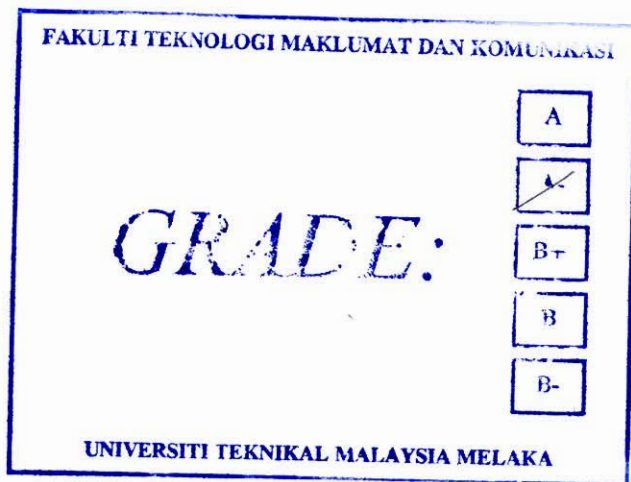


THE OPTICAL CHARACTER RECOGNITION (OCR) SYSTEM WITH TEMPLATE MATCHING

MOHD SHAHRIZAN MOHAMAD IESA



UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

TITEL: **OPTICAL CHARACTER RECOGNITION (OCR)**

TAHAP PENGAJIAN: **2008/2011**

NAMA: **MOHD. SHAHRIZAN BIN MOHD. IESA**

Perpustakaan ingin membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hakmilik Universiti Teknikal Malaysia Melaka
2. Perpustakaan Universiti Teknikal Malaysia Melaka dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Universiti Teknikal Malaysia Melaka dibenarkan membuat salinan tesis 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



(TANDATANGAN PENULIS)



(TANDATANGAN PENYELIA)

Alamat tetap: NO 79, KAMPUNG
JAK, JALAN BATU KAWA, 93050
MIRI, SARAWAK.

Penyelia: DR. GEDE PRAMUDYA ANANTA

Tarikh: 14/07/2011

Tarikh: 14/07/2011

PERHATIAN: *Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)

** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

**OPTICAL CHARACTER RECOGNITION (OCR)
WITH TEMPLATE MATCHING**

MOHD SHAHRIZAN BIN MOHAMAD IESA

**This report is submitted in partial fulfillment of the requirements for the
Bachelor of Computer Science (Artificial Intelligence)**

**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA**

2011



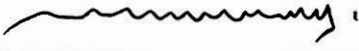
DECLARATION

I hereby declare that this project entitled

THE OPTICAL CHARACTER RECOGNITION (OCR) SYSTEM

is written by me and is my own effort and that no part has been plagiarized
without citations.

STUDENT :  _____ Date: 14/07/2011
(MOHD SHAHRIZAN BIN MOHD IESA)

SUPERVISOR :  _____ Date: 14/07/2011
(DR. GEDE PRAMUDYA ANANTA)

DEDICATION

To my parents, Mohamad Iesa Bin Bakong and Tuyah Binti Gundi

ACKNOWLEDGEMENTS

I will praise to my God Allah S.W.T who always guide me to the correct path along the project session. Who always help me to success this Projek Sarjana Muda (PSM).

I am most indebted to my beloved family; my parents who always give me ideas, feedbacks, and mental and spiritual support regarding this projects.

I am also very grateful for the supervision from Dr. Gede Pramudya Ananta, who gives me knowledge about Intelligent Tutoring Systems and intensive guidance throughout this project development.

I am also wishes to acknowledge my best friends in BITI UTeM, and Indonesian students for their support, critics, and feedback, also helping me in presentation practice.

ABSTRACT

The Optical Character Recognition (OCR) System is a simple system that can be used to translate the image that contains text to the computerized text that display in note pad file. This system is using template matching technique that is use the training data to compare the input to generate the output. So the training data will learn the input until it matches the right pattern to generate the text in note pad file.

The system is dedicated for education and research purpose and more focus to people who are work with writing things.

ABSTRAK

Sistem Optical Character Recognition (OCR) adalah satu sistem yang mudah yang boleh digunakan untuk menterjemahkan imej yang mengandungi teks kepada teks yang bertaip di komputer yang dipaparkan di notepad. Sistem ini menggunakan teknik menyamakan template dengan menggunakan data yang telah ditetapkan sebagai rujukan (Training Data) dengan input untuk menghasilkan output. Jadi data yang telah ditetapkan akan mempelajari corak input sehingga menghasilkan corak yang sama dengan corak input untuk menghasilkan output yang akan dipaparkan di notepad.

Sistem ini khusus digunakan untuk tujuan pendidikan dan penyelidikan dan lebih diberi tumpuan kepada orang-orang yang banyak melakukan aktiviti menulis ketika berkerja.

