

**PAYMENT RECEIPTS VALIDATION THROUGH DUPLICATE
ELIMINATION USING OPTICAL CHARACTER
RECOGNITION**

Vinoch Selvarathinam

(189355L)

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

July 2020

**PAYMENT RECEIPTS VALIDATION THROUGH DUPLICATE
ELIMINATION USING OPTICAL CHARACTER
RECOGNITION**

Vinoch Selvarathinam

(189355L)

Dissertation submitted in partial fulfilment of the requirements for the degree

Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

July 2020

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

Name: S.Vinoch

The supervisor/s should certify the thesis/dissertation with the following declaration.

I certify that the declaration above by the candidate is true to the best of my knowledge and that this report is acceptable for evaluation for MSc Thesis.

Signature of the supervisor:

Date:

Name: Dr. Indika Perera

ACKNOWLEDGEMENTS

My sincere appreciation goes to my family for the continuous support and motivation given to make this thesis a success. I also express my heartfelt gratitude to Dr. Indika Perera, my supervisor, for the supervision and advice given throughout to make this research a success. I also thank my parents, brothers for their heartfelt support. Last but not least I also thank my friends who supported me in this whole effort.

ABSTRACT

Optical Character Recognition (OCR) is the method of digital image retrieval of the characters. The idea behind OCR is to obtain an image or pdf format document and extract the characters from that image and present it in an editable format to the user.

This thesis focused on research related to extract the information such as vendor name, category of the receipt (food related, travel related etc.) and amount from the receipt which can be printed and hand written. Further to identifying the mentioned information, expanded the research on identifying the duplicate receipts as well.

Petty Cash is an accessible store of money kept by organizations for expenditure on small items. When an employee wants to reimburse the amount that he/she spent, they need to fill a voucher with the date of the expense, amount, vendor, reason of the expense and attach the supporting documents (receipts) which will consume papers. In this digital world, easily we can automate this process using digital platforms and tools.

Mobile phones are significantly playing major roles in our day-to-day life more than ever and the usage of mobile phones are increasing drastically compare to desktop computers. In order to reduce the carbon foot print, we can take necessary steps to reduce the paper usage. Building a mobile application which can automate the petty cash process which includes OCR capability on receipts would engage the users to use it in their organizations.

This is the first time, OCR on receipts and duplicate identifier is researched and done. There are no researches conducted on this.

Keywords: Image Processing, Optical Character Recognition, Neural Network

TABLE OF CONTENTS

| | |
|---|-----|
| Declaration | i |
| Acknowledgements | II |
| Abstract | III |
| Table Of Contents | IV |
| List Of Figures | VII |
| List Of Abbreviations | IX |
| 1. Introduction | 1 |
| 1.1 Use Of Petty Cash In The Organizations | 1 |
| 1.2 Cash Frauds In Petty Cash | 2 |
| 1.3 Research Problem | 3 |
| 1.4 Main Challenges In Optical Character Recognition | 3 |
| 1.5 Motivation For The Research | 6 |
| 1.6 Application Of Ocr And Duplicate Identifier Process | 6 |
| 1.7 Objective Of The Research | 8 |
| 1.8 Contribution Of The Research | 9 |
| 2. Literature Review | 10 |
| 2.1 Necessity Of Petty Cash In Organizations | 10 |
| 2.2 History Of Optical Character Recognition | 11 |
| 2.3 Growth Of Optical Character Recognition | 12 |
| 2.4 Types Of Optical Character Recognitions System | 14 |
| 2.5 Text Identification And Extraction From Image Using Optical Character Recognition | 15 |
| | iv |

| | |
|---|----|
| 2.5.1 Histogram Based Approach | 15 |
| 2.6 Techniques In Optical Character Recognition | 16 |
| 2.6.1 Pre-Processing Phase | 16 |
| 2.6.2 Segmentation Phase | 17 |
| 2.6.3 Normalization Phase | 19 |
| 2.6.4 Feature Extraction Phase | 19 |
| 2.6.5 Classification Phase | 20 |
| 2.6.6 Postprocessing Phase | 21 |
| 2.7 Ocr Applications | 22 |
| 2.7.1 Handwriting Recognition | 22 |
| 2.7.2 Receipt Imaging | 22 |
| 2.7.4 Legal Industry | 22 |
| 2.7.4 Banking | 23 |
| 2.7.5 Healthcare | 23 |
| 2.7.6 Captcha | 23 |
| 2.8 Optical Character Recognition Reading By Tesseract Open Source Tool | 24 |
| 2.9 Ocr Engine To Extract Food Items From Receipts | 26 |
| 2.10 Text Extraction On Bills And Invoices | 28 |
| 2.11 Summary Table | 31 |
| 2.12 Deep Learning Techniques Compare With Traditional Ocr Methods | 31 |
| 3. Methodology | 33 |
| 3.1 Identifying The Category Of A Receipt | 33 |
| 3.1.1 Convolutional Neural Network | 33 |
| 3.1.2 Microsoft Azure Custom Vision | 34 |

