

B071710802

BACHELOR OF COMPUTER ENG. TECH. (COMPUTER SYSTEMS)

2020 UTeM

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



NUR HAKIMAH BT ZAKARI
اونيورسي تيكنيكل ماليسيا ملاك
B071710802

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2020



UNIVERSITI TEKNIKAL MALAYSIA MELAKA

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

This report is submitted in accordance with the requirement of the Universiti Teknikal Malaysia Melaka (UTeM) for the Bachelor of Computer Engineering Technology (Computer System) with Honours

اونيورسيتي تيكنيكل مليسيا ملاك By

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

NUR HAKIMAH BT ZAKARI

B071710802

981012-11-5288

FACULTY OF ELECTRICAL AND ELECTRONIC ENGINEERING

TECHNOLOGY

2020

BORANG PENGESAHAN STATUS LAPORAN PROJEK SARJANA MUDA

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

Sesi Pengajian: 2020

Saya **Nur Hakimah bt Zakari** mengaku membenarkan Laporan PSM ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka (UTeM) dengan syarat-syarat kegunaan seperti berikut:

1. Laporan PSM adalah hak milik Universiti Teknikal Malaysia Melaka dan penulis.
2. Perpustakaan Universiti Teknikal Malaysia Melaka dibenarkan membuat salinan untuk tujuan pengajian sahaja dengan izin penulis.
3. Perpustakaan dibenarkan membuat salinan laporan PSM ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. **Sila tandakan (X)

- SULIT* Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia sebagaimana yang termaktub dalam AKTA RAHSIA RASMI 1972.
- TERHAD* Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan.
- TIDAK TERHAD

Yang benar,

Disahkan oleh penyelia:


Nur Hakimah bt Zakari



Alamat Tetap:

3653 Kampung Duyong Besar

20100 Kuala Terengganu

Terengganu Darul Iman

Tarikh: 15.02.2021



DR. ROSTAM AFFENDI BIN HAMZAH
Penasarah Kanan
Jabatan Teknologi Kejuruteraan Elektronik & Komputer
Fakulti Teknologi Kejuruteraan Elektrik & Elektronik
Universiti Teknikal Malaysia Melaka

Tarikh: 15.02.2021

*Jika Laporan PSM ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan PSM ini

DECLARATION

I hereby, declared this report entitled OPTICAL CHARACTER RECOGNITION (OCR) ON IMAGES USING TEMPLATE-MATCHING AND IMAGE CORRELATION is the results of my own research except as cited in references.

Signature:



Author : Nur Hakimah bt Zakari

Date: 15.02.2021

APPROVAL

This report is submitted to the Faculty of Electrical and Electronic Engineering Technology of Universiti Teknikal Malaysia Melaka (UTeM) as a partial fulfilment of the requirements for the degree of Bachelor of Computer Engineering Technology (Computer Systems) with Honours. The member of the supervisory is as follow:



ABSTRAK

Sepanjang tahun, Pengesanan 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 Pengesanan 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. Untuk templat ini, 6 jenis huruf komputer dipilih antaranya Cambria, Lucida Sans, Dubai Medium, Microsoft YaHei, Malgun Gothic dan Lato melalui 4 jenis sampel dan ketepatan dianalisis berdasarkan pengesanan yang betul (%). Sistem OCR ini akan diimplikasikan menggunakan perisian pengiraan iaitu MATLAB dengan Image Processing Toolbox. Tujuan projek ini ialah untuk mendapat teks dari gambar yang mempunyai maklumat yang sangat penting di mana teks tersebut akan keluar melalui buku nota.

ABSTRACT

Throughout the year, Optical Character Recognition (OCR) has become a demand as the high usage in images that converted images into editable machine-coded text within multimedia or digital field. Optical Character Recognition (OCR) techniques is a process to extract the text by improving the quality of the digital image. OCR system includes Pre-processing, Segmentation, Feature Extraction and Classification. For classification, template matching and correlation used to identify the font and extract the text. For this template, 6 fonts have been chosen which are Cambria, Lucida Sans, Dubai Medium, Microsoft YaHei, Malgun Gothic and Lato through 4 various sample to match the template and analyses the accuracy based on the correct recognition (%). This OCR system will be implied in MATLAB software with Image Processing Toolbox. The purpose of this project is to obtain the text within the images that contain crucial important information where the text will be output through notepad.

