# Project Documentation: Automated Tabular Text Extraction

#### 1. Introduction

This project aims to automate the extraction of tabular data from images, converting it into a structured format like CSV or a DataFrame. The challenge is to accurately detect and extract text from cells within a table, even when the table structure is complex or when the text may be slightly misaligned. Various OCR (Optical Character Recognition) models were tested, including open-source solutions and cloud-based services, to identify the most effective method for this task. AWS Textract was ultimately selected due to its superior accuracy and reliability in extracting tabular data.

#### 2. Problem Statement

Given an image containing a table, the goal is to extract the text from each cell and present it in a structured tabular format. The output should include both a CSV file and a visual representation of the detected table with bounding boxes around each cell.

I.	Signature	Printed Name Printed Name  Man L Conquillo	Street and Number 1234 Anywhere St. 2573 S 48th FanWick	City Kent KEn+	Phone Number 555-555-5555 U12 28/8/80	Email you@somewhere.com Mar. 1 - MI & 40@gmai,	Date m/d/y 07 Angzy
2	25/2	Zachary Sn.th	23420 9LT Ave South	Kent	425 - 789-463	zhawk350Ggmail	OT AUL 24
Ζ.	ng	GUEN PITTS	625W meeter ST	Kent	206715-1153	NP	08-9-24
4.	4	BRIAN BOILERA	21920 KENT N	KENT	206 231 7022		8/8/24
5.	Ry Ky	Ruy Bony	23240 88th puts	Kent	(206) 3846169	NA	8/8/24
6.	Michael L.	Michael Felix	11305 SE fent Kangly Rd	Kent	258 753-5579	MA	8-8-24
7.	SurJah	Sunie Jalil	13317SC 236 TOP1	Kent	258-314-1495	allsquie@Ichado	
8.	Rit d	Rita Lourence	1615 W Smith St,	9 Kent	253-317-7805	Ritalawrence 35 Egn	
9.	Juny Johnson	Terry Johnson	23330 51st Ave S	Keut	106-145-3240		08-08-24
10.	11.00 2-	Roy Bousley	24212 63 Ways	Kent	775787 118	8 Roy Bourley	11957 = 50

# 3. Technology and Tools

• Programming Language: Python

#### • Libraries:

- Boto3 for AWS Textract integration
- OpenCV-python for image processing
- o PIL (Python Imaging Library) for image handling
- o Pandas for DataFrame creation and CSV operations
- NumPy for numerical operations

#### • OCR Models Tested:

- Pytesseract: Open-source OCR tool with good accuracy on clean, highcontrast text.
- o **Google Vision**: A cloud-based OCR service with advanced capabilities for text detection.
- EasyOCR: A lightweight OCR tool, particularly effective on images with clear text.
- AWS Textract: Cloud-based service by Amazon specifically designed for document analysis, particularly strong in extracting data from tables and forms.

# 4. Implementation Details

# 4.1 Preprocessing

Before passing the image to AWS Textract, basic preprocessing was performed to enhance text visibility and line detection. This involved converting the image to grayscale, applying adaptive thresholding, and using morphological operations to highlight horizontal and vertical lines.

# **4.2 Text Extraction Using AWS Textract**

The core of the project utilizes AWS Textract, which analyses the image to detect blocks of text, including tables. The service returns a detailed JSON response containing the detected elements and their bounding boxes.

# 4.3 Post-processing

After text extraction, the following steps were carried out:

- Table Structure Reconstruction: Based on the row and column indices provided by AWS Textract, the extracted text was organized into a matrix representing the table structure.
- **DataFrame and CSV Creation**: The matrix was converted into a pandas DataFrame and saved as a CSV file for easy data manipulation and export.

#### 4.4 Visualization

Bounding boxes were drawn around detected cells in the original image to visualize the accuracy of table detection.

# 5. Comparative Analysis

During the development process, several OCR models were tested to evaluate their effectiveness for this specific task:

• **Pytesseract**: Struggled with tables where text alignment was slightly off or when lines were not perfectly horizontal or vertical. Required significant preprocessing for acceptable results.

