PROJECT REPORT

Data Retrieval from Tax Invoices using OCR

ANIMESH PADHY,
MLL INTERN

TABLE OF CONTENTS

1. Introduction:

Computer vision has gained attention as a data-reliant method for feature extraction. Visualization techniques aid machines in understanding images, with Optical Character Recognition (OCR) being a key technology. OCR swiftly extracts and recognizes text from images using an existing database. While various OCR methods exist, they often fail to deliver the desired text format. Among global OCR options, Keras OCR's line segmentation and Easy OCR's space issues are less preferred, with Tesseract OCR being the top open-source choice, often used with Python's pytesseract library. Tesseract OCR excels at extracting text based on invoice formats.

Key OpenCV techniques include thresholding for image segmentation, pixel-based computation for black-and-white conversion, and border blurring for emphasis. Smoothing evens out rough edges for text blending. Regular expressions further clean extracted characters. The cleaned text is converted into JSON and CSV formats for better understanding.

JSON returns objects from the server for web developers, commonly used to display output on web pages. CSV, a comma-separated format, creates tabular columns akin to Excel sheets.

2. Problem Statement:

In manual financial data entry processes, the conversion of tax invoice images into structured data formats like Excel spreadsheets is time-consuming and error-prone. The need for accurate extraction and formatting of information from these images poses a challenge. There is a requirement for an automated solution that can efficiently extract text from tax invoice images, organize it into a structured layout, and map it to a digital format like an Excel spreadsheet.

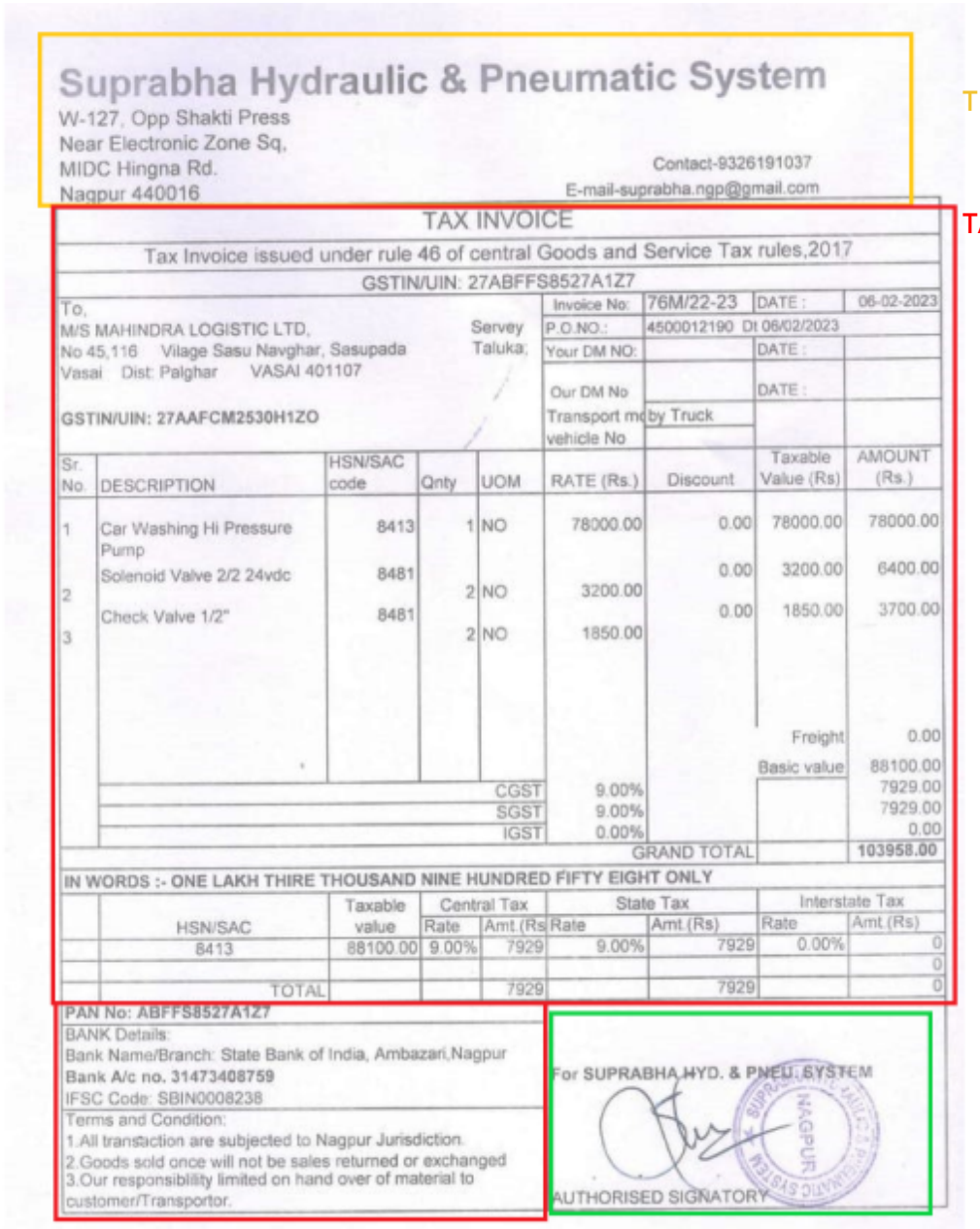
3. Objective:

The primary objective of this project is to develop an automated system that utilizes Optical Character Recognition (OCR) technology to convert tax invoice images into an Excel spreadsheet format. The project aims to streamline the process of handling financial documents.

4. Working:

4.1. Layout Analysis

In the project's initial phase, layout analysis is done using the PP-picodet model from PaddleDetection. This supports automatically segmenting tax invoice images, and identifying key elements like text, titles, and tables. The layout detection model is integrated to identify layouts from images, forming a strong basis for further data processing and conversion.