DEDICATION

Alhamdulillah, praise to the Almighty Allah S.W.T

This thesis is dedicated to:

My Parents,

Mr Zakari bin Ngah and Mrs Siti Rohani bt Ismail



ACKNOWLEDGEMENTS

In the Name of Allah, the Most Gracious, the Most Merciful

Alhamdulillah, thank you Allah because of His blessing, I finally complete and finish my final year project successfully.

During the process to complete my project objective, I do a lot of research either by using internet, reading past year thesis and reference books and with the guidance and support from peoples around me, I finally complete the project due to the time given. Here, I want to give credit to those who helped me to achieve what I had achieved in my final year project.

First and foremost, I would like to express my deep sense of gratitude and acknowledgement to my supervisor TS DR Rostam Affendi bin Hamzah for his timely guidance, advices, valuable and constructive suggestions during the planning and developing of this project. In addition, thanks for her support and encouragement throughout this final year project. I would like to show appreciation everyone who is involved in this project either directly or indirectly for their helps and co-operation, and to my family. Without their support, I would not have been able to finish my final year project.

TABLE OF CONTENTS

	PAGE
TABLE OF CONTENTS	xi
LIST OF TABLES	xv
LIST OF FIGURES	xvi
LIST OF APPENDICES	xix
LIST OF SYMBOLS	xx
LIST OF ABBREVIATIONS	xxi
CHAPTER 1 INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	3
1.3 Research Objective	4
1.4 Scope of Research	5
1.5 Thesis Outline	6
CHAPTER 2 LITERATURE REVIEW	7
2.1 Introduction	7
2.2 Overview of optical character recognition	7
2.3 Related Researches	8
2.3.1 Optical Character Recognition using MATLAB	8

2.3.2	A Text Detection, Localization and Segmentation System for OCR in Images	10
2.3.3	Performance of Document Image OCR Systems for Recognizing Video Texts on Embedded Platform	13
2.3.4	Detecting Symbol on Road Surface for Mapping and Localization using OCR	15
2.3.5	OCR Based Image Text to Speech Conversion Using MATLAB	17
2.3.6	Ote-OCR Based Text Recognition Extraction from Video Frames	19
2.3.7	Natural Scene Text Localization and Recognition	20
2.3.8	Text Localization, Enhancement and Binarization in Multimedia Documents	21
2.3.9	“I” – A novel algorithm for Optical Character Recognition (OCR)	23
2.3.10	Optical Character Recognition: An Encompassing Review	27
2.3.11	Character Recognition Using MATLAB’s Neural Network Toolbox	28
2.3.12	Improving OCR performance with Background Image Elimination	29
2.4	Summary	31

CHAPTER 3	METHODOLOGY	34
3.1	Introduction	34
3.2	Workflow to Implementation Project	34
3.3	Flowchart represent process of the project	36
3.4	Block Diagram of Project	37
3.5	Software Implementation	38
3.6	Optical Character Recognition	40
3.6.1	Pre-processing	40
3.6.2	Segmentation	41
3.6.2.1	Line Segmentation	41
3.6.2.2	Word Segmentation	42
3.6.2.3	Character Segmentation	43
3.6.3	Feature Extraction	43
3.6.4	Feature Template-matching and Image Correlation	44
CHAPTER 4	RESULT AND DISCUSSION	46
4.1	Introduction	46
4.2	Software simulation	46
4.2.1	The coding for the process of the system	46
4.3	Result simulation	51
4.3.1	Image Processing	51