	1.	110	Mand Bonguillo	2357263 ALT 484h Fentrica	ent.	412 ZP1 8180		DI Ang 1
	2	250	Zachary ph	23420 alst Ave south	Kent	425 789-463	Thank 350 Conail	O7 AUL 24
•	3.	M	GUEN PITS	625W Meeter 57	Kewt Kent	206715-1153	NP	08-9-24
	4.	4	BRIAN BOURAS	21920 KENT N	KENT	206 237 7027		8/8/24
	5.	Ry Ky	Ruy Bony	23240 88th puts	Kent	(206) 3846169 253-5574	NA	8/8/24
	6.	Michael LA.	Michael Felix	11305 SE Lent Kangly Rd	Kent	253 753-5579	MA	8-8-24
	7.	Ser Jah	Squie Jalil	13317 SC 234" PI	Kent	253-314-1495	allsquie@Ichade	a w 1
	8.	Richard	Rita Laurence	1615 W Smith Bt 180	M Kent	253-317-7805	Ritalawrence 35 Rga	
	9.	Jeny Johnson	Terry Johnson	23330 5100 Ave, S	Keut	-206-245-3240		08-08-24
	10.	An Bargo	Roy Bousley	24212 63 Way	Kent	775787 112	8 Roy Bowley	1957 esona
	-/	"//- ()					53847	

• Google Vision: Performed well in text detection but was less consistent in maintaining table structure. Better suited for extracting text blocks rather than structured tables.

Α	В	С	D	E	F	G
Signature	Printed Name	Street and Numb	City	Phone Number	Email	Date
M	Manie Boquiño	1234 Anywhere	Kent	555-555-5555	you@somewher	m/d/y
?	Manie Boquiño	15735 49th Fenr	Kent	412 281 8180	[email address r	07 Aug 24
?	Zachary Smith	23420 9UT Ave	Kent	425-789-4663	[email address r	07 Aug 24
?	Gwen Pitts	625W MeekET S	Kent	206 715-1153	NA	08-09-24
?	Brian Bokhan	21920 Kent	Kent	206 235 7022	NA	8/8/24
?	Ray Pony	23240 88th PUE	Kent	(206) 3846169	NA	8/8/24
?	Mikel Michael Fe	11305 SE kent k	Kent	23753-5579	NA	8-8-24
?	Ball Segnie	13317SP. 236 hp	Kent	253-3141495	allsynie@lcadia	4.8.24
?	Rita Laurence	1615 W smith st	Kent	25-3-317-7805	Rita [email addre	8-524
?	Terry Johnson	23330 51st Ave	Kent	206-245-3240	NA	08-08-24
?	Rousey	24212 63 ways I	Kent	775-2857 1188	NA	NA
?	Roy Bousley	1957 - sha	Kent	725-125-1255	[email address r	NA

- **EasyOCR**: Efficient for simple table structures but struggled with complex or dense tables.
- **AWS Textract**: Outperformed other models in accurately detecting table structures and extracting text, even from challenging layouts.

Signature	- Printh-Inth-	Show the rest to be the	City	O hely compression	diam'i	Date
1 regrature	PERSONAL PROPERTY OF THE PERSONAL PROPERTY OF	STONE TO STONE TO STONE	<b>A</b>	555 AT 50 552 P. 1	somewhere com	m/s i Anazaz
1. // / 4	Man's Ponguille	25735 HOTH Frankick	Kn†	12 281 8180	Man e Mla acagno	A SAME ID DI
2. 35	Zachary Sp. th	23420 abiliave Couth	Ken	425 - 789 - 463	zhawi 350 200	
3. M	GWen Ellis	625W Meeker 57	KeNI KeNI	2010715-1152	NH	04,5-2
4. Our	BRIAN BULLARA	21920 CSW 1	KEP#	206 231 7026		\$ 16 15
5 Ry Fy	Ruy Poor	23240 8844 MES	Kent	(206) 3846/69	·MA	8/8/24
6 Marked LA	Michael Felix	17305 FE Leat Kandy R.	na yn Ko	268 753-5579	MA	8-8-24
7 2 2 7	Crane Toll	72178C 2244Pl	Kent.	253-314-1495	Ollsquie@Ichad	4.6.24
8	Total Language	WIS W Smith St Well	Keni	253-317-7805	Rita lawren cc 35 Pgn	11601
9 1 A harton	Frair Johnson	12421 20 at Aug 5	KINANDO	SAR STARS	MA	08-08-24
10 7 10	Roy Bousley	24212 63 Ways	Kent	775787 112	8 Roy Bourley	1957 = 5019

