


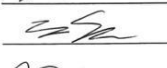


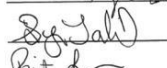
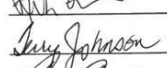
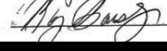



Project Documentation: Automated Tabular Text Extraction

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.

| | Signature | Printed Name Printed Name | Street and Number 1234 Anywhere St. | City Kent | Phone Number 555-555-5555 | Email you@somewhere.com | Date m/d/y |
|-----|---|------------------------------|--|--------------|------------------------------|-----------------------------|---------------|
| 1. |  | Marie Ramonillo | 25735 48th Fernwick | Kent | 612 281 8180 | Marie.M.Ramonillo@gmail.com | 07 Aug 24 |
| 2. |  | Zachary Smith | 23420 91st Ave South | Kent | 425-789-4663 | zhaw350@gmail.com | 07 Aug 24 |
| 3. |  | Gwen Pitts | 625W Meeker St | Kent | 206 715-1153 | NA | 08-9-24 |
| 4. |  | Brian Buchanan | 21920 Kent Ave | Kent | 206 235 7022 | | 8/8/24 |
| 5. |  | Ray Bailey | 23240 88th Ave S | Kent | (206) 5846167 | NA | 8/8/24 |
| 6. |  | Michael Felix | 11305 SE Kent/Kangly Rd | Kent | 253 753-5579 | NA | 8-8-24 |
| 7. |  | Segnie Jalil | 13317 SE 236th Pl | Kent | 253-314-1495 | allsegnie@icloud.com | 8-8-24 |
| 8. |  | Rita Lawrence | 1615 W Smith St, Kent | Kent | 253-317-7805 | Ritalawrence35@gmail.com | 8-8-24 |
| 9. |  | Terry Johnson | 25330 51st Ave S | Kent | 206-245-3240 | NA | 08-08-24 |
| 10. |  | Roy Bousley | 24212 68 ways | Kent | 253-252-1138 | Roy Bousley 757-2512 | |

3. Technology and Tools

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

- **OCR Models Tested:**

- **Pytesseract:** Open-source OCR tool with good accuracy on clean, high-contrast text.
- **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. | | Manie Boquillo | 23420 ALT 49th Fernick | Kent | 412 281 8180 | Manie Boquillo@gmail.com | 07 Aug 24 |
| 2. | | Zachary Smith | 23420 9UT Ave Kent | Kent | 425-789-4663 | zhaw135@gmail.com | 07 Aug 24 |
| 3. | | Gwen Pitts | 625W MeekET S Kent | Kent | 206 715-1153 | NA | 08-9-24 |
| 4. | | Brian Bokhan | 21920 Kent Kent | Kent | 206 235 7022 | NA | 8/8/24 |
| 5. | | Ray Pony | 23240 88th PUE Kent | Kent | (206) 3846169 | NA | 8/8/24 |
| 6. | | Michael Felix | 11305 SE kent k Kent | Kent | 23753-5579 | NA | 8-8-24 |
| 7. | | Segnie Jalil | 13317SP. 236 h Kent | Kent | 253-3141495 | allsynie@icadial.com | 4.8.24 |
| 8. | | Rita Laurence | 1615 W smith st Kent | Kent | 25-3-317-7805 | Rita Lawrence@gmail.com | 8-8-24 |
| 9. | | Terry Johnson | 23330 51st Ave Kent | Kent | 206-245-3240 | NA | 08-08-24 |
| 10. | | Roy Bousley | 24212 63 ways Kent | Kent | 775-2857 1188 | NA | 08-08-24 |

- **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.

| A | B | C | D | E | F | G |
|-----------|------------------|-----------------|------|---------------|-------------------|-----------|
| Signature | Printed Name | Street and Numt | City | Phone Number | Email | Date |
| M | Manie Boquillo | 1234 Anywhere | Kent | 555-555-5555 | you@somewhere | m/d/y |
| ? | Manie Boquillo | 15735 49th Fern | Kent | 412 281 8180 | [email address n | 07 Aug 24 |
| ? | Zachary Smith | 23420 9UT Ave | Kent | 425-789-4663 | [email address n | 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 h | Kent | 253-3141495 | allsynie@icadial | 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 | Kent | 775-2857 1188 | NA | NA |
| ? | Roy Bousley | 1957 - sha | Kent | 725-125-1255 | [email address n | 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.

