Mobile Captured Receipt OCR and Information Extraction

Tran Minh Huy; Nguyen Tien Dat
Advisor: Nguyen Quoc Trung

March 18, 2023

Abstract

Storing the information of receipts or receipts has long been one of the most vital tasks in shops, restaurants, or businesses, etc. For human, extracting information is a tedious task while faced with hundreds or thousands of identical sheets of paper. Furthermore, this is also error-prone, one small mistake can lead to a serious consequence which can cause a lost up to millions dollar. So building a system that does these tasks is worth paying attention to. "Receipts information extraction" (RIE) covers important aspects related to the automated analysis of scanned receipts. The RIE tasks play a key role in many document analysis systems and hold significant commercial potential. This paper proposed an OCR method built and trained on a dataset of over thousands of receipt images and annotations collected from "The Mobile capture receipts Optical Character Recognition" (MC-OCR) challenge. This paper approach the problems with 3 main steps including receipt recognition, text recognition, text extraction.

Keywords: NLP, OCR, RIE

1 Introduction

The widespread use of mobile devices, particularly smartphones, has made it possible for people to carry out various tasks conveniently and efficiently. One such task is the capture and storage of digital receipts, which has become increasingly popular in recent years [1]. The ability to store receipts digitally eliminates the need for paper receipts, reduces clutter, and makes it easier to manage expenses. However, despite the convenience of digital receipts, the process of extracting relevant information from them can be a daunting and time-consuming task.

Extract information from receipts is important for various purposes, including expense tracking, accounting, tax reporting, and more. Manually extracting information from receipts is not only time-consuming but also prone to errors, such as missing data or incorrect entries. Therefore, there is a need for an automated solution that can extract information accurately and efficiently. This is where Optical Character Recognition (OCR) technology comes in [2].

OCR technology involves the use of software to recognize and extract text from images. In the context of receipt extraction, OCR algorithms can be used to extract relevant information such as the date of the transaction, the amount spent, the name of the merchant, and other details [2]. OCR technology has been used in various applications, including digitizing documents, converting scanned images to editable text, and more recently, receipt extraction [3].

2 Related Work

One of the challenges in document analysis is to extract structured information from unstructured or semi-structured documents, such as receipts. Receipts contain key information such as merchant name, date, total amount, and tax that can be useful for various applications such as expense management, accounting, and auditing. However, receipts vary widely in format, quality, and language, making it difficult to apply a single method for information extraction. In this paragraph, we review some of the related work on receipts information extraction in scientific literature.

One approach for receipts information extraction is to use optical character recognition (OCR) techniques to recognize text from scanned images and then apply natural language processing (NLP)

techniques to extract key information from the recognized text. For example, describes a system that uses Azure Form Recognizer OCR-powered receipt data extraction to analyze and extract key information from sales receipts. The system combines powerful OCR capabilities with deep learning models to handle various formats and quality of receipts. The system also supports multiple languages and locales.

Another approach for receipts information extraction is to use graph convolutional networks (GCNs) to model the spatial and semantic relationships between text regions in scanned images [5]. For example, proposes a method that uses GCNs to extract key information from receipts by exploiting the graph structure of text regions. The method first detects text regions using an object detection model and then constructs a graph based on their spatial positions and semantic similarities. The method then applies GCNs to propagate features across the graph nodes and classify them into different categories of key information.