	Signature	Printed Name Printed Name	Street and Number 1234 Anywhere St.	City Kent	Phone Number 555-555-555	Email you@somewhere.co	Date m/d/y
1	no	Marie Ranquilo	25735 48th Fennick	KEnt	612 281 8180	Mare	07. Angzy
	use	Zachary Smith	23420 abst Ave South	Kent	425-789-4663	zhawk 35a Zon mail	07 AUL2L
2. 3.		Gwen PITTS	625W meeket ST	KeNT	206 715-1153	NA	08-09-2024
4	M	BRIAN	21920 KENT a	KENT	206 235 7022		08-08-2024
5	Ry K	Ray Parky	23240 88th ANES	Kent	(206) 3846167	NA	08-08-2024
6		Michael Felix	11305 SE kent Kangly Rd	Kent	253 753-5579	NA	08-08-2024
7	Sy SaliD	Jalil	1331 SC 236thPl	Kent	253-314-1495		
8	Rit h	Signie Rita Laurence	1615 W Smith st,	Kent	253-317-7805	Ritalawrena135@gmail.com	08-08-2024
9		Terry Johnson	23330 51st Ave S	Keut	206-245-3240	NA	08-08-2024
10	Ag Buy They Johnson	Roy Bousley	24212 63 ways	Kent	725287 1188	Roy	

#### 6. Results

AWS Textract provided the most reliable results, with minimal errors in text extraction and table structure preservation. The output included correctly formatted CSV files and visualizations with bounding boxes around table cells, highlighting the accurate detection of rows and columns.

#### 7. Conclusion

For projects requiring the extraction of tabular data from images, AWS Textract is the recommended tool due to its high accuracy and robustness in handling complex table structures. While open-source alternatives can be effective in simpler scenarios, Textract's specialized capabilities in document analysis make it the superior choice for tasks like these.

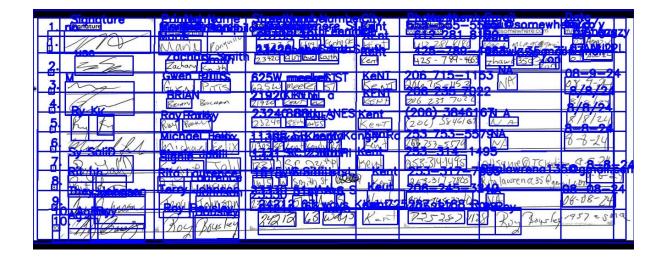
#### 8. Future Work

Potential improvements include further optimizing the preprocessing steps to handle a broader range of table layouts and experimenting with other machine learning models to enhance text detection in specific use cases.

#### 9. Appendix

- Code Snippets:
  - o AWS Textract Integration: Detailed in the provided code.
  - Preprocessing Techniques: Includes thresholding and morphological operations.
  - Comparison of OCR Models: Code for each OCR model tested is available upon request.
- Sample Images and Output Files: Attached or available in the project repository.