Suprabha Hydraulic & Pneumatic System
W-127, Opp Shakti Press
Near Electronic Zone Sq,
MIDC Hingna Rd.
Nagpur 440016
Contact-9326191037
E-mail-suprabha.ngp@gmail.com

TAX INVOICE

Tax Invoice issued under rule 46 of central Goods and Service Tax rules,2017

GSTIN/UIN: 27ABFFS8527A1Z7

To,
M/S MAHINDRA LOGISTIC LTD,
No 45,116 Vilage Sasu Navghar, Sasupada
Vasai Dist: Palghar VASAI 401107
GSTIN/UIN: 27AAFCM2530H1ZO

Servey
Taluka:

Invoice No: 76M/22-23 DATE : 06-02-2023
P.O.NO.: 4500012190 Dt 06/02/2023
Your DM NO: DATE :
Our DM No DATE :
Transport mode by Truck
vehicle No

Sr. No.	DESCRIPTION	HSN/SAC code	Qnty	UOM	RATE (Rs.)	Discount	Taxable Value (Rs)	AMOUNT (Rs.)
1	Car Washing Hi Pressure Pump	8413	1	NO	78000.00	0.00	78000.00	78000.00
2	Solenoid Valve 2/2 24vdc	8481	2	NO	3200.00	0.00	3200.00	6400.00
3	Check Valve 1/2"	8481	2	NO	1850.00	0.00	1850.00	3700.00
							Freight	0.00
							Basic value	88100.00
				CGST	9.00%			7929.00
				SGST	9.00%			7929.00
				IGST	0.00%			0.00
							GRAND TOTAL	103958.00

IN WORDS :- ONE LAKH THIRE THOUSAND NINE HUNDRED FIFTY EIGHT ONLY

HSN/SAC	Taxable value	Central Tax Rate	Central Tax Amt (Rs	State Tax Rate	State Tax Amt (Rs)	Interstate Tax Rate	Interstate Tax Amt (Rs)
8413	88100.00	9.00%	7929	9.00%	7929	0.00%	0
							0
TOTAL			7929		7929		0

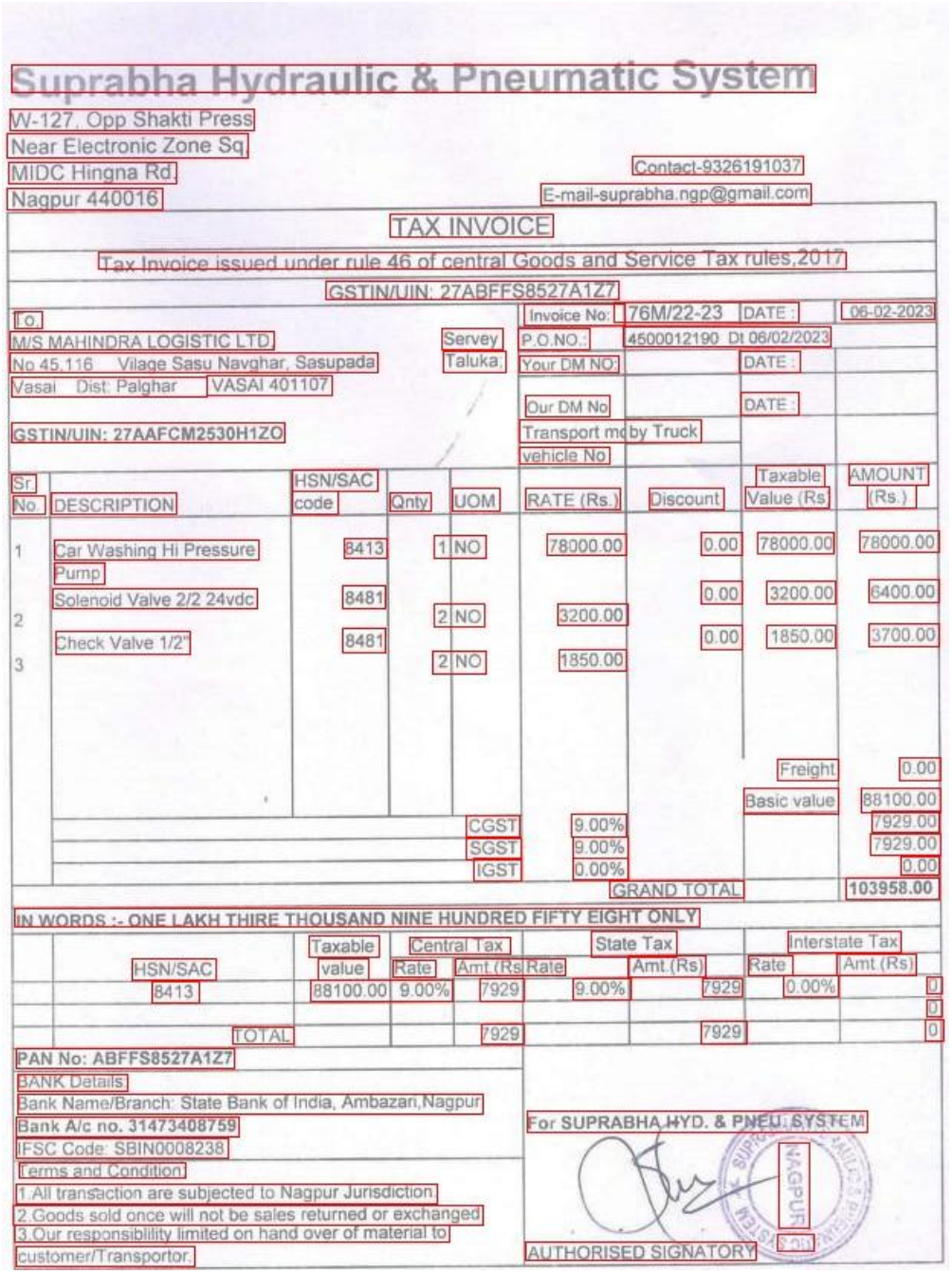
PAN No: ABFFS8527A1Z7
BANK Details:
Bank Name/Branch: State Bank of India, Ambazari,Nagpur
Bank A/c no. 31473408759
IFSC Code: SBIN0008238

Terms and Condition:
1.All transaction are subjected to Nagpur Jurisdiction.
2.Goods sold once will not be sales returned or exchanged
3.Our responsibility limited on hand over of material to customer/Transportor.

