# Table Detection and Text Extraction using YOLO and Tesseract OCR

## Introduction

### Project Overview

This project was undertaken to explore and develop a system for extracting structured information from scanned invoices. By leveraging YOLO (You Only Look Once) and Tesseract OCR (Optical Character Recognition), the aim was to detect tables within the invoices and extract relevant data, structuring it into a user-friendly format.

## Technology and Methodology

### YOLO for Table Detection

YOLO, a real-time object detection system, is utilized to detect tables within scanned invoices. This deep learning algorithm was used for its capabilities to perform both classification and localization simultaneously. In this project, YOLO's specific advantage lies in its ability to detect tables and their boundaries, making it an important part of the data extraction process.

### Image Preprocessing

Once the tables are detected, they undergo a series of preprocessing steps to optimize them for text extraction. These include:

- Resizing: Scaling the image using the Lanczos4 interpolation method to enhance the resolution.
- Grayscaling: Converting the colored image to grayscale, simplifying the image and emphasizing textual content.
- Blurring: Applying a Gaussian Blur to remove noise and smooth the image.
- Thresholding: Utilizing the OTSU thresholding method to binarize the image, isolating the text for OCR.

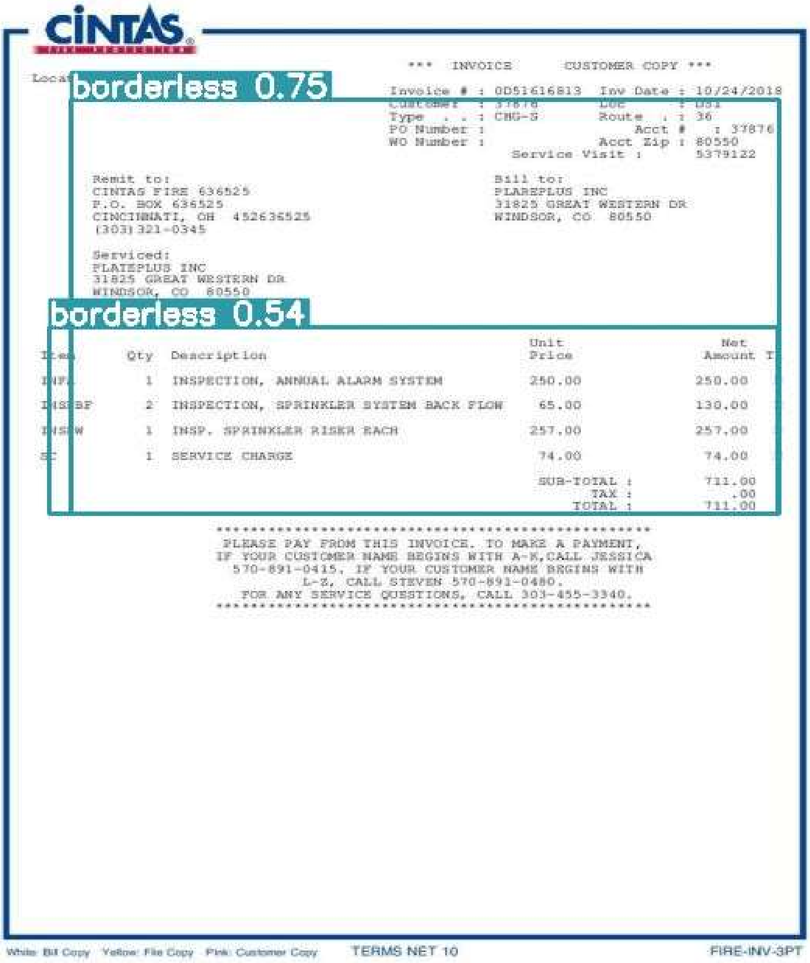
### Text Extraction using Tesseract

Tesseract OCR is an open-source OCR engine that has been trained in various languages and text formats. In this project, it's utilized to extract text from preprocessed images. By employing the OCR Engine Mode 3 (OEM 3), the algorithm considers both the image and any existing language models to enhance the recognition accuracy.