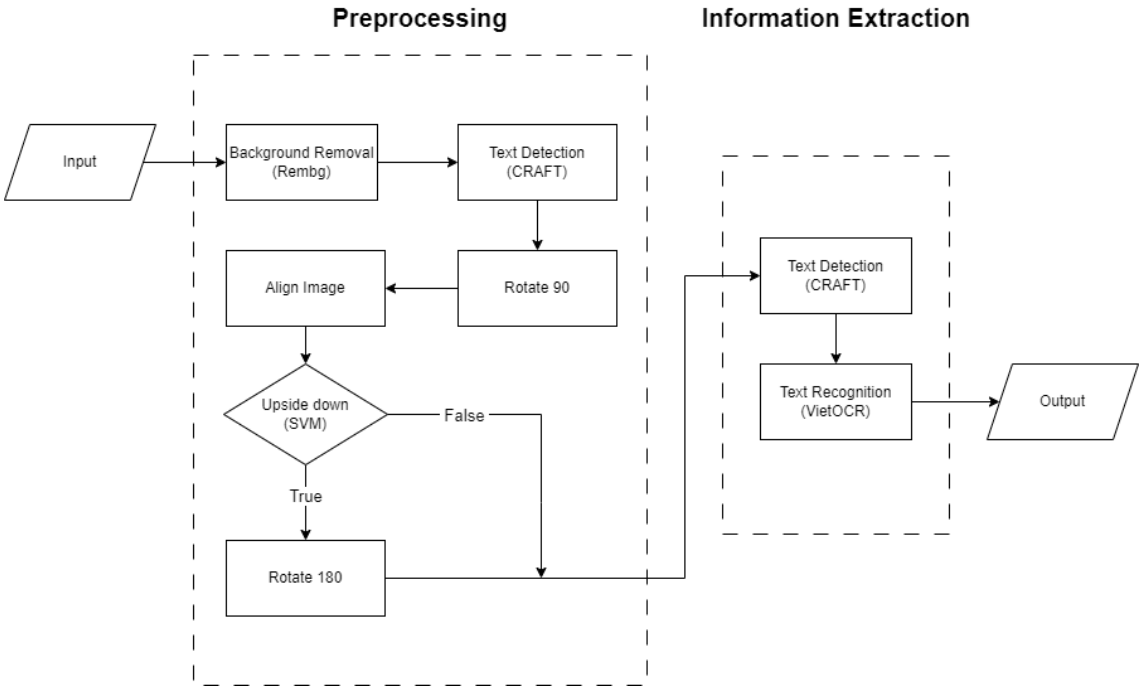
A third approach for receipts information extraction is to use a combination of OCR and GCNs to leverage both textual and visual features of scanned images [4]. For example, presents a survey on scanned receipts OCR and information extraction that compares different methods based on their performance on a benchmark dataset called SROIE. The survey shows that methods that use both OCR and GCNs achieve better results than methods that use only one of them.

3 Data Preparation

The dataset used for this project is obtained from RIVF2021 MC-OCR Competition with over 1100 images of receipts taken under different conditions, in addition we also collected another 450 receipts with total of 1550 receipts. After EDA (Exploratory Data Analysis) we realize that the receipt usually has very large size up to 3000x4000 pixel but the information still not clearly because of the light condition, angle, etc... We also notice that the names of some receipts in the dataset are already labeled with the name of the store, with is very good for us to distinct the different between them and come up with a better solution for later on.

4 System Design

So we created a system to solve this problem. Our pipeline summary is divided into two stages: Preprocessing and Information Extraction.



4.1 Preprocessing

4.1.1 Remove Background

After the EDA process, we found that removing the background was necessary because it took up a lot of space in addition to making data loading and processing more difficult. And if the background contains text, it will make the text recognition process worse. We found Rembg and use it for testing. It is a library published on Github that recognizes the main object and removes the background directly from the image. After running, we get a fairly high accuracy of 99%, the receipt photos are completely backgrounded without losing information. We also convert the photos to black and white because receipts are usually simple, which reduces the data without affecting its actual quality.

4.1.2 Align Image

After background removal we have many images with wrong angle or inverted and to solve the problem we use CRAFT to recognize the text and then use boxes labeled from detected text to calculate and find out the receipt with wrong angle, based on the height and width of the box and then rotated 90 degrees. To deal with a tilted receipt, we use a formula to calculate the tilt of the receipt based on the three longest boxes compared to overall edges of the image. However, after rotated 90 degrees and tilted the image, some of them is still inverted. So we choose Support Vector Machine (SVM) to help us identify what is inverted. We doubled the dataset into 50% normal and 50% inverted with total of 3000 images. After testing, the results are up to 96% accuracy but 4% which mean we still loss around 100 images.