TABLE OF CONTENTS

CHAPTER	SUBJECT	PAGE
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGEMENTS	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
CHAPTER I	INTRODUCTION	
	1.1 Project Background	1
	1.2 Problem Statement	2
	1.3 Objectives	2
	1.4 Scope	3
	1.5 Project Significance	3
	1.6 Expected Output	4
	1.7 Conclusion	4
CHAPTER II	LITERATURE REVIEW AND PROJECT METHODOLOGY	
	2.1 Introduction	5
	2.2 Facts and Finding	5
	2.2.1 Domain	6
	2.2.2 Existing System	6
	2.2.3 Technique	7
	2.3 Project Methodology	8

2.4	Project Requirements	8
	2.4.1 Software Requirement	8
	2.4.2 Hardware Requirement	8
2.5	Project Schedule and Milestones	9
2.6	Conclusion	10

CHAPTER III ANALYSIS

3.1	Introduction	11
3.2	Problem analysis	12
	3.2.1 Proposed system	14
3.3	Requirement analysis	15
	3.3.1 Data Requirement	15
	3.3.2 Functional Requirement	15
	3.3.3 Non-functional Requirement	16
3.4	Conclusion	17

CHAPTER IV DESIGN

4.1	Introduction	18
4.2	High-Level design	19
	4.2.1 System architecture	19
	4.2.2 User interface design	21
	4.2.3 Navigation design	22
	4.2.4 Input design	24
	4.2.5 Technical design	25
	4.2.6 Output design	25
4.3	Database design	26
4.4	Detailed design	26
	4.4.1 Software design	26
	4.4.2 Physical Database design	28
4.5	Conclusion	28

CHAPTER V IMPLEMENTATION

5.1	Introduction	30
5.2	Software or Hardware Development Environment setup	31

5.3	Software or Hardware Configuration Management	32
	5.3.1 Configuration environment setup	32
	5.3.2 Version Control Procedure	33
5.4	Implementation Status	33
5.5	Conclusion	35

CHAPTER VI TESTING

6.1	Introduction	36
6.2	Test plan	37
	6.2.1 Test Organization	37
	6.2.2 Test Environment	38
	6.2.3 Test Schedule	39
6.3	Test Strategy	40
	6.3.1 Classes of Tests	40
6.4	Test Implementation	41
	6.4.1 Test Description	41
	6.4.2 Test Data	41
6.5	Test Result and Analysis	42
6.6	Conclusion	44

CHAPTER VII PROJECT CONCLUSION

7.1	Observation on Weakness and Strengths	45
	7.1.1 Strengths	46
	7.1.2 Weakness	46
7.2	Propositions for Improvement	47
7.3	Contribution	48
7.4	Conclusion	48

Bibliography	50
---------------------	-----------

Appendix	51
-----------------	-----------

The source code	51
-----------------	----

LIST OF FIGURE

FIGURE	TITLE	PAGE
2.1	Project Schedule & Milestone	9
3.1	Segmentation Process	12
3.2	Pattern Classification Process	13
3.3	UML Diagram System Flow	14
4.1	The OCR's system architecture	19
4.2	Flow chart for the OCR's system	20
4.3	Sample database design.	28
5.1	Single-tier Architecture	31

LIST OF TABLE

TABLE	TITLE	PAGE
4.1	User interface	21
4.2	Input design for the OCR system	24
4.3	Technical design for the OCR's system	25
5.1	Environment Setup	33
5.2	Implementation Progress	34
6.1	Test Organization	38
6.2	Software and Hardware requirement	38
6.3	Test Schedule	39
6.4	Result	43

CHAPTER I

INTRODUCTION

1.1 Project Background

Nowadays, everyone likes to read a journal, novel or other reading materials to gain knowledge, but not everyone like to read the whole reading materials to find the information that their want. Maybe they will only searching the information that their want at the certain part or at the certain pages in the journal. It is easy for them after their found the information that their want, by writing it on the note or on a piece of paper.

So, to make the converting process from the text inside the reading materials to the computerized writing in the computer is more easier, the Optical Character recognition (OCR) will help the user to solve the problem. Optical Character recognition (OCR) is design to translate the pattern of the text on the reading materials for example the text from book or journal to a computerized text without need the user to type the

text. The system will automatically write in the notepad or Microsoft word in the computer according to the text in the reading materials that had been select by the user.

Optical Character recognition (OCR) is a simple system that can be run easily by the user, just select the text that we want to copy then wait the system to translate the text. Optical Character recognition (OCR) is not a new thing that have been invent today, a lot of system or system have been invent same as the Optical Character recognition (OCR) function that is translate the text into computerized text but Optical Character recognition (OCR) it is design to more user friendly, easy to use, easy to care and make the user comfortable with it.

1.2 Problem Statement

From the background, we can know that the problem statement for this project is how to create an intelligent system that can detect the text from the book and translate the pattern of the text into computer digitalized.

1.3 Objectives

- Create a system that can scan the pattern of the text and translate the pattern of the text.
- Provide an easy to use system to the user when doing their daily routine.
- Helping those who face the problem when writing.

1.4 Scope

Optical Character recognition (OCR) is specialized to detect/learn the pattern of the text. This system is a part of image processing in Artificial intelligence field which is specialized to translate pattern.

This project's scope shows the development of Optical Character recognition (OCR) system, which is included with how the system works. This project's target will be ranged from junior high school student to high school student or someone who works with writing job for example journalist or reporter.