### Data Organization and Export

Post text extraction, the data is processed and organized into categories like Quantity, Description, Unit Price, and Net Amount. This organization is facilitated by using Python's Pandas library to create a DataFrame, which is then exported to a CSV file. The structured format aids in future data analysis and manipulation.

## Visual Representation and Annotation



CINTAS

Locat

*** INVOICE CUSTOMER COPY ***

Invoice # : 0D51616813 Inv Date : 10/24/2018
Customer : 37876 Loc : D51
Type . . : CHG-S Route . : 36
PO Number : Acct # : 37876
WO Number : Acct Zip : 80550
Service Visit : 5379122

Remit to:
CINTAS FIRE 636525
P.O. BOX 636525
CINCINNATI, OH 452636525
(303)321-0345

Bill to:
PLAREPLUS INC
31825 GREAT WESTERN DR
WINDSOR, CO 80550

Serviced:
PLATEPLUS INC
31825 GREAT WESTERN DR
WINDSOR, CO 80550

| Item | Qty | Description | Unit<br>Price | Net<br>Amount | T |
|---|---|---|---|---|---|
| INFA | 1 | INSPECTION, ANNUAL ALARM SYSTEM | 250.00 | 250.00 | |
| INSBF | 2 | INSPECTION, SPRINKLER SYSTEM BACK FLOW | 65.00 | 130.00 | |
| INSRW | 1 | INSP. SPRINKLER RISER EACH | 257.00 | 257.00 | |
| SC | 1 | SERVICE CHARGE | 74.00 | 74.00 | |
| | | | SUB-TOTAL : | 711.00 | |
| | | | TAX : | .00 | |
| | | | TOTAL : | 711.00 | |

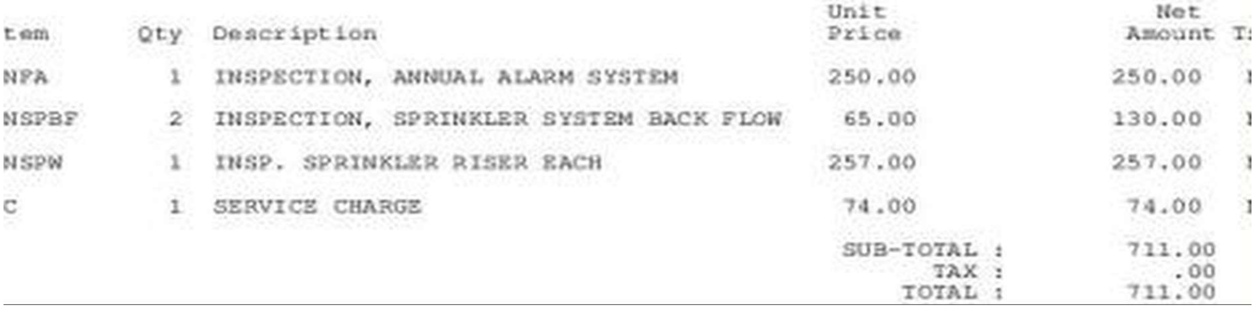
****************************************************
PLEASE PAY FROM THIS INVOICE. TO MAKE A PAYMENT,
IF YOUR CUSTOMER NAME BEGINS WITH A-K,CALL JESSICA
570-891-0415. IF YOUR CUSTOMER NAME BEGINS WITH
L-Z, CALL STEVEN 570-891-0480.
FOR ANY SERVICE QUESTIONS, CALL 303-455-3340.
****************************************************

