Faculty of Electrical and Electronic Engineering Technology

OPTICAL CHARACTER RECOGNITION (OCR) ON IMAGES USING TEMPLATE-MATCHING AND IMAGE CORRELATION

NURUL HASYA BINTI ROJEHAN

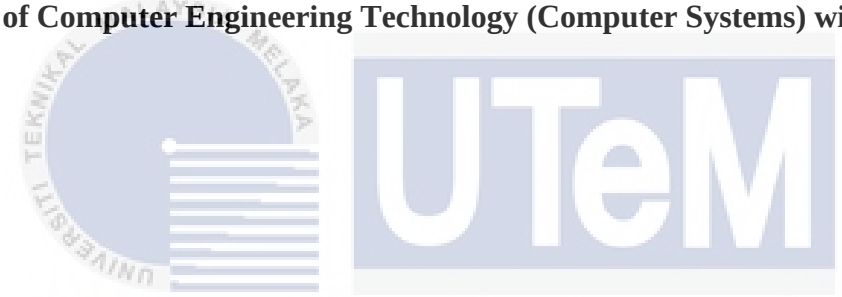
Bachelor of Computer Engineering Technology (Computer Systems) with Honours

2022

OPTICAL CHARACTER RECOGNITION (OCR) ON IMAGES USING TEMPLATE-MATCHING AND IMAGE CORRELATION

NURUL HASYA BINTI ROJEHAN

A project report submitted
in partial fulfillment of the requirements for the degree of
Bachelor of Computer Engineering Technology (Computer Systems) with Honours



Faculty of Electrical and Electronic Engineering Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2022

UNIVERSITI TEKNIKAL MALAYSIA MELAKA
FAKULTI TEKNOLOGI KEJUTERAAN ELEKTRIK DAN ELEKTRONIK

BORANG PENGESAHAN STATUS LAPORAN
PROJEK SARJANA MUDA II

Tajuk Projek: Optical Character Recognition (OCR) on Images using Template-matching and Image Correlation

Sesi Pengajian: 2021/2022

Saya **Nurul Hasya Binti Rojehan** mengaku membenarkan laporan Projek Sarjana Muda ini disimpan di Perpustakaan dengan syarat-syarat kegunaan seperti berikut:

1. Laporan adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan laporan ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (✓):

☐	**SULIT***	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
☐	**TERHAD***	(Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
☑	**TIDAK TERHAD**	

Disahkan oleh:

(TANDATANGAN PENULIS)
Alamat Tetap: No 125,
Jalan Berlian 6,
RPT Jelapang Baru,
30020 Ipoh,
Perak

(COP DAN TANDATANGAN PENYELIA)

Ts. DR. ROSTAM AFFENDI BIN HAMZAH
Dean
Faculty of Electrical & Electronic Engineering Technology
Universiti Teknikal Malaysia Melaka

Tarikh: 7/1/2022

Tarikh: 7/1/2022

*CATATAN: Jika laporan ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali tempoh laporan ini perlu dikelaskan sebagai SULIT atau TERHAD.

DECLARATION

I declare that this project report entitled “Optical Character Recognition on Images using Template-Matching and Image Correlation” is the result of my own research except as cited in the references. The project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : 

Student Name : NURUL HASYA BINTI ROJEHAN

Date : 7/1/2022

APPROVAL

I hereby declare that I have checked this project report and in my opinion, this project report is adequate in terms of scope and quality for the award of the degree of Bachelor of Computer Engineering Technology (Computer Systems) with Honours.

Signature : 

Supervisor Name : TS. DR. ROSTAM AFFENDI BIN HAMZAH

Date : 7/1/2022

DEDICATION

Alhamdulillah, praise to the Almighty Allah SWT.

