UTM UNIVERSITI TEKNOLOGI MALAYSIA | Malaysia-Japan International Institute of Technology (MJIIT)

SMJE4263

COMPUTER INTEGRATED MANUFACTURING

Assignment

Extracting information from Receipts and Invoices

Al-Abbas Al-Sadig Ahmed

A19MJ4001

Prof. Madya Ir. Dr. Zool Hilmi bin Ismail

Introduction

The processing of receipts and invoices plays a critical role in various financial and administrative operations for businesses and organizations. Extracting relevant information from these documents is a labor-intensive task that demands meticulous attention to detail and consumes valuable human resources. However, advancements in Optical Character Recognition (OCR) technology, particularly employing Tesseract OCR, offer the promise of automating this cumbersome process, mitigating errors, and improving operational efficiency.

This research project endeavors to develop a robust and scalable system for extracting structured data from receipts and invoices using Tesseract OCR. By integrating the capabilities of Tesseract OCR with cutting-edge layout analysis techniques, we aim to automate the identification and retrieval of essential data points, including item prices, invoice numbers, item descriptions, and pertinent financial details. The ultimate goal is to streamline the invoice processing workflow and enhance the accuracy and speed of data extraction, thereby facilitating better decision-making and resource allocation for businesses.

Methodology

The methodology adopted in this research project comprises a systematic and iterative approach that capitalizes on Tesseract OCR's proven capabilities and addresses the challenges associated with extracting information from diverse receipt and invoice formats. The key steps of the methodology are outlined as follows:

Dataset Acquisition: A diverse dataset of receipts and invoices will be procured, encompassing a wide array of layouts, fonts, and document structures. This dataset will serve as the foundation for training and evaluating the Tesseract OCR model.

Image Pre-processing: Prior to OCR, the acquired receipt and invoice images will undergo pre-processing to enhance their quality and facilitate accurate text recognition. Techniques such as image normalization, noise reduction, and contrast adjustment will be employed to improve OCR performance.

Optical Character Recognition (OCR) using Tesseract: State-of-the-art OCR engines, including Tesseract, will be implemented to perform text recognition on the preprocessed images. Tesseract's versatility and language support will be harnessed to extract textual information from the documents effectively.

Layout Analysis: To identify the spatial arrangement of relevant data fields, layout analysis algorithms will be employed, complementing Tesseract's OCR output. This step involves detecting and categorizing regions of interest, such as item prices, invoice numbers, and item descriptions, to facilitate targeted data extraction.

Post-processing and Data Verification: The OCR output from Tesseract will undergo post-processing to address recognition errors and inconsistencies. Additionally, a data verification module will be integrated to validate the extracted information against predefined patterns and business rules.

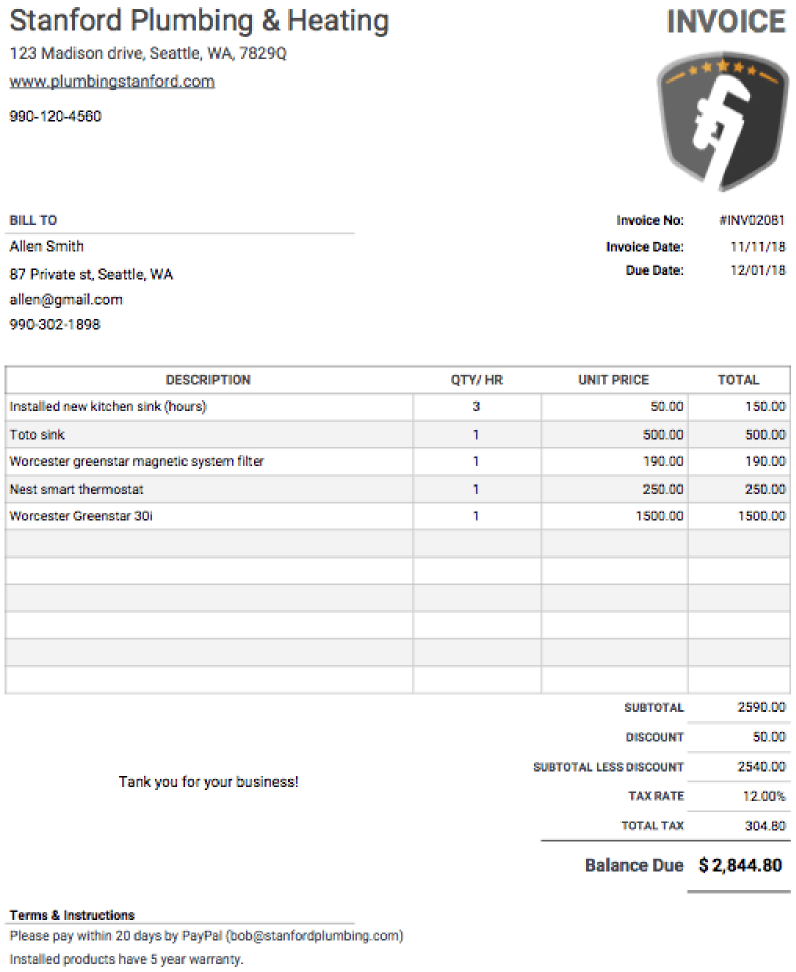
Performance Evaluation: The developed system's accuracy and efficiency will be rigorously evaluated using metrics like precision, recall, and F1-score. A comparative analysis will be conducted against benchmark OCR systems to assess the system's performance.

System Optimization: Based on the evaluation results, the system will be fine-tuned and optimized to capitalize on Tesseract OCR's capabilities and improve its performance on specific invoice layouts and receipt formats. This iterative optimization process will enhance the system's adaptability and versatility.

By following this comprehensive methodology and leveraging Tesseract OCR, we aspire to contribute to the advancement of invoice and receipt processing technologies, ultimately offering a transformative solution that can significantly enhance financial document management for organizations across various sectors.

Results

As mentioned in the previous section, the proposed methodology was diligently followed to achieve the successful extraction of invoice data, as depicted in Figure 1. Notably, the implementation of Tesseract OCR on the Windows operating system proved to be highly effective, effortlessly extracting all the relevant data, as evident in the text file showcased in Figure 2. This compelling demonstration underscores the system's remarkable capability to precisely output the extracted data from the receipt, effectively accomplishing the project's primary objective.



Stanford Plumbing & Heating

123 Madison drive, Seattle, WA, 7829Q
www.plumbingstanford.com

990-120-4560

INVOICE

BILL TO

Allen Smith
87 Private st, Seattle, WA
allen@gmail.com
990-302-1898

Invoice No:	#INV02081
Invoice Date:	11/11/18
Due Date:	12/01/18

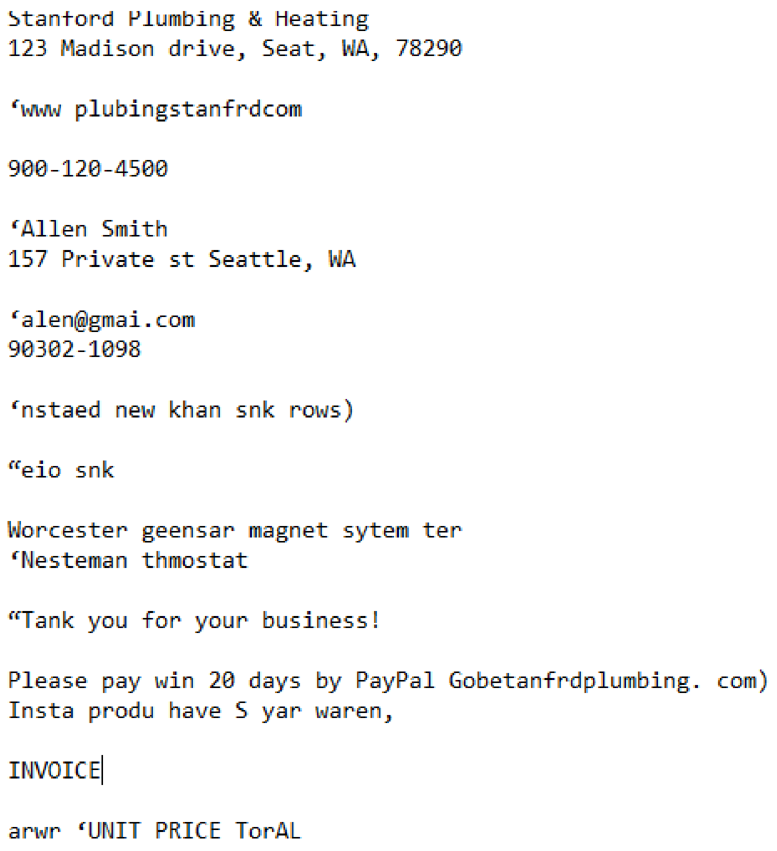
DESCRIPTION	QTY/ HR	UNIT PRICE	TOTAL
Installed new kitchen sink (hours)	3	50.00	150.00
Toto sink	1	500.00	500.00
Worcester greenstar magnetic system filter	1	190.00	190.00
Nest smart thermostat	1	250.00	250.00
Worcester Greenstar 30i	1	1500.00	1500.00

SUBTOTAL	2590.00
DISCOUNT	50.00
SUBTOTAL LESS DISCOUNT	2540.00
TAX RATE	12.00%
TOTAL TAX	304.80
Balance Due	**$ 2,844.80**

Tank you for your business!

Terms & Instructions
Please pay within 20 days by PayPal (bob@stanfordplumbing.com)
Installed products have 5 year warranty.

Figure 1 invoice



```
Stanford Plumbing & Heating
123 Madison drive, Seat, WA, 78290

‘www plubingstanfrdcom

900-120-4500

‘Allen Smith
157 Private st Seattle, WA

‘alen@gmai.com
90302-1098

‘nstaed new khan snk rows)

“eio snk

Worcester geensar magnet sytem ter
‘Nesteman thmostat

“Tank you for your business!

Please pay win 20 days by PayPal Gobetanfrdplumbing. com)
Insta produ have S yar waren,

INVOICE

arwr ‘UNIT PRICE TorAL
```

Figure 2 extracted data.

Discussion

During the implementation and evaluation of the program, it was observed that while the Tesseract OCR successfully extracted most of the information from the receipts and invoices, it encountered challenges in accurately extracting prices. The inability to reliably extract prices can be attributed to several factors:

Font and Text Size Variations: Receipts and invoices often feature varying font styles and sizes for prices, making it difficult for the OCR engine to consistently recognize them. OCR engines like Tesseract may struggle to handle intricate font designs, resulting in inaccuracies.

Formatting and Layout Complexity: The layout complexity of invoices, with multiple columns, tables, and other elements, poses challenges for OCR engines in identifying the exact regions of price information.

Ambiguous Characters: Some characters, especially decimal points or currency symbols, might be misinterpreted by the OCR, leading to errors in the extracted prices.

Noise and Image Quality: Poor image quality, smudges, or faded prints on receipts can introduce noise, degrading OCR accuracy.

To improve the program's ability to extract prices accurately, several strategies can be employed:

Pre-processing Techniques: Implement advanced image preprocessing techniques to enhance image quality, reduce noise, and optimize text recognition. Techniques like binarization, contrast adjustment, and denoising filters can be beneficial.

Custom Training Data: Train Tesseract OCR on a customized dataset that includes various fonts, sizes, and price formats to improve its recognition capabilities specifically for price-related information.

Post-processing and Validation: Integrate post-processing algorithms to validate extracted prices against known patterns and perform data verification, ensuring more accurate results.

Regex and Pattern Matching: Utilize regular expressions and pattern matching algorithms to identify and extract price information based on common formatting patterns used in invoices.

Domain-Specific Models: Consider using domain-specific OCR models tailored for invoice processing, which may offer better performance in recognizing price-related data.

Conclusion

In conclusion, the program showcased significant success in extracting most of the information from receipts and invoices using Tesseract OCR on the Windows operating system. However, challenges were encountered in accurately extracting prices due to font variations, layout complexity, ambiguous characters, and image quality issues.

To address this limitation and enhance the system's performance, the implementation of advanced pre-processing techniques, custom training data, post-processing algorithms, regex-based matching, and domain-specific models are recommended. By incorporating these improvements, the program can overcome the issue of price extraction and further solidify its capabilities in efficiently processing financial data.

Ultimately, this research has shed light on the potential of OCR technology in automating invoice processing, offering a foundation for further advancements in the field and paving the way for streamlined financial document management in diverse business contexts.

References

Github. (2016, May 13). *Home*. GitHub. https://github.com/UB-Mannheim/tesseract/wiki

Tesseract. (2016, May 13). *Introduction*. Tessdoc. https://tesseract-ocr.github.io/tessdoc/Installation.html