White: Bill Copy Yellow: File Copy Pink: Customer Copy TERMS NET 10 FIRE-INV-3PT

Figure 1: Table Detection Process

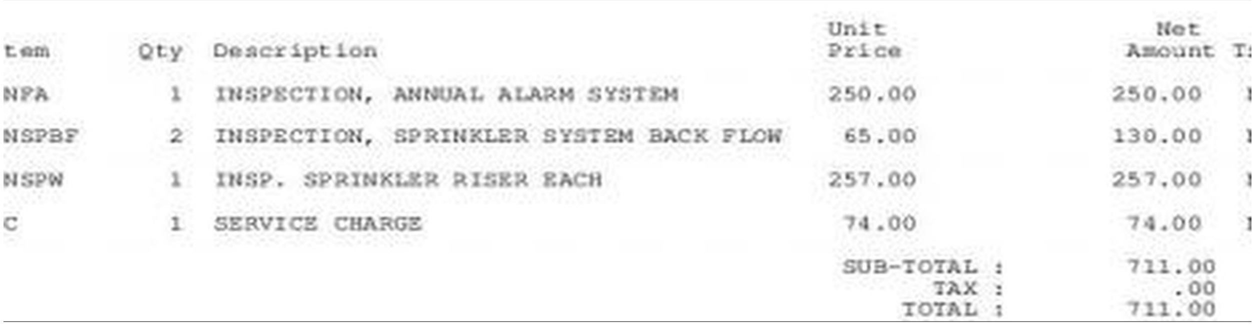
This image represents the initial step of identifying and extracting the table from the source document. The code utilizes boundary detection algorithms to identify the exact coordinates of the table, ensuring a precise crop. This process is important as it isolates the relevant information from any potential noise in the document.

As mentioned earlier, there were several preprocessing steps utilized in the desired region to, eventually, extract the relevant text needed out of the invoice. First, the image is resized to a higher resolution to improve the clarity of text and enable better detection. Next, the colored image is converted to shades of gray, reducing the complexity of the image, and focusing on the essential features. Afterwards, a Gaussian blur was applied to remove noise and smooth the image. Next, thresholding is used to separate the text from the background, creating a binary image where the text is black, and the background is white. The image is then inverted to make the text white and the background black, preparing it for text extraction. Shown below are the figures for the preprocessing steps.



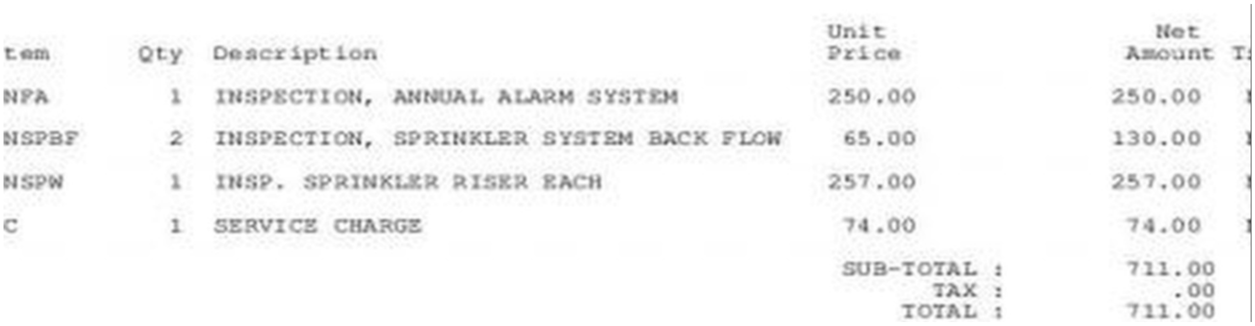
| | | | Unit | Net | |
|---|---|---|---|---|---|
| tem | Qty | Description | Price | Amount | T: |
| NFA | 1 | INSPECTION, ANNUAL ALARM SYSTEM | 250.00 | 250.00 | I |
| NSPBF | 2 | INSPECTION, SPRINKLER SYSTEM BACK FLOW | 65.00 | 130.00 | I |
| NSPW | 1 | INSP. SPRINKLER RISER EACH | 257.00 | 257.00 | I |
| C | 1 | SERVICE CHARGE | 74.00 | 74.00 | I |
| | | | SUB-TOTAL : | 711.00 | |
| | | | TAX : | .00 | |
| | | | TOTAL : | 711.00 | |