This thesis is dedicated to:

My mother,

Mrs Rozilah Binti Harun

ABSTRACT

Optical Character Recognition (OCR) has grown in popularity as a result of its widespread use in converting photos into editable machine-coded text in the multimedia and digital fields. Optical Character Recognition (OCR) is a method for extracting content from a digital image by increasing its quality. Pre-processing, Segmentation, Feature Extraction, and Classification are all part of the OCR system. To identify the font and retrieve the text, template matching and correlation were utilised. The goal of this project is to extract text from photos that contain critical information, with the text serving as the output.

ABSTRAK

Sepanjang tahun, Pengesahan Huruf Optik (OCR) telah menjadi satu permintaan yang tinggi disebabkan penggunaanya dalam menukarkan imej kepada huruf komputer yang boleh diubah dalam bidang digital dan multimedia. Teknik Pengesahan Huruf Optik (OCR) adalah proses untuk mendapat teks dengan memperbaiki kualiti imej digital. Sistem OCR terdiri daripada Pre-pemprosesan, Pengekstrakan ciri, Pengelasan dan Pemprosesan Pasca. Untuk pemprosesan pasca, padanan templat dan korelasi digunakan untuk mengenalpasti huruf komputer dan mengekstrak teks tersebut. Tujuan projek ini ialah untuk mendapat teks dari gambar yang mempunyai maklumat yang sangat penting di mana teks tersebut akan keluar melalui buku nota.

ACKNOWLEDGEMENTS

Predominantly, I am inclined to express my gratitude to my supervisor, Ts. Dr Rostam Affendi Bin Hamzah for his precious admonishment, words of insight and tolerance all over this project.

I am also honour-bound to Universiti Teknikal Malaysia Melaka (UTeM) for the monetary support which validates me to achieve the project. Not slipping my colleagues for their enthusiasm for sharing their thoughts and opinions regarding the project.

My highest admiration goes to my parents, and family members for their fondness and prayer during the span of my study. Without their support, I would not have been qualified to complete my final year project.

Finally, I would like to thank everyone who is involved in this project either directly or indirectly, associates and classmates, the faculty members, as well as other solitaries who are not listed here for being cooperative and obliging.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\sum x$ - Total of x

LIST OF ABBREVIATIONS

A	-	The template grey level image
$\bar{A}$	-	The average grey level in the template image
B	-	The source image section
$\bar{B}$	-	The average grey level in the source image

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

The term OCR is an acronym for Optical Character Recognition. It is a technique for detecting text within a digital image. Text recognition in scanned documents and photographs is a typical application. OCR software can turn a physical paper document or an image into a text-searchable electronic form. OCR, which is implemented in MATLAB software, is a widely used programme that is well-known for its many features in the image processing toolkit. OCR is a text recognition technology that uses photos or videos to recognise text in a multimedia document. Many studies and developments have been conducted on OCR to produce a precise text extraction result. Many approaches are used in the development of OCR, including image processing, pre-processing, and other approaches that are introduced and integrated into the OCR system. All of these strategies have their own set of advantages and disadvantages when it comes to text extraction. In addition, the OCR is subjected to a variety of methods and algorithms to obtain the desired output.

This project intends to create an OCR system in MATLAB programme to extract the text using an OCR system. MATLAB was chosen because it is simple to implement, has free software access for students, and is simple to understand the process in OCR. MATLAB stands for MATrixLABoratory, a programme that converts images into matrixes so that subsequent processes can be performed quickly. As a result, the colours, edges, intensity, texture, and pattern in the image can be easily recognised and identified. Pre-processing, segmentation, feature extraction, and classification are the four main components of an OCR

system. The images will be pre-processed, resulting in images that are smoother than the originals. Each character is converted from the segmented word. The feature extraction will then take the character's information and compare it to the existing template for classification.

1.2 Problem Statement

In today's technological world, advanced multimedia technology has become a source of concern for daily living, as it can be confusing for some people. People like to take pictures, but it is risky to maintain them because there could be problems with them. Despite this, an architecture for an OCR system was developed to address the issue. However, to share the image with others, the text information included within the image must be typed manually. The processes inside the OCR system can assist users in quickly sharing information in text format by converting digital photos into extractable text. However, there could be an issue if the text were extracted incorrectly. To avoid this, the OCR system's performance must be evaluated using a variety of photos to ensure that the text extraction result is accurate.