4.3.2	Sample 1	53
4.3.3	Sample 2	55
4.3.4	Sample 3	55
4.3.5	Sample 4	57
4.4	Result Analysis	58
4.4.1	An analysis based on accuracy of text extraction	58
4.4.2	Analysis based on sum of accuracy	64
4.5	Summary	66
CHAPTER 5	CONCLUSION	67
5.1	Introduction	67
5.2	Summary	67
5.3	Future Work	68
REFERENCES		69
APPENDIX		73



LIST OF TABLES

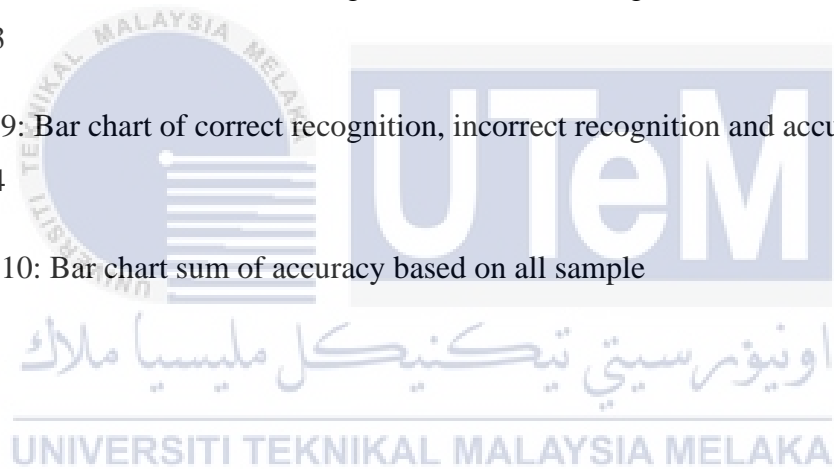
TABLE	TITLE	PAGE
Table 2.1:	Comparison of past related projects	33
Table 4.1:	Image pre-processing	53
Table 4.2:	Text extraction of Sample 1	54
Table 4.3:	Text extraction of Sample 2	55
Table 4.4:	Text extraction of Sample 4	56
Table 4.5:	Text extraction of Sample 4	57
Table 4.6:	The percent of accuracy for correct recognition in text extraction	60
Table 4.7:	Sum of Accuracy	64

LIST OF FIGURES

FIGURE	TITLE	PAGE
Figure 2.1:	Parts of an OCR System	8
Figure 2.2:	Handwritten sample and the output	9
Figure 2.3:	Text image sample and the output	9
Figure 2.4:	The main processing steps of the system	11
Figure 2.5:	The result of for the TDL module	12
Figure 2.6:	Result after (a) localization, (b) binarization and (c) recognition process.	12
Figure 2.7:	Comparison of OCR Software	13
Figure 2.8:	The screen shot output of FineReader	14
Figure 2.9:	Effect of Italic Fonts on Tesseract Accuracy	14
Figure 2.10:	Overview of the proposed algorithm	15
Figure 2.11:	Example of top-view images.	16
Figure 2.12:	Tesseract working process	17
Figure 2.13:	OCR Framework	18
Figure 2.14:	TTS system	18
Figure 2.15:	Edge Detection before and after dilation	19
Figure 2.16:	Edge Detection before and after dilation	19

Figure 2.17: Output for the video sequence	20
Figure 2.18: Text localization and recognition results on SVT dataset	21
Figure 2.19: Morphological operation	22
Figure 2.20: The input image(a), the gradient(b), the binarized image(c) and image after morphological post processing(d)	22
Figure 2.21: Main block diagram	23
Figure 2.22: Segment data format	24
Figure 2.23: Data fields in the segment data format	25
Figure 2.24: Different neighbor pixel patterns	26
Figure 2.25: Original image and processed image	30
Figure 3.1: Planning project flowchart	35
Figure 3.2: Project development flowchart	36
Figure 3.3: Block diagram of project	37
Figure 3.4: MATLAB development environment	39
Figure 3.5: MATLAB operation	39
Figure 3.6: Image after binarization	41
Figure 3.7: Horizontal and Vertical Projection	42
Figure 3.8: Correct character segmentation using vertical projection	43
Figure 3.9: Feature extraction of character h	44
Figure 3.10: Extracted character with the matching template	44
Figure 4.1: The code for image pre-processing in MATLAB	47

Figure 4.2: The code for image horizontal segmentation in MATLAB	48
Figure 4.3: The code for image vertical segmentation in MATLAB	49
Figure 4.4: The code for template creation in MATLAB	49
Figure 4.5: The code for template matching in MATLAB	50
Figure 4.6: Bar chart of correct recognition, incorrect recognition and accuracy in Sample 1	60
Figure 4.7: Bar chart of correct recognition, incorrect recognition and accuracy in Sample 2	61
Figure 4.8: Bar chart of correct recognition, incorrect recognition and accuracy in Sample 3	62
Figure 4.9: Bar chart of correct recognition, incorrect recognition and accuracy in Sample 4	63
Figure 4.10: Bar chart sum of accuracy based on all sample	65



LIST OF APPENDICES

APPENDIX	TITLE	PAGE
Appendix 1:	Coding for Optical Character Recognition	74
Appendix 2:	Coding for horizontal segmentation	75
Appendix 3:	Coding for horizontal segmentation	76
Appendix 4:	Coding for clipping	76
Appendix 5:	Coding for template creation	78
Appendix 6:	The used template	78
Appendix 7:	Coding for template matching	81



LIST OF SYMBOLS

A	-	The template gray level image
\bar{A}	-	The average gray level in the template image
B	-	Source image section
\bar{B}	-	the average gray level in the source image
R	-	Template image size



LIST OF ABBREVIATIONS

OCR	-	Optical Character Recognition
TDL	-	Text detection localization
CSB	-	Character segmentation and binarization
STB	-	Set top box
TIFF	-	Tagged image file format
YUV	-	Luminance-bandwidth-chrominance
Ote	-	Optical Text Extraction
RGB	-	Red-Green-Blue
SVM	-	Support vector machine
SVT	-	Street view text



CHAPTER 1

INTRODUCTION

1.1 Background

Optical character recognition (OCR) is the main aspect that need to be implemented in order to extract the text from the images. OCR implemented in MATLAB software is one of the common usage software and famously knows as for its various features in image processing toolbox. OCR is used to recognize the texts based on multimedia document either in images or videos. Many researches and development have been carried out on OCR in order to obtain the precise result of text extraction. The development of OCR involves many techniques to use like image processing, pre-processing and more of other techniques introduced and implemented to the OCR system. All these techniques have its own pros and cons to the text extraction. Besides, various method and algorithms are applied to the OCR in order to get the result.

The aim of this project is to extract text using OCR system by implementing the system in MATLAB software. MATLAB is chosen because it easy to implement, free accessing the software for student and easy to understand how it work in optical character recognition (OCR). MATLAB stands for MATrixLABoratory which image can be converted into matrixes first and the other techniques can perform easily. Thus, the images can be easily recognized and identified the colours, edges, intensity, texture or pattern in image. OCR system consists of four main components which is pre-processing, segmentation, feature extraction and classification. The images will undergo pre-processing method which the images will be smoother than the original images. The

segmented word converts into each character individually. Then, the feature extraction will gain the information of the character and match it with the existing template for classification.

The result from this project shows that the OCR system able to extract text from the images at the same time upgrade the picture quality.



1.2 Problem Statement

In the technology world, the advanced technology regarding multimedia become a concern to the daily life as it may complicated to some people. People tend to capture images, but it may vulnerable to keep the images regarding there may be issues within it. Despite that, a framework for optical character recognition (OCR) system introduced to solve the problem. However, the text information that contain within the image need to type manually in order to share it to others. By using OCR system, the techniques within the system can help people in order to fast sharing it in text format by changing the digital images into extractable text. Nevertheless, there may be some issue as the text maybe wrongly extracted. In order to avoid that, the performance of OCR system must be analysis using various images to make sure that the text extraction result is accurate.