Figure 2: Resized Image of Detected Region



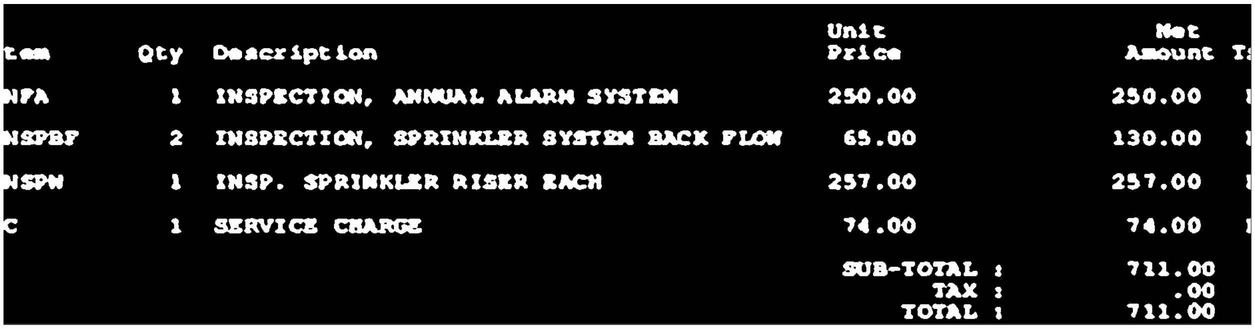
| | | | Unit | Net | |
|---|---|---|---|---|---|
| tem | Qty | Description | Price | Amount | T: |
| NFA | 1 | INSPECTION, ANNUAL ALARM SYSTEM | 250.00 | 250.00 | I |
| NSPBF | 2 | INSPECTION, SPRINKLER SYSTEM BACK FLOW | 65.00 | 130.00 | I |
| NSPW | 1 | INSP. SPRINKLER RISER EACH | 257.00 | 257.00 | I |
| C | 1 | SERVICE CHARGE | 74.00 | 74.00 | I |
| | | | SUB-TOTAL : | 711.00 | |
| | | | TAX : | .00 | |
| | | | TOTAL : | 711.00 | |

Figure 3: Gray scaling Image of Detected Region



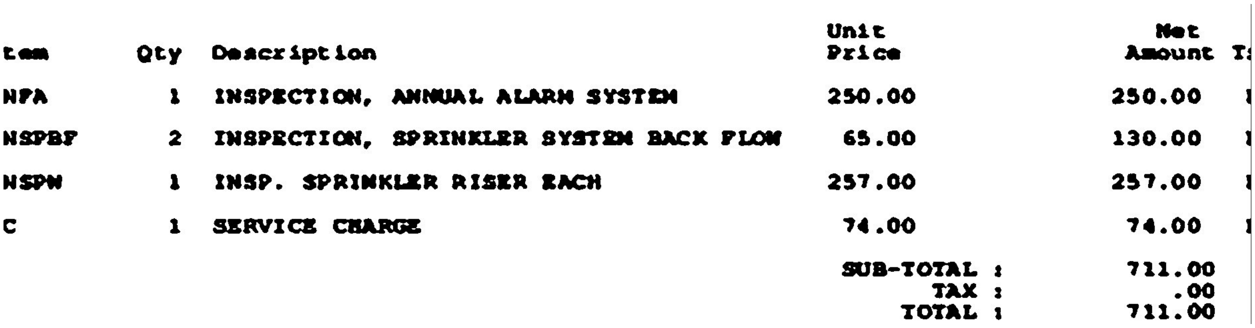
| | | | Unit | Net | |
|---|---|---|---|---|---|
| tem | Qty | Description | Price | Amount | T: |
| NFA | 1 | INSPECTION, ANNUAL ALARM SYSTEM | 250.00 | 250.00 | I |
| NSPBF | 2 | INSPECTION, SPRINKLER SYSTEM BACK FLOW | 65.00 | 130.00 | I |
| NSPW | 1 | INSP. SPRINKLER RISER EACH | 257.00 | 257.00 | I |
| C | 1 | SERVICE CHARGE | 74.00 | 74.00 | I |
| | | | SUB-TOTAL : | 711.00 | |
| | | | TAX : | .00 | |
| | | | TOTAL : | 711.00 | |