1.3 Project Objective

These are the three goals outlined below:

a) To present a template-matching and picture correlation-based framework for OCR.
b) To match the fonts on an image to the proposed project.
c) To evaluate the performance of an OCR system with various types of fonts found in photos.

1.4 Scope of Project

The approach of extracting text from photos is focused on a specific group of people with specific needs. First, it focuses on individuals who want to avoid keeping information within photos, as well as others who may harm device storage utilisation. Next, this strategy focuses on a wide community, particularly in multimedia and digital, to make the process of exchanging information via social media or a website more efficient.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

For extracting text from photos, optical character recognition is now widely used. Many studies turned out to be conducted to boost and enhance the current system's production. This chapter examines earlier optical character recognition research papers for the past five years.

2.2 Overview

Optical Character Recognition (OCR) has become well-known, and its application in extracting text from photos has enticed individuals to adopt it, particularly those with a passion for multimedia. People are increasingly using smartphones to capture photographs to save the information contained in the image as technology advances.

In general, OCR is a type of data entry method that enters data into a system in alphabetic, numeric, or symbolic form. Because OCR is widely employed in the digital world, demand for OCR has continued to rise till now. People learn the simplest technique to save data by using OCR, which converts images into text, which can then be copied and simply forwarded to others.

2.3 Related Research

2.3.1 Optical Character Recognition Systems

This study described the procedures used by the existing OCR system in detail, as well as the source and remedy to the problem. A typical OCR system is made up of various parts [1]. After the images have been printed, this system will be turned off. Handwritten and printed characters can be accomplished; however, the quality of the scanned image must be considered. Figure 2.1 shows the elements of an OCR system.

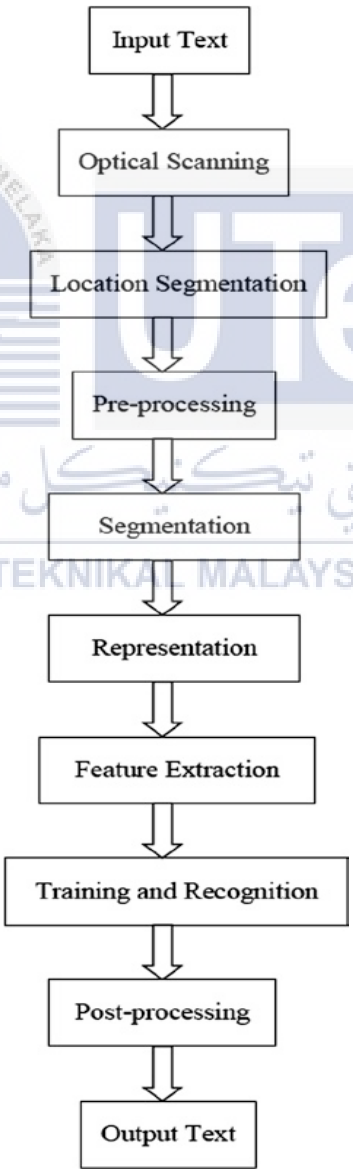


Figure 2.1 Components of OCR System

The OCR system, which is based on MATLAB (MATrixLABoratory), is well-known for transforming images into matrices and allowing numerous processes to be applied to the images to obtain the desired result. The images' colours, intensity, edges, texture, and pattern may all be detected using this software [2]. The scanner is used as a medium to capture images into the computer by using the OCR process. During the segmentation phase, the text region in the image will be recognised, and the symbol will be extracted [3]. To go on to the next stage, the recognised symbol will go through a series of steps that include pre-processing and noise removal. Each symbol was identified by comparing the extracted symbol classes from the preceding procedure.