For SUPRABHA HYD. & PNEU. SYSTEM
AUTHORISED SIGNATORY

4.2. Text Detection and Recognition

We proceeded with text detection and recognition using PaddleOCR. This step involves extracting text content from the previously identified tables. This process streamlined the data acquisition process, reducing manual efforts and enhancing the accuracy and reliability of the extracted text data.



Suprabha Hydraulic & Pneumatic System

W-127, Opp Shakti Press
Near Electronic Zone Sq
MIDC Hingna Rd,
Nagpur 440016

Contact-9326191037
E-mail-suprabha.ngp@gmail.com

TAX INVOICE

Tax Invoice issued under rule 46 of central Goods and Service Tax rules,2017

GSTIN/UIN: 27ABFFS8527A1Z7

To,
M/S MAHINDRA LOGISTIC LTD
No 45,116 Vilage Sasu Navghar, Sasupada
Vasai Dist: Palghar VASAI 401107

GSTIN/UIN: 27AAFCM2530H1ZO

Servey
Taluka

Invoice No: 76M/22-23 DATE : 06-02-2023
P.O.NO.: 4500012190 Dt 06/02/2023
Your DM NO: DATE :
Our DM No DATE :
Transport mode by Truck
vehicle No

Sr No	DESCRIPTION	HSN/SAC code	Qnty	UOM	RATE (Rs.)	Discount	Taxable Value (Rs	AMOUNT (Rs.)
1	Car Washing Hi Pressure Pump	8413	1	NO	78000.00	0.00	78000.00	78000.00
2	Solenoid Valve 2/2 24vdc	8481	2	NO	3200.00	0.00	3200.00	6400.00
3	Check Valve 1/2"	8481	2	NO	1850.00	0.00	1850.00	3700.00
							Freight	0.00
							Basic value	88100.00
				CGST	9.00%			7929.00
				SGST	9.00%			7929.00
				IGST	0.00%			0.00
						GRAND TOTAL		103958.00

IN WORDS :- ONE LAKH THIRE THOUSAND NINE HUNDRED FIFTY EIGHT ONLY

HSN/SAC	Taxable value	Central Tax Rate	Central Tax Amt.(Rs	State Tax Rate	State Tax Amt.(Rs)	Interstate Tax Rate	Interstate Tax Amt.(Rs)
8413	88100.00	9.00%	7929	9.00%	7929	0.00%	0
							0
TOTAL			7929		7929		0

PAN No: ABFFS8527A1Z7

BANK Details
Bank Name/Branch: State Bank of India, Ambazari,Nagpur
Bank A/c no. 31473408759
IFSC Code: SBIN0008238

Terms and Condition
1.All transaction are subjected to Nagpur Jurisdiction.
2.Goods sold once will not be sales returned or exchanged.
3.Our responsiblility limited on hand over of material to customer/Transportor.

