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 per-son 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

LIST OF FIGURES

LIST OF TABLES

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