| 1. | Signature | Printed Name | Street and Numt | City | Phone Number | Email | Date |
|-----|-----------|----------------|------------------------|------|---------------|--------------------------|-----------|
| 1. | | Manie Boquillo | 23420 ALT 49th Fernick | Kent | 412 281 8180 | Manie Boquillo@gmail.com | 07 Aug 24 |
| 2. | | Zachary Smith | 23420 9UT Ave Kent | Kent | 425-789-4663 | zhaw135@gmail.com | 07 Aug 24 |
| 3. | | Gwen Pitts | 625W MeekET S Kent | Kent | 206 715-1153 | NA | 08-9-24 |
| 4. | | Brian Bokhan | 21920 Kent Kent | Kent | 206 235 7022 | NA | 8/8/24 |
| 5. | | Ray Pony | 23240 88th PUE Kent | Kent | (206) 3846169 | NA | 8/8/24 |
| 6. | | Michael Felix | 11305 SE kent k Kent | Kent | 23753-5579 | NA | 8-8-24 |
| 7. | | Segnie Jalil | 13317SP. 236 h Kent | Kent | 253-3141495 | allsynie@icadial.com | 4.8.24 |
| 8. | | Rita Laurence | 1615 W smith st Kent | Kent | 25-3-317-7805 | Rita Lawrence@gmail.com | 8-8-24 |
| 9. | | Terry Johnson | 23330 51st Ave Kent | Kent | 206-245-3240 | NA | 08-08-24 |
| 10. | | Roy Bousley | 24212 63 ways Kent | Kent | 775-2857 1188 | NA | 08-08-24 |

| | Signature | Printed Name Printed Name | Street and Number 1234 Anywhere St. | City Kent | Phone Number 555-555-5555 | 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 | N A | 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.





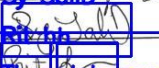
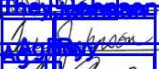
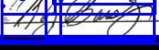

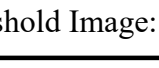
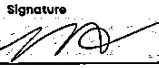
8. Future Work

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.

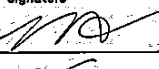
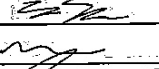

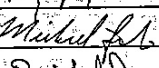
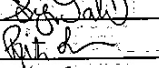


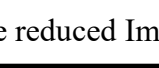
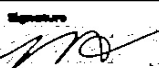
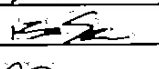
9. Appendix

- **Code Snippets:**
 - **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.

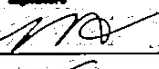
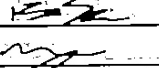

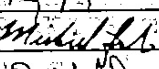
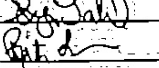
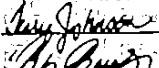
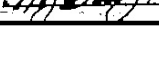



AWS Tesseract:

| | Signature | Printed Name Printed Name | Street and Number 1234 Anywhere St. | City Kent | Phone Number 555-555-5555 | Email you@somewhere.com | Date m/d/y |
|----|---|------------------------------|--|--------------|------------------------------|----------------------------|---------------|
| 1 |  | Marie Kontino | 2573 S 49th Ferndale | Kent | 612 281 8180 | marie.kontino@gmail.com | 07 Aug 21 |
| 2 |  | Zachary Smith | 23420 91st Ave South | Kent | 425-799-4663 | zhaw35@gmail.com | 07 Aug 21 |
| 3 |  | Gwen Pitts | 625W Market ST | Kent | 206 715-1153 | NA | 08-9-24 |
| 4 |  | Brian Buchanan | 21920 Kent Ave | Kent | 206 235 7022 | NA | 8/8/24 |
| 5 |  | Roy Bousley | 23249 58th Ave S | Kent | (206) 3846167 | NA | 8/8/24 |
| 6 |  | Michael Felix | 11385 Kentway Rd | Kent | 253 253-5579 | NA | 8-8-24 |
| 7 |  | Signe Jalil | 13317 SP 236th Pl | Kent | 858 314 4495 | allsigne@icloud.com | 8-8-24 |
| 8 |  | Rita Lawrence | 1615 W Smith St | Kent | 206 317-7805 | Ritalawrence35@gmail.com | 8-8-24 |
| 9 |  | Terry Johnson | 23330 51st Ave S | Kent | 106-245-3240 | NA | 08-08-24 |
| 10 |  | Roy Bousley | 23212 63 Way S | Kent | 225-252-1138 | Roy Bousley 757-0819 | 08-08-24 |