Figure 4: Blurring Image of Detected Region



| | | | Unit | Net | |
|---|---|---|---|---|---|
| tem | Qty | Description | Price | Amount | T: |
| NFA | 1 | INSPECTION, ANNUAL ALARM SYSTEM | 250.00 | 250.00 | I |
| NSPBF | 2 | INSPECTION, SPRINKLER SYSTEM BACK FLOW | 65.00 | 130.00 | I |
| NSPW | 1 | INSP. SPRINKLER RISER EACH | 257.00 | 257.00 | I |
| C | 1 | SERVICE CHARGE | 74.00 | 74.00 | I |
| | | | SUB-TOTAL : | 711.00 | |
| | | | TAX : | .00 | |
| | | | TOTAL : | 711.00 | |

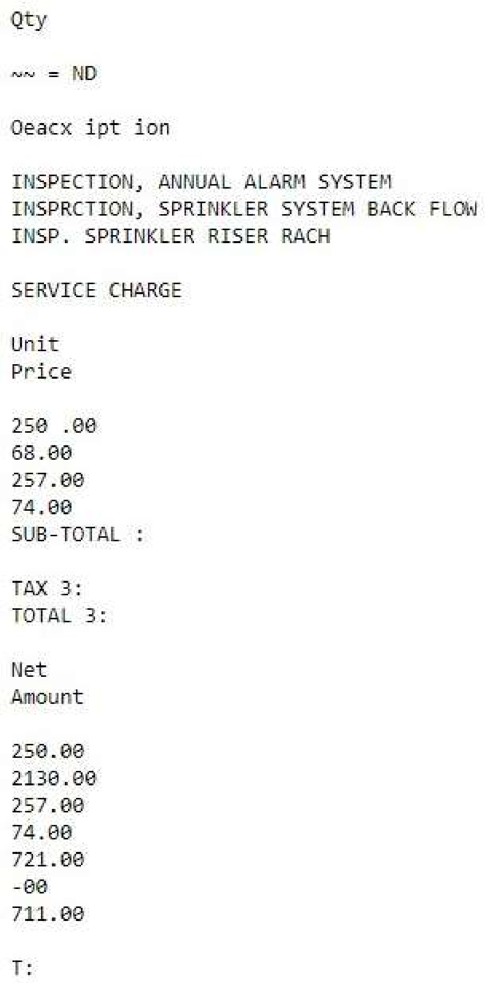
Figure 5: Thresholding of Image of Detected Region



| tem | Qty | Description | Unit<br>Price | Net<br>Amount | T: |
|---|---|---|---|---|---|
| NFA | 1 | INSPECTION, ANNUAL ALARM SYSTEM | 250.00 | 250.00 | |
| NSPBF | 2 | INSPECTION, SPRINKLER SYSTEM BACK FLOW | 65.00 | 130.00 | |
| NSPW | 1 | INSP. SPRINKLER RISER EACH | 257.00 | 257.00 | |
| C | 1 | SERVICE CHARGE | 74.00 | 74.00 | |
| | | | SUB-TOTAL : | 711.00 | |
| | | | TAX : | .00 | |
| | | | TOTAL : | 711.00 | |