For SUPRABHA HYD. & PNEU. SYSTEM

AUTHORISED SIGNATORY

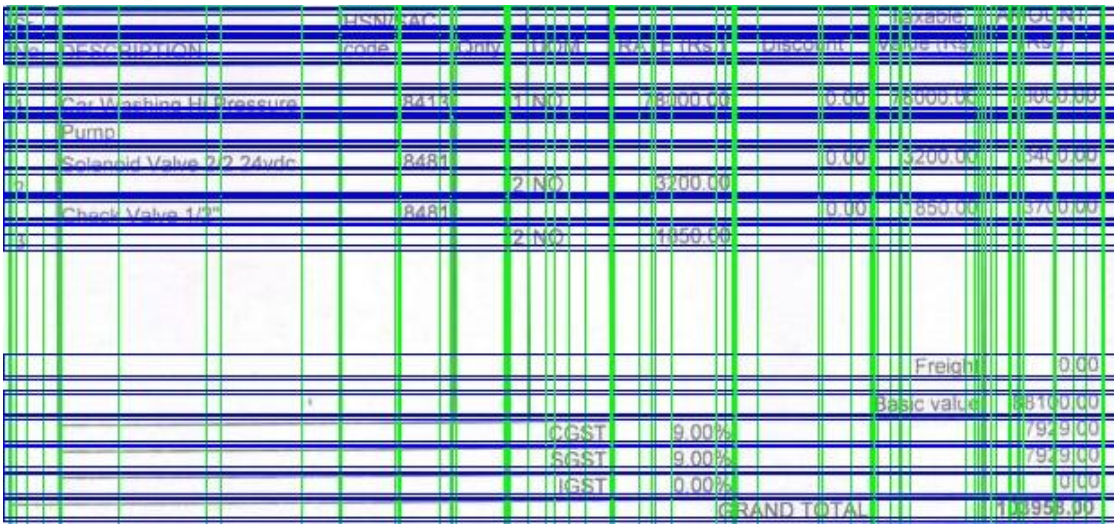
```
[[[3.0, 5.0], [22.0, 5.0], [22.0, 18.0], [3.0, 18.0]], ('Sr.', 0.9957569241523743)]
[[[188.0, 2.0], [243.0, 2.0], [243.0, 16.0], [188.0, 16.0]], ('HSN/SAC', 0.9949851632118225)]
[[[494.0, 1.0], [536.0, 1.0], [536.0, 12.0], [494.0, 12.0]], ('Taxable', 0.9966048002243042)]
[[[553.0, 1.0], [603.0, 1.0], [603.0, 11.0], [553.0, 11.0]], ('AMOUNT', 0.9885193705558777)]
[[[4.0, 20.0], [22.0, 20.0], [22.0, 32.0], [4.0, 32.0]], ('No', 0.9693977236747742)]
[[[31.0, 18.0], [113.0, 18.0], [113.0, 32.0], [31.0, 32.0]], ('DESCRIPTION', 0.9972743391990662)]
[[[187.0, 17.0], [218.0, 17.0], [218.0, 31.0], [187.0, 31.0]], ('code', 0.9899513721466064)]
[[[252.0, 17.0], [279.0, 17.0], [279.0, 29.0], [252.0, 29.0]], ('Qnty', 0.9940733909606934)]
[[[292.0, 15.0], [324.0, 15.0], [324.0, 29.0], [292.0, 29.0]], ('UOM', 0.9874367713928223)]
[[[340.0, 15.0], [402.0, 15.0], [402.0, 26.0], [340.0, 26.0]], ('RATERs.', 0.9217668771743774)]
[[[420.0, 15.0], [471.0, 13.0], [472.0, 27.0], [421.0, 29.0]], ('Discount', 0.9973067045211792)]
[[[487.0, 15.0], [541.0, 15.0], [541.0, 26.0], [487.0, 26.0]], ('Value Rs', 0.9287118911743164)]
[[[565.0, 13.0], [596.0, 13.0], [596.0, 27.0], [565.0, 27.0]], ('Rs.)', 0.857263445854187)]
[[[32.0, 49.0], [166.0, 49.0], [166.0, 62.0], [32.0, 62.0]], ('Car Washing Hi Pressure', 0.9650688767433167)]
[[[220.0, 46.0], [250.0, 46.0], [250.0, 61.0], [220.0, 61.0]], ('8413', 0.9985027313232422)]
[[[282.0, 46.0], [315.0, 46.0], [315.0, 60.0], [282.0, 60.0]], ('1NO', 0.9767632484436035)]
[[[354.0, 45.0], [407.0, 45.0], [407.0, 59.0], [354.0, 59.0]], ('78000.00', 0.9944804310798645)]
[[[456.0, 45.0], [483.0, 45.0], [483.0, 57.0], [456.0, 57.0]], ('0.00', 0.9933117628097534)]
[[[494.0, 45.0], [543.0, 45.0], [543.0, 56.0], [494.0, 56.0]], ('78000.00', 0.995461106300354)]
[[[560.0, 43.0], [612.0, 43.0], [612.0, 57.0], [560.0, 57.0]], ('78000.00', 0.9944612979888916)]
[[[31.0, 64.0], [64.0, 64.0], [64.0, 77.0], [31.0, 77.0]], ('Pump', 0.9910174012184143)]
[[[31.0, 80.0], [166.0, 79.0], [166.0, 93.0], [31.0, 94.0]], ('Solenoid Valve 2/224vdc', 0.938427746295929)]
[[[220.0, 78.0], [249.0, 78.0], [249.0, 93.0], [220.0, 93.0]], ('8481', 0.9972835183143616)]
[[[456.0, 77.0], [483.0, 77.0], [483.0, 89.0], [456.0, 89.0]], ('0.00', 0.9837182760238647)]
[[[499.0, 75.0], [545.0, 75.0], [545.0, 89.0], [499.0, 89.0]], ('3200.00', 0.9952859878540039)]
[[[566.0, 74.0], [612.0, 74.0], [612.0, 89.0], [566.0, 89.0]], ('6400.00', 0.9923862814903259)]
[[[280.0, 90.0], [315.0, 90.0], [315.0, 105.0], [280.0, 105.0]], ('2NO', 0.9723656177520752)]
[[[361.0, 90.0], [407.0, 90.0], [407.0, 104.0], [361.0, 104.0]], ('3200.00', 0.99577397108078)]
[[[31.0, 108.0], [121.0, 107.0], [121.0, 121.0], [32.0, 122.0]], ('Check Valve 1/2', 0.9563330411911011)]
[[[221.0, 108.0], [249.0, 108.0], [249.0, 121.0], [221.0, 121.0]], ('8481', 0.9984059929847717)]
[[[454.0, 106.0], [483.0, 103.0], [484.0, 117.0], [455.0, 120.0]], ('0.00', 0.9875277280807495)]
[[[500.0, 104.0], [545.0, 104.0], [545.0, 118.0], [500.0, 118.0]], ('1850.00', 0.9945277571678162)]
[[[565.0, 103.0], [613.0, 103.0], [613.0, 117.0], [565.0, 117.0]], ('3700.00', 0.9917264580726624)]
[[[7.0, 126.0], [13.0, 126.0], [13.0, 135.0], [7.0, 135.0]], ('3', 0.6382533311843872)]
[[[281.0, 120.0], [315.0, 120.0], [315.0, 135.0], [281.0, 135.0]], ('2NO', 0.9792035222053528)]
[[[363.0, 121.0], [406.0, 121.0], [406.0, 132.0], [363.0, 132.0]], ('1850.00', 0.9924020767211914)]
 [[[505.0, 192.0], [547.0, 192.0], [547.0, 206.0], [505.0, 206.0]], ('Freight', 0.9902230501174927)]
 [[[586.0, 192.0], [612.0, 192.0], [612.0, 204.0], [586.0, 204.0]], ('0.00', 0.9942625761032104)]
 [[[485.0, 214.0], [545.0, 214.0], [545.0, 225.0], [485.0, 225.0]], ('Basic value', 0.963576078414917)]
 [[[561.0, 212.0], [612.0, 212.0], [612.0, 225.0], [561.0, 225.0]], ('88100.00', 0.9946956634521484)]
 [[[302.0, 228.0], [338.0, 228.0], [338.0, 243.0], [302.0, 243.0]], ('CGST', 0.9944390058517456)]
 [[[372.0, 228.0], [407.0, 228.0], [407.0, 242.0], [372.0, 242.0]], ('9.00%', 0.9570962190628052)]
 [[[568.0, 226.0], [612.0, 226.0], [612.0, 240.0], [568.0, 240.0]], ('7929.00', 0.9968467950820923)]
 [[[302.0, 243.0], [338.0, 243.0], [338.0, 258.0], [302.0, 258.0]], ('SGST', 0.9952542781829834)]
 [[[372.0, 243.0], [408.0, 243.0], [408.0, 257.0], [372.0, 257.0]], ('9.00%', 0.9980270266532898)]
 [[[568.0, 240.0], [613.0, 240.0], [613.0, 254.0], [568.0, 254.0]], ('7929.00', 0.9950212240219116)]
 [[[307.0, 258.0], [339.0, 258.0], [339.0, 272.0], [307.0, 272.0]], ('IGST', 0.9959328174591064)]
 [[[372.0, 257.0], [408.0, 257.0], [408.0, 271.0], [372.0, 271.0]], ('0.00%', 0.9962207078933716)]
 [[[585.0, 254.0], [613.0, 254.0], [613.0, 269.0], [585.0, 269.0]], ('0.00', 0.9526457190513611)]
 [[[397.0, 271.0], [485.0, 271.0], [485.0, 284.0], [397.0, 284.0]], ('GRAND TOTAL', 0.9727221131324768)]
 [[[550.0, 271.0], [612.0, 271.0], [612.0, 282.0], [550.0, 282.0]], ('103958.00', 0.9967718124389648)]
```