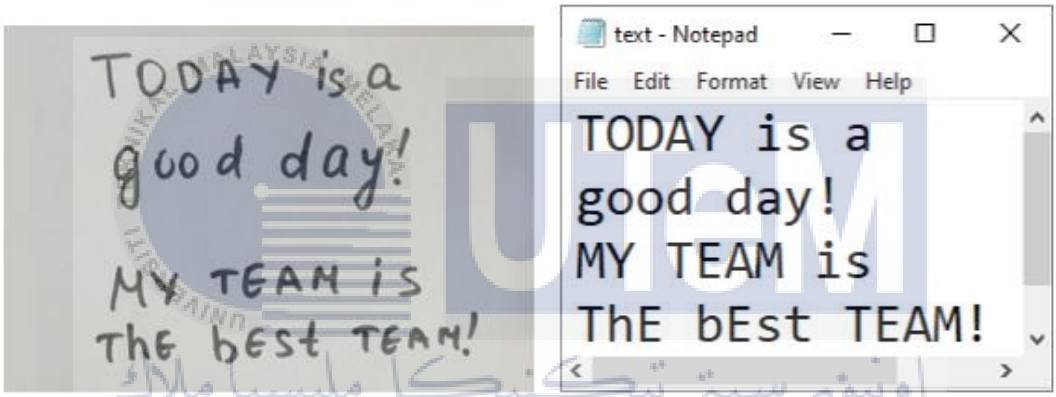


Figure 2.2 Handwritten Sample and Output

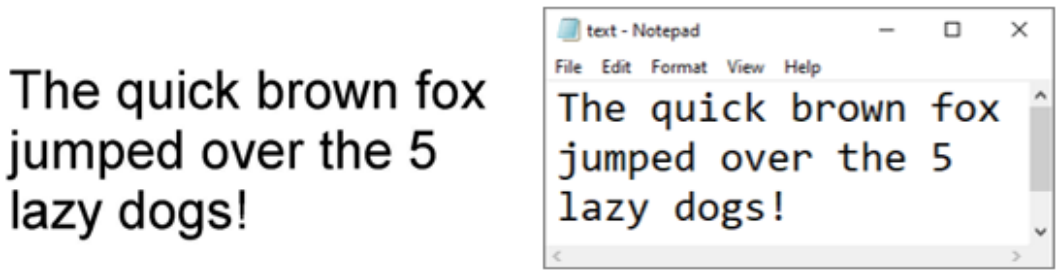


Figure 2.3 Text Image Sample and Output

To evaluate the outcome, the system was reviewed based on three characteristics throughout the process. Character classification is proportionately proportioned in terms of recognition rate. There are some characters that the system does not recognise, which

contributes to the rejection rate. The fail character, on the other hand, is highlighted by the OCR system, making it easier to spot for manual repair. In terms of error rate, the system will not notice a misclassified character, but the error can be detected and corrected manually. Even though the OCR system is cost-effective, the time spent detecting and correcting the OCR system's errors is more essential [4]. There will be only a 1% error indicated in a 99 per cent correct identification rate.

2.3.2 Steps Involved in Text Recognition in OCR

In this study, various steps are reviewed including text recognition, classification of manuscript OCR systems according to text type, and application-oriented recent OCR research. Reducing noise, contrast, enhancement, and image sharpening are examples of low-level image processing operations [5]. Pre-processing is a crucial step before moving on to feature extraction since it ensures that the results are appropriate for the subsequent phases. Binary or grey pictures are used in the majority of OCR applications. Without conducting the pre-processing stage, the photos may have watermarks or a non-uniform backdrop, making recognition harder.

The filters are used to cancel out the image's high or low frequency. Smoothing is the process of removing high frequencies from an image while boosting or edge detection is the process of removing low frequencies [6]. The following Figure 2.4 displays the primary image and 2.5 displays the image that has been applied with Prewitt and Canny edge detection methods.