Figure 6: Inverting the Image of Detected Region

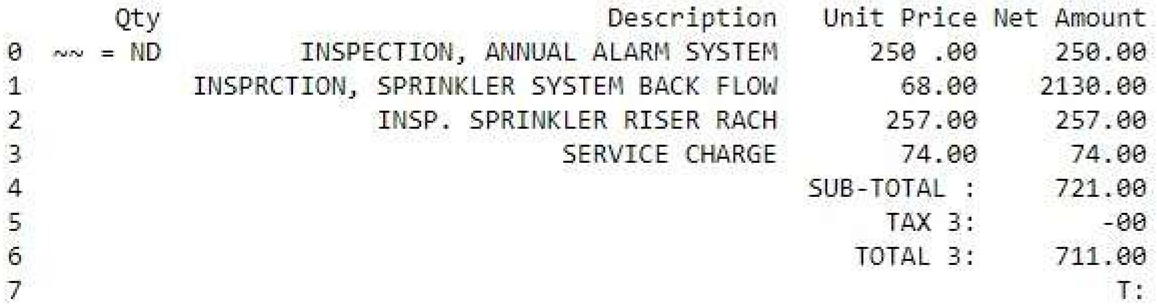
The image shown below is the result of applying Optical Character Recognition (OCR) using Tesseract to the preprocessed table image. Here, characters are recognized and converted into machine-readable text. Despite some potential inaccuracies (e.g., "~~ = ND"), the OCR process has successfully extracted the primary textual content, including quantities, descriptions, unit prices, and net amounts.



```
Qty

~~ = ND

Oeacx ipt ion

INSPECTION, ANNUAL ALARM SYSTEM
INSPRCTION, SPRINKLER SYSTEM BACK FLOW
INSP. SPRINKLER RISER RACH

SERVICE CHARGE

Unit
Price

250 .00
68.00
257.00
74.00
SUB-TOTAL :

TAX 3:
TOTAL 3:

Net
Amount

250.00
2130.00
257.00
74.00
721.00
-00
711.00

T:
```

Figure 7: Extracted Text from Detected Region

The final output is a DataFrame that neatly organizes the extracted information into columns and rows. It shows a successful conversion of the unstructured text data into a structured format. The extracted information was processed, and any inaccuracies were removed for the most part, but there were just a few things that weren’t detected and very few inaccuracies were translated into the final output, as shown below.



```
       Qty                             Description   Unit Price Net Amount
0  ~~ = ND         INSPECTION, ANNUAL ALARM SYSTEM      250 .00     250.00
1           INSPRCTION, SPRINKLER SYSTEM BACK FLOW        68.00    2130.00
2                       INSP. SPRINKLER RISER RACH       257.00     257.00
3                                   SERVICE CHARGE        74.00      74.00
4                                                   SUB-TOTAL :     721.00
5                                                        TAX 3:        -00
6                                                      TOTAL 3:     711.00
7                                                                       T:
```

Figure 8: Final Output DataFrame

# Critical Analysis

## Challenges

- YOLO Installation: The installation process for YOLO was a little challenging and time-consuming. It required careful configuration and compatibility checks, causing delays in the initial development phase.
- Handling Various Invoice Formats: Different invoice structures and layouts posed a significant challenge in developing a generalized solution. Custom adjustments and meticulous tuning were necessary to make the system adaptable.
- Preprocessing Inconsistencies: Certain preprocessing steps were difficult to get right, especially in maintaining the balance between noise reduction and preserving essential details. Trial-and-error and a lot of iteration were required to fine-tune the process.

## Successes

- Accuracy in Results: Despite the complexity of the scanned documents, the solution achieved commendable accuracy in extracting text and numerical information. Key details were captured with minimal errors, showcasing the robustness of the approach.

## Limitations

- Resource Constraints: The inability to train YOLO from scratch due to hardware limitations led to relying on pre-trained models. While effective, this approach may not be as tailored to specific types of invoices, potentially affecting accuracy.
- Sensitivity to Preprocessing Parameters: The solution's performance can vary based on the preprocessing steps' parameters. Setting up the settings in the wrong order can lead to noise retention or essential detail loss, requiring careful fine-tuning for each case.

## Conclusion

This project has demonstrated a successful integration of technologies to extract relevant information from scanned documents and invoices in this case. Utilizing YOLO for object detection and Tesseract OCR for text recognition, these faced several challenges to provide an efficient, accurate, and scalable solution. However, some limitations were identified, such as resource constraints and sensitivity to preprocessing. These areas present opportunities for future enhancement and refinement, such as implementing additional preprocessing strategies or training on specific invoice formats to further improve accuracy.