4.3. Table Reconstruction

The task of reconstructing tables from extracted text boxes is essential in automating data entry and document analysis. The methodology employed for table reconstruction consisted of the following steps:

4.3.1. Boundary Extension:

The initial step involved extending the boundaries of the text boxes both row-wise and column-wise. This extension helped group related text boxes together, forming the basis for table reconstruction.

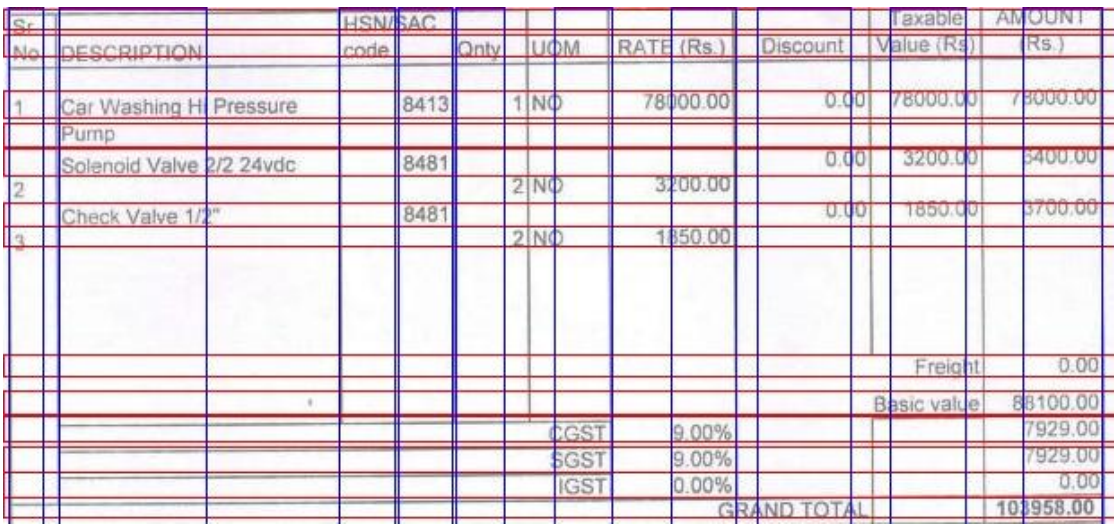


4.3.2. Overlapping Boxes:

Extending the boundaries often resulted in overlapping text boxes, making the reconstruction process more complex. The challenge was to identify the most probable text box for each cell in the table.

4.3.3. Probability Comparison:

To address the issue of overlapping boxes, a comparison of probabilities was performed. The probabilities associated with each text box were evaluated, and the box with the highest probability was selected for each cell.



4.4. Convert into CSV

The reconstructed tables in textual format were converted into Comma-Separated Values (CSV) files, preserving cell content placement. Subsequently, these CSV files were loaded into Pandas Dataframes, facilitating efficient data handling. Using the Pandas Dataframe, an Excel file was generated, organizing data into separate sheets for each table.