| | | |
|-------|---|----|
| 3.2 | Extract The Amount From A Receipt | 36 |
| 3.2.1 | Recurrent Neural Networks | 36 |
| 3.2.2 | Long Short-Term Networks | 36 |
| 3.2.3 | Microsoft Azure Computer Vision – Cognitive Service | 39 |
| 3.2.4 | Identifying The Amount Using Regular Expression | 40 |
| 3.3 | Identifying The Duplicate Receipts | 42 |
| 3.3.1 | Text Similarity | 42 |
| 3.3.2 | Jaccard Similarity | 42 |
| 3.3.3 | Comparing The Texts Of The Receipts | 43 |
| 3.3.4 | Comparison Between Date And Time, Vendor, Amount And Category | 43 |
| 4. | Solution Architecture And Implementation | 45 |
| 4.1 | Load, Extract, Transform (Etl) | 45 |
| 4.2 | Database Preparation | 51 |
| 4.3 | Mobile App Development | 51 |
| 4.3.1 | Developed Mobile App Interfaces | 52 |
| 5. | Data & Analysis | 55 |
| 5.1 | Custom Vision Training | 55 |
| 5.2 | Identifying The Amount Using Regex | 56 |
| 5.3 | Duplicate Receipts Identification | 58 |
| 6. | General Discussion & Conclusion | 60 |
| 6.1 | General Discussion On The Case Study | 60 |
| 6.2 | Conclusion | 61 |
| 6.3 | Future Work | 62 |
| | References | 63 |

LIST OF FIGURES

| | |
|---|----|
| Figure 2.1: Architecture steps of Tesseract OCR | 24 |
| Figure 2.2: Tesseract Optical Character Recognition Result Analysis | 25 |
| Figure 2.3: Walmart receipts before and after image background removal | 26 |
| Figure 2.4: Final text retrieved from Walmart receipt | 27 |
| Figure 2.5: Example output image after the canny edge detection | 28 |
| Figure 2.6: Line segmentation process | 29 |
| Figure 2.7: Word segmentation process | 30 |
| Figure 2.8: Character segmentation process | 30 |
| Table 2.2: OCR Applications and the accuracy | 31 |
| Figure 3.1: Azure Custom Vision Architecture [39] | 35 |
| Figure 3.2: An unrolled recurrent neural network | 37 |
| Figure 3.3: The repeating module in a standard RNN | 38 |
| Figure 3.4: For interacting layers in LSTM repeating module | 39 |
| Figure 3.5: Sample receipt and the scanned value from Azure computer vision | 40 |
| Figure 3.6: LSTM Architecture | 41 |
| Figure 3.6: Duplicate Receipt Identifier Architecture | 44 |
| Figure 4.1: Tagging of the receipt in Azure custom vision | 46 |
| Figure 4.2: Tagged images set in Azure Custom Vision | 47 |
| Figure 4.3: Checking the performance after training | 48 |
| Figure 4.4: Selection interface of the training type | 49 |
| Figure 4.5: Testing the prediction after the training of the model | 50 |
| Figure 4.6: Sample receipt obtained from Keells | 52 |
| Figure 4.7: Output result after scanning the image | 53 |
| Figure 4.8: Identified duplicate receipts are shown in list | 54 |
| Figure 5.1: The duplicate prediction output by the developed system in percentage | 59 |

LIST OF TABLES

| | |
|--|----|
| Table 2.1: Some important pre-processing operations | 16 |
| Table 2.2: OCR Applications and the accuracy | 31 |
| Table 5.1: Microsoft Custom Vision training output results | 55 |

LIST OF ABBREVIATIONS

| Abbreviation | Description |
|--------------|--|
| OCR | Optical Character Recognition |
| CAGR | Compound Annual Growth Rate |
| CAPTCHA | Completely Automatic Public Turing Test to Tell Computers and Humans Apart |
| ANPR | Automatic Number Plate Recognition |
| GISMO | Geographic Information Systems and Mapping Operations |
| ANSI | American National Standards Institute |
| DIA | Document Image Analysis |
| DPI | Dots Per Inch |
| SVM | Support Vector Machine |
| LSTM | Long Short-Term Memory |
| RNN | Recurrent Neural Network |
| GRU | Gated Recurrent Unit |
| REGEX | Regular Expression |
| CNN | Convolutional Neural Network |
| ETL | Extract, Transform and Load |
| CV | Custom Vision |