**AWS Tesseract:** 



# Threshold Image:

	Signature	Printed Name Printed Name	Street and Number 1234 Anywhere St. 2573 S. 48th Fertylick	City Kent	Phone Number 555-555-5555 (412 28/ 8/80	Email you@somewhere.com Man L Ala Kagan	Date midly 07. Avg2 y
	25	Zachary Snith	23420 glot huc south	Kent	42 <i>5 -</i> 789 - <del>16</del> 3	zhavr3500zmail	07 ALL 24
ž.	2	GUEN PITTS	1625W meeter ST	KENT	206715-1153	NR	04-9-21
4.	4	Beard Buchan	21920 KONT N	KENT	206 231 7022	<u> </u>	8/8/24
<b>5</b> ,	Ry Ky	Ruy Board	23240 88th pits	. Kent	(206) 3846169	WA.	8/8/24
6.4	Milul fl	Michael Felix	11305 SE Lent Kanyly Rd	<u>"Keñt"</u>	253 753-5579	MA	8-3-24
<b>Ž</b> .)	Chetres	Synie Jolil	138178C 236491	Rent	ass-314-1495'	allsquie@Ichalo	
.8.	Wit do	Rita Laurence	1615 W. Smith st, 600	Kent Keut	253-317-7805 -106-145-3240	Ritalawrence35 Pga	08-08-74
9	Suy Johnson	Jerry Johnson	24212 63 ways		7257525711	1/200	
10.	( Ah (Bour)	Roy Bousley	1,001	4	<u>V.</u> N	- Aug - Eug	200 - 200 - A - 1 - 2 - 2 - 2 - 2 - 2 - 2 - 2 - 2 - 2

Noice reduced Image: Poor quality and bad in text detection

Mo	Primal House Primal House WA DWY L. FORN'S	MANUFORE LANGE	Caty Name HEnt	10 22 28 8180	Man & Man A Comme	Down matthy 07 Amp2 y 1
	. Zalan Snith	123 20 girt he with	Kerl	:425:2789-468	zhank35002001	07 84 84
	GUEN PITTS	CASW MORTER ST	Kent	206715-1153	NR	04-9-21
	Book Burne	21920 Ber w	turns !	26 237 7022		5/8/24
Ryty	Pen Breez	2129 10 Phs	Ke-T	(200) 3844/69	WA	8/8/24
Milal L1.	Michael Felix	113 55 fort Roby Rd	1 Kmt	25 757-5579	MA:	8-3-24
2 Ruy M	Cinie Tolil	18317SC 244	Kent	AS8-314-1498'	all sque Otted	₩.4-451
RAT A	Rita Lairunce	1615 Wishits,	Ked!	263-617-7805	Rita la men a so Pia	al Com
They Topico	Jerry Johnson	13330 , 51st Are 5	Vent	-106-145-3240	NA	08-08-24
of Ap Burgo	Ray Boukey	उद्धाव ८३ १४५	X-rt	725750	18 235 Buck	757 - 341

# Vertical and Horizontal Column Detection:

j 87	Signature	Printed Name Printed Name    Variable   Committee	Street and Number 1234 Anywhere St. 2573 S. 48th For	ni Ci	City Kent: モュー	Phone Number 555-555-5555 U12 281 8180	Emoil you@sone	where com	Date midly 07 Angry
2	25 <u>.</u> :	Zachary South	23420 9LT hue south	3	Cent	:42 <i>5</i> - 7 89 - 463	zhavk	35032° ail	07 ALZC
· 3	1 ·	GUEN PITTS	1625W meeter 51	, in		206715-1153	Vlex	-	08-9-21
: 43 :43	a-	Beand Burner	21920 KENT N	,	Kent	206 231 7022		Ш	8/8/z4
(5,	Ry Ky	Ruy Roun	23240 88th puts	4	Kent	(206) 3846169	W.	<u>:                                    </u>	8/8/24
õ., ;	Miled LA	Michael Felix	11305 SE Lent Knigly R	j ,	Kent_	25% 753-5579	MA	-	8-3-24
<b>z</b> .:	Suy M	Sunce Tall	1381781. 2364.4		rent	as <u>8.314.149</u> 5	allsgu	e0 rchd	and-921
.8.	Rit L	Rita Laurence	1615 W Smith St,	(Q. 3	Kent	253-317-7805	Ritalaun	ence 35 fign	ned com
9.	Jung Johnson	Terry Johnson	23330 51st Ave	5	Kent-	106-145-3240	MA		18-08-74
10.	An Brugo .	Boy Bousley	24212 63 W	9	K- AT	7-35-28-2 112	γ ' <u>'</u> ',	g Daysle	1793/25419

# 10. References

- AWS Textract Documentation
- OpenCV Documentation
- Pytesseract GitHub Repository
- Google Vision API Documentation
- EasyOCR Documentation