1.5 Project Significance

Optical Character recognition (OCR) target will be ranged from junior high school student to high school student or someone who works with writing job

This system will make the user easy to write or copy something while reading. It is no need for the user to write the information on a paper, but directly store the information that their want in the computer. It is more safety for the user because sometime, when we write something on a piece of paper, we will always forget where we put the note after few days writing the note. By having this Optical Character recognition (OCR) all the writing and copying things will be done easily.

1.6 Expected Output

Optical Character recognition (OCR) is developed based on several existing image processing that work with pattern recognition. This is not a new system, but this project is expected to add a new innovation to the current existing technology. In this case, we proposed Optical Character recognition (OCR), to work better and more flexible when detect different pattern such as symbol.

In the other hand, to make the product of this project ready to be implemented and directly capable to help student or someone who works with writing job, it is expected that this product is easy to install, easy to use, and it is easy for user to understand how the system run. The most important thing this system can be run at any desktop without need a high-performance computer or high-performance laptop to run it.

1.7 Conclusion

Optical Character recognition (OCR) is developed to help the user in writing and copying process from the text in reading materials will be translate in computerized writing. Besides that, it helps the user to minimize the time that use for writing or copying text.

Optical Character recognition (OCR) also can show to us that artificial intelligence can give us a lot of benefits when we can apply it in our daily routine. It can help the student to think and also give them idea to invent something that can help them self although it is just a simple thing.

In the next chapter, Literature review will be discussed and the explanation of project methodology about the Optical Character recognition (OCR) will be proceed.

CHAPTER II

LITERATURE REVIEW AND PROJECT METHADODOLOGY

2.1 Introduction

In this chapter, literature review, project methodology and project requirements will be discussed. This chapter will explain and provide a guideline to satisfy the objective of the project. All of the reference that relevant to the Optical Character recognition (OCR) also will be explained for this chapter.

2.2 Facts and Finding

Some research must be made to search some facts and finding about the project to make me clearly understand about my project before I plan and decide my technique and methodology.

2.2.1 Domain

The Optical Character recognition (OCR) can be classified as an Intelligent Program, a developed sub domain under Artificial Intelligence in image processing field. The program represents a more specific type of image translation or pattern recognition.

2.2.2 Existing System

The early developments The Optical Character recognition (OCR) is in 1929 where Gustav Tauschek obtained a patent on OCR in Germany. He has design a machine with a mechanical device that used templates and a photo detector. The OCR become more helpful in 1949 when Radio Corporation of America (RCA) engineers worked on the first primitive computer-type OCR it help blind people for the US Veterans Administration, but instead of converting the printed characters to machine language, their device converted it to machine language and then spoke the letters. The first commercial system was installed at the Reader's Digest in 1955. Nowadays, there are many OCR-based projects that aimed to help people when converting an image text to machine language. For example:

- Tesseract (Created by Hewlett-Packard; under further development by Google, 2010)
- ABBYY FineReader (ABBYY, 2009)
- Image to OCR Converter (Soft Solutions Limited of India, 2010)
- OmniPage (Philip Bernzott, John Dilworth, David George, Bryan Higgins, and Jeremy Knight, 2009)

Many techniques have been applied to complete the OCR project, although some techniques are different but the process or the framework is still the same.

2.2.3 Technique

The earliest technique that always been use in Optical Character Recognition (OCR) is something called matrix, or pattern matching. Before all the matching process proceeds, segmentation of the image is the most important part should be done first. Segmentation is the process that separates the text with other text so the inputs can easily matching with the template one by one per letter.

OCR programs which use the pattern matching method have bitmaps stored for every character of each of the different font and type sizes. By comparing a database of stored bitmaps distributed to the bitmaps of the scanned letters the program attempts to recognize the letters. This technique was only really successful using non-proportional fonts like Courier where letters are spaced regularly and are easier to identify.

Feature extraction was the next step in OCR's development. This attempted to recognize characters by identifying their universal features, the goal being to make OCR typeface-independent. If all characters could be identified using rules defining the way that loops and lines join each other, then individual letters could be identified regardless of their typeface. For example: the letter 'a' is made from a circle, a line on the right side and an arc over the middle. The arc over the middle is optional. So, if a scanned letter had these features it would be correctly identified as the letter a by the OCR program.

No OCR software ever recognizes 100% of the scanned letters. Some OCR programs use the matrix/pattern matching and/or feature extraction methods to recognize as many characters as possible - and complement this by using spell checking on the hitherto unrecognized letters.

2.3 Project Methodology

The Optical Character recognition (OCR) is developed using Object Oriented Analysis and Design approach. The program is learning from the default template or training data to train a new input to generate output.

2.4 Project Requirements

Project requirement are the equipment that I needed in completing this project. There are two types of requirement that is hardware and software requirement.

2.4.1 Software Requirement

There are several software which are used to develop the system, and its report

- Microsoft Office Professional 2007
- Matlab R2009a
- Notepad

2.4.2 Hardware Requirement

- Scanner

2.5 Project Schedule and Milestones

