How OCR technology developed, from oneadvanced.com/news-and-opinion/opticalcharacter-recognition-ocr-technology-a-briefhistory/
[7] The beginning of OCR: 1960s-1980s from, veryfi.com/ocr-api-platform/history-of-ocr/
[8] The beginning of OCR: 1990s- Early 2000s from, veryfi.com/ocr-api-platform/history-of-ocr/
[9] Some examples of Techniques used for Text Extraction from Image, from amygb.ai/blog/extracttext-from-images
[10] https://aws.amazon.com/what-is/ocr/
[11] https://deci.ai/blog/history-yolo-object-detectionmodels-from-yolov1-yolov8/
[12] https://aws.amazon.com/what-is/nlp/