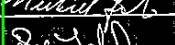




Threshold Image:

| | Signature | Printed Name Printed Name | Street and Number 1234 Anywhere St. | City Kent | Phone Number 555-555-5555 | Email you@somewhere.com | Date m/d/y |
|----|---|------------------------------|--|--------------|------------------------------|----------------------------|---------------|
| 1 |  | Marie Kontino | 2573 S 49th Ferndale | Kent | 612 281 8180 | marie.kontino@gmail.com | 07 Aug 21 |
| 2 |  | Zachary Smith | 23420 91st Ave South | Kent | 425-799-4663 | zhaw35@gmail.com | 07 Aug 21 |
| 3 |  | Gwen Pitts | 625W Market ST | Kent | 206 715-1153 | NA | 08-9-24 |
| 4 |  | Brian Buchanan | 21920 Kent Ave | Kent | 206 235 7022 | NA | 8/8/24 |
| 5 |  | Roy Bousley | 23249 58th Ave S | Kent | (206) 3846167 | NA | 8/8/24 |
| 6 |  | Michael Felix | 11385 Kentway Rd | Kent | 253 253-5579 | NA | 8-8-24 |
| 7 |  | Signe Jalil | 13317 SP 236th Pl | Kent | 858 314 4495 | allsigne@icloud.com | 8-8-24 |
| 8 |  | Rita Lawrence | 1615 W Smith St | Kent | 206 317-7805 | Ritalawrence35@gmail.com | 8-8-24 |
| 9 |  | Terry Johnson | 23330 51st Ave S | Kent | 106-245-3240 | NA | 08-08-24 |
| 10 |  | Roy Bousley | 23212 63 Way S | Kent | 225-252-1138 | Roy Bousley 757-0819 | 08-08-24 |

Noice reduced Image: Poor quality and bad in text detection

| | Signature | Printed Name Printed Name | Street and Number 1234 Anywhere St. | City Kent | Phone Number 555-555-5555 | Email you@somewhere.com | Date m/d/y |
|----|---|------------------------------|--|--------------|------------------------------|----------------------------|---------------|
| 1 |  | Marie Kontino | 2573 S 49th Ferndale | Kent | 612 281 8180 | marie.kontino@gmail.com | 07 Aug 21 |
| 2 |  | Zachary Smith | 23420 91st Ave South | Kent | 425-799-4663 | zhaw35@gmail.com | 07 Aug 21 |
| 3 |  | Gwen Pitts | 625W Market ST | Kent | 206 715-1153 | NA | 08-9-24 |
| 4 |  | Brian Buchanan | 21920 Kent Ave | Kent | 206 235 7022 | NA | 8/8/24 |
| 5 |  | Roy Bousley | 23249 58th Ave S | Kent | (206) 3846167 | NA | 8/8/24 |
| 6 |  | Michael Felix | 11385 Kentway Rd | Kent | 253 253-5579 | NA | 8-8-24 |
| 7 |  | Signe Jalil | 13317 SP 236th Pl | Kent | 858 314 4495 | allsigne@icloud.com | 8-8-24 |
| 8 |  | Rita Lawrence | 1615 W Smith St | Kent | 206 317-7805 | Ritalawrence35@gmail.com | 8-8-24 |
| 9 |  | Terry Johnson | 23330 51st Ave S | Kent | 106-245-3240 | NA | 08-08-24 |
| 10 |  | Roy Bousley | 23212 63 Way S | Kent | 225-252-1138 | Roy Bousley 757-0819 | 08-08-24 |

Vertical and Horizontal Column Detection:

| | Signature | Printed Name Printed Name | Street and Number 1234 Anywhere St. | City Kent | Phone Number 555-555-5555 | Email you@somewhere.com | Date m/d/y |
|-----|---|------------------------------|--|--------------|------------------------------|----------------------------|---------------|
| 1. |  | Marie Rousley | 2573 S 4th Fernick | Kent | 612 281 8180 | Marie.rousley@gmail.com | 07 Aug 24 |
| 2. |  | Zachary Smith | 23420 9th Ave South | Kent | 425-789-4663 | zhawr35@gmail.com | 07 Aug 24 |
| 3. |  | Gwen Pitts | 625W Market St | Kent | 206 715-1153 | N/A | 08-9-24 |
| 4. |  | Brian Bowman | 21920 Kent Ave | Kent | 206 235 7022 | | 8/8/24 |
| 5. |  | Rui Rousley | 23240 9th Ave S | Kent | (206) 384-6187 | N/A | 8/8/24 |
| 6. |  | Michael Felix | 1305 SE Kent Ravely Rd | Kent | 253 757-5579 | NA | 8-8-24 |
| 7. |  | Kymie Jall | 3317 SP 236th | Kent | 253-344-495 | kallsgumie@icloud.com | 8-8-24 |
| 8. |  | Rita Lawrence | 1615 W Smith St, Kent | Kent | 253-317-7805 | Ritalawrence35@gmail.com | 8-8-24 |
| 9. |  | Terry Johnson | 23230 51st Ave S | Kent | 206-245-3240 | NA | 08-08-24 |
| 10. |  | Rui Rousley | 24212 68 Way | Kent | 253-352-1138 | Rui Rousley | 8-8-24 |

10. References

- [AWS Textract Documentation](#)
- [OpenCV Documentation](#)
- [Pytesseract GitHub Repository](#)
- [Google Vision API Documentation](#)
- [EasyOCR Documentation](#)