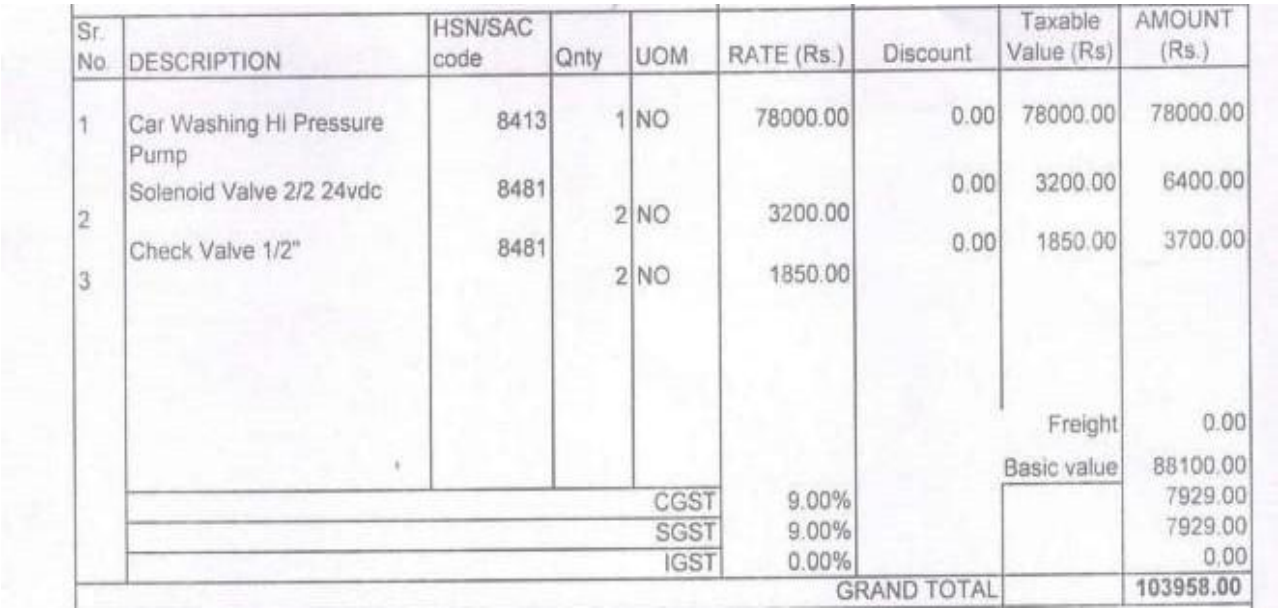
This Excel file provided a structured representation of the tables, enhancing data manipulation tasks.

```
[['Sr.', '', 'HSN/SAC', 'HSN/SAC', '', '', '', '', 'Taxable', 'AMOUNT'],
 ['No', 'DESCRIPTION', 'code', '', 'Qnty', 'UOM', 'RATERs.', 'Discount', 'Value Rs', 'Rs.)'],
 ['', 'Car Washing Hi Pressure','', '8413','','1NO','78000.00','0.00','78000.00', '78000.00'],
 ['', 'Pump', '', '', '', '', '', '', '', '6400.00'],
 ['', 'Solenoid Valve 2/224vdc', '', '8481', '', '', '', '0.00', '3200.00', '6400.00'],
 ['', 'Check Valve 1/2', '', '8481', '', '', '', '0.00', '1850.00', '3700.00'],
 ['3', '', '', '', '', '2NO', '1850.00', '', '', ''],
 ['', '', '', '', '', '', '', '', 'Freight', '0.00'],
 ['', '', '', '', '', '', '', '', 'Basic value', '88100.00'],
 ['', '', '', '', '', 'CGST', '9.00%', '', '', '7929.00'],
 ['', '', '', '', '', 'SGST', '9.00%', '', '', '7929.00'],
 ['', '', '', '', '', 'IGST', '0.00%', '', '', '0.00'],
 ['', '', '', '', '', '', '', 'GRAND TOTAL', '', '103958.00']]
```

	0	1	2	3	4	5	6	7	8	9
0	Sr.		HSN/SAC	HSN/SAC					Taxable	AMOUNT
1	No	DESCRIPTION	code		Qnty	UOM	RATERs.	Discount	Value Rs	Rs.)
2		Car Washing Hi Pressure		8413		1NO	78000.00	0.00	78000.00	78000.00
3		Pump								6400.00
4		Solenoid Valve 2/224vdc		8481				0.00	3200.00	6400.00
5		Check Valve 1/2		8481				0.00	1850.00	3700.00
6	3					2NO	1850.00			
7									Freight	0.00
8									Basic value	88100.00
9						CGST	9.00%			7929.00
10						SGST	9.00%			7929.00
11						IGST	0.00%			0.00
12								GRAND TOTAL		103958.00

5. Demonstration:

Input 1 :



Sr. No	DESCRIPTION	HSN/SAC code	Qnty	UOM	RATE (Rs.)	Discount	Taxable Value (Rs)	AMOUNT (Rs.)
1	Car Washing Hi Pressure Pump	8413	1	NO	78000.00	0.00	78000.00	78000.00
2	Solenoid Valve 2/2 24vdc	8481	2	NO	3200.00	0.00	3200.00	6400.00
3	Check Valve 1/2"	8481	2	NO	1850.00	0.00	1850.00	3700.00
							Freight	0.00
							Basic value	88100.00
				CGST	9.00%			7929.00
				SGST	9.00%			7929.00
				IGST	0.00%			0.00
						GRAND TOTAL		103958.00

Outputs:

Text Detection and Recognition

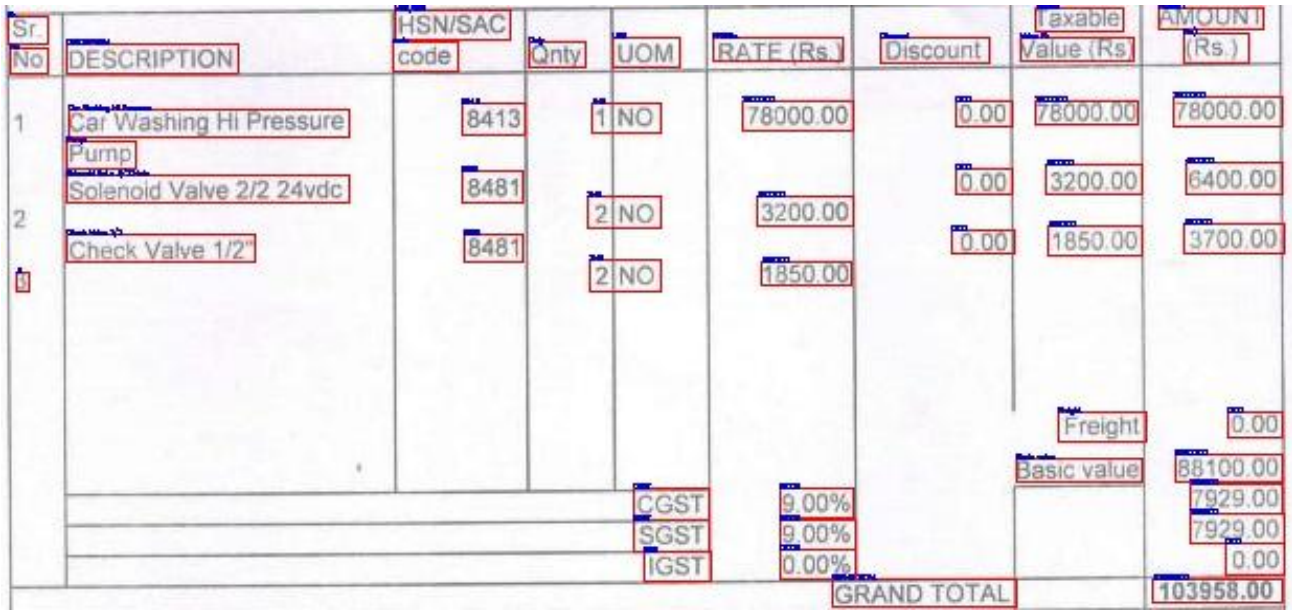
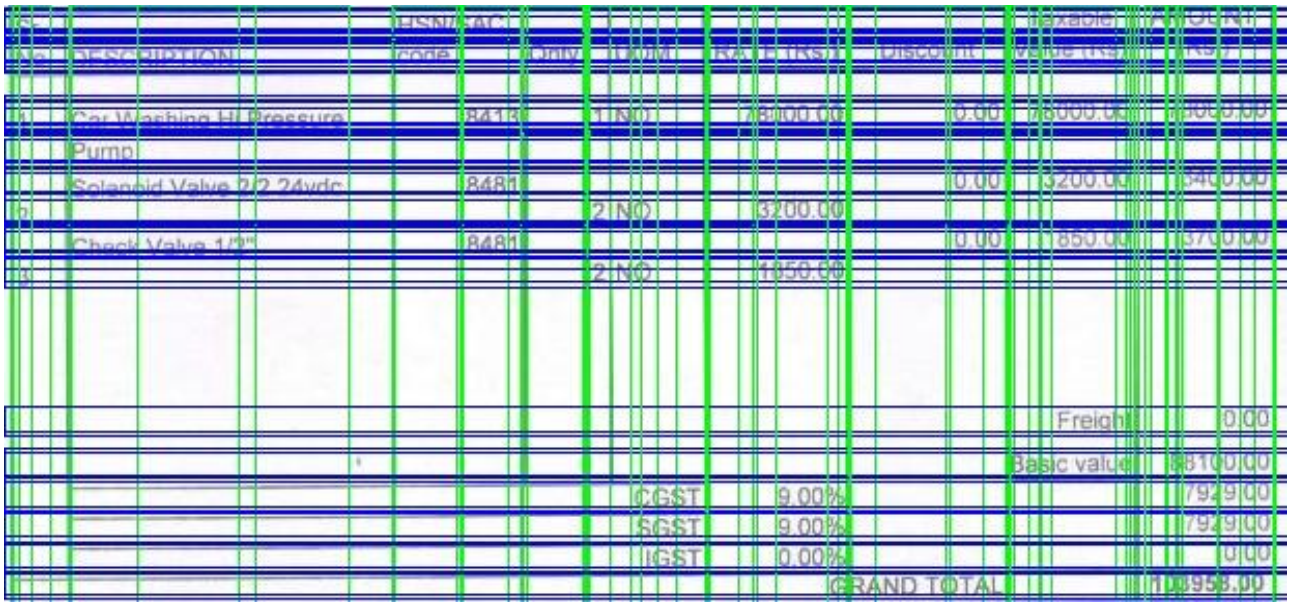
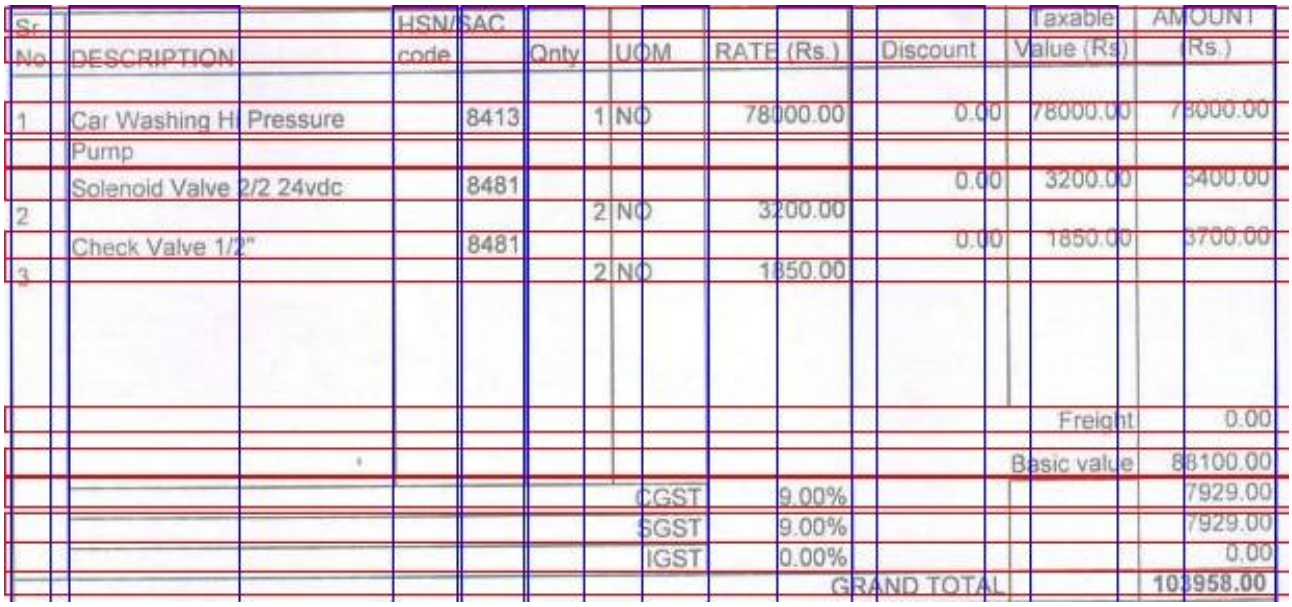


Table Reconstruction

Boundary Extension



Coordinates and Probability Comparision



Sr No	DESCRIPTION	HSN/SAC code	Qnty	UOM	RATE (Rs.)	Discount	Taxable Value (Rs)	AMOUNT (Rs.)
1	Car Washing Hi Pressure Pump	8413	1	NO	78000.00	0.00	78000.00	78000.00
	Solenoid Valve 2/2 24vdc	8481				0.00	3200.00	6400.00
2			2	NO	3200.00			
	Check Valve 1/2"	8481				0.00	1850.00	3700.00
3			2	NO	1850.00			
							Freight	0.00
							Basic value	88100.00
					CGST	9.00%		7929.00
					SGST	9.00%		7929.00
					IGST	0.00%		0.00
						GRAND TOTAL		103958.00

Load into Pandas Dataframe

	0	1	2	3	4	5	6	7	8	9
0	Sr.		HSN/SAC	HSN/SAC					Taxable	AMOUNT
1	No	DESCRIPTION	code		Qnty	UOM	RATERs.	Discount	Value Rs	Rs.)
2		Car Washing Hi Pressure		8413		1NO	78000.00	0.00	78000.00	78000.00
3		Pump								6400.00
4		Solenoid Valve 2/224vdc		8481				0.00	3200.00	6400.00
5		Check Valve 1/2		8481				0.00	1850.00	3700.00
6	3					2NO	1850.00			
7									Freight	0.00
8									Basic value	88100.00
9						CGST	9.00%			7929.00
10						SGST	9.00%			7929.00
11						IGST	0.00%			0.00
12								GRAND TOTAL		103958.00

Convert into Excel spreadsheet

	A	B	C	D	E	F	G	H	I	J
1	0	1	2	3	4	5	6	7	8	9
2	No	DESCRIPTION	code		Qnty	UOM	RATERs.	Discount	Value Rs	Rs.)
3		Car Washing Hi Pressure		8413	1	1NO	78000.00	0.00	78000.00	78000.00
4		Pump								
5		Solenoid Valve 2/224vdc		8481	2			0.00	3200.00	6400.00
6		Check Valve 1/2		8481	2			0.00	1850.00	3700.00
7						CGST	9.00%			7929.00
8						SGST	9.00%			7929.00
9						IGST	0.00%			0.00

Input 2:

Invoice No	Particulars Of Goods (said to contain)	No Of Pkg	Remarks
7002866312, 7002866313, 7002866314, 7002866315, 7002866316, 7002866317	P-UP	6	N3J97663, N3J98052, N3J98100, N3J98101, N3J98103, N3J98104
Declared Value (INR): 5,620,584		Delivery Instruction:	

Outputs:

Text Detection and Recognition

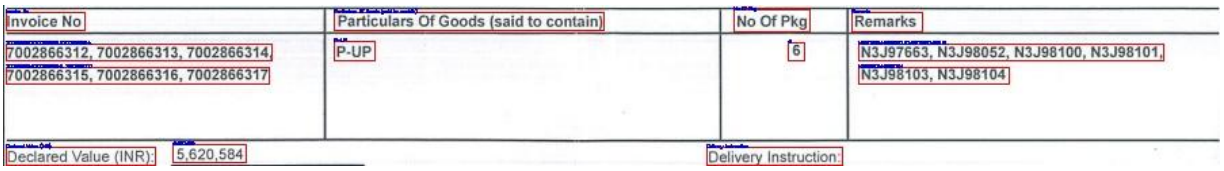
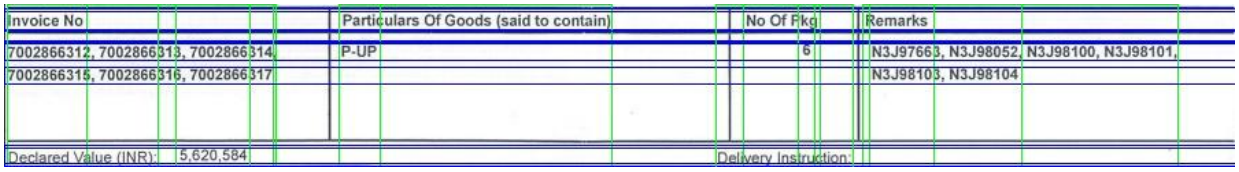
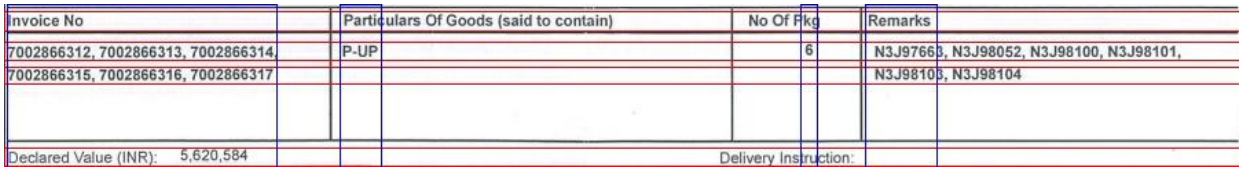


Table Reconstruction

Boundary Extension



Coordinates and Probability Comparision



Load into Pandas Dataframe

	0	1	2	3
0	Invoice No	Particulars Of Goods (said to contain)	No Of Pkg	Remarks
1	7002866312,7002866313,7002866314,	P-UP	6	N3J97663,N3J98052,N3J98100N3J98101,
2	7002866315,7002866316, 7002866317			N3J98103,N3J98104
3	5,620,584	Delivery Instruction:		

6. Overcoming Issues:

Whenever the image is not clear enough, the model struggles to identify tables within the image accurately. In such cases, it becomes necessary to enhance the image's contrast while preserving its quality. Alternatively, cropping the input image to isolate the table before feeding it into the algorithm can also be a viable approach. This highlights the need for robust preprocessing techniques to handle image quality variations effectively.

7. Source Code:

https://colab.research.google.com/drive/1rDsWdBWgHH_sBB9hPOVbEGxYrKWmcsS_?usp=sharing

8. References:

- **PaddlePaddle:** https://www.paddlepaddle.org.cn/
  An open-source deep learning platform derived from industry practice committed to making the innovation and application of deep learning technology easier
- **PaddlePaddle / PaddleOCR**: release/2.7 and release/2.5 documentation
- **neuralearn:** https://neuralearn.podia.com/
  Deep Learning for Computer Vision