4.2 Information Extraction

The receipts are now better after being processed then we choose a model called VietOCR to identify information in Vietnamese invoices. VietOCR is a model that public in Github with purpose of Vietnamese OCR. In VietOCR there is consists of two models vgg-transformer and vgg-seq2seq, the vgg-transformer is much slower compare to the vgg-seq2seq while there is just a bit difference in accuracy, so for the optimal we choose vgg-seq2seq. At this step we use CRAFT model again to recognize the text and then zoom into the boxes coordinates with zero to five pixels padding to give a better extraction for VietOCR model. For short, the VietOCR model to extract the information in the box areas of the CRAFT model detected. After running this last step we get the output with information of the receipt being stored in the output folder.

5 Result

After the process, I have a folder that fully extracts receipt information based on the input folder with an accuracy rate of about 92%. Information is extracted according to a clear standard, the information in the same row will still be together to create accuracy when looking for a certain information. Errors in information extraction must be mentioned, most of which are due to the poor image quality of the invoice, so even through the quality process it is difficult to improve. However, it does not affect the results too much.

6 Discussion

Through solving the problem of extracting information, we have improved compared to previous research with the following advantages. Firstly, it's simpler and faster to remove the background in the image. Secondly, the black background had been removed and we also converted the image to grayscale to reduce the image size and make the image clearer since receipt don't need much color. Thirdly, the image adjustment process is changed when pushing the tilt rotation to before the reversed rotation, this make the reversed rotation become more easy because we decrease the case of normal, tilt, reverse into normal and reverse only. In the process of reducing the size when removing the background, the training model rotation has quite high results. The image after going through the data pre-processing step is brought back to a size just enough for accurate identification. In addition, model simplification and optimization improve the speed of information extraction significantly.

Information is extracted with high accuracy although there are still some errors mainly due to poor image quality. In the end, the results were almost the same or better than the previous study. Although the data are not as good as expected and the study time is limited so the final results may not be as expected, it is still a worthwhile study and has many positive benefits for the community.

7 Conclusion

In this paper, we developed a multi-step system to extract information from receipts. Upon designing this system, the preprocessing step is one of the most important factors in order for the final model to extract the text. Many papers paid almost all attention to the task of recognizing the text and overlooked the other ones. However, after testing on several models, the most vital factor contributed to the overall accuracy of the model come from how elaborate the image is preprocessed. In further study, we will focus more in this section and see whether there is better way to archive the better performance.

8 Appendix

8.1 Project Management Plan

Team members:

- Nguyen Tien Dat
- Tran Minh Huy

In this project, we uses the Agile model to manage our project because it can quickly adapt to changing requirements, so it's very useful when our project is small and we can communication and collaboration among team members and across different roles.

To build our project management plan, we use 6 step

- Step 1: Identify the project
- Step 2: Determine the desired outcome
- Step 3: Delineate each of the project's component tasks
- Step 4: Identify the the member responsible for the tasks
- Step 5: Determine a timeline
- Step 6: Review, revise and reallocate

With each task, we have some member responsible for this task. They will monitoring progress and aggregate the results. However they don't do it alone, all of us will join in all tasks and we will do it together.

Based on what we found out, we have the timeline below. This is the maximum time limit we have, however progress can be much faster, depending on the request of our class.

Timeline:

- 12/02 - 16/02: Synthesize necessary knowledge for the project
- 17/02 - 25/02: Prepare and Analysis the dataset
- 26/02 - 11/03: Build model
- 12/03 - 20/03: Final result

References

[1] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. E-commerce in your inbox: Product recommendations at scale. pages 1809–1818, 2015.

[2] Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. Optical character recognition. 1999.

[3] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin. A survey of ocr applications. *International Journal of Machine Learning and Computing*, 2(3):314, 2012.

[4] Renshen Wang, Yasuhisa Fujii, and Ashok C Popat. Post-ocr paragraph recognition by graph convolutional networks. pages 493–502, 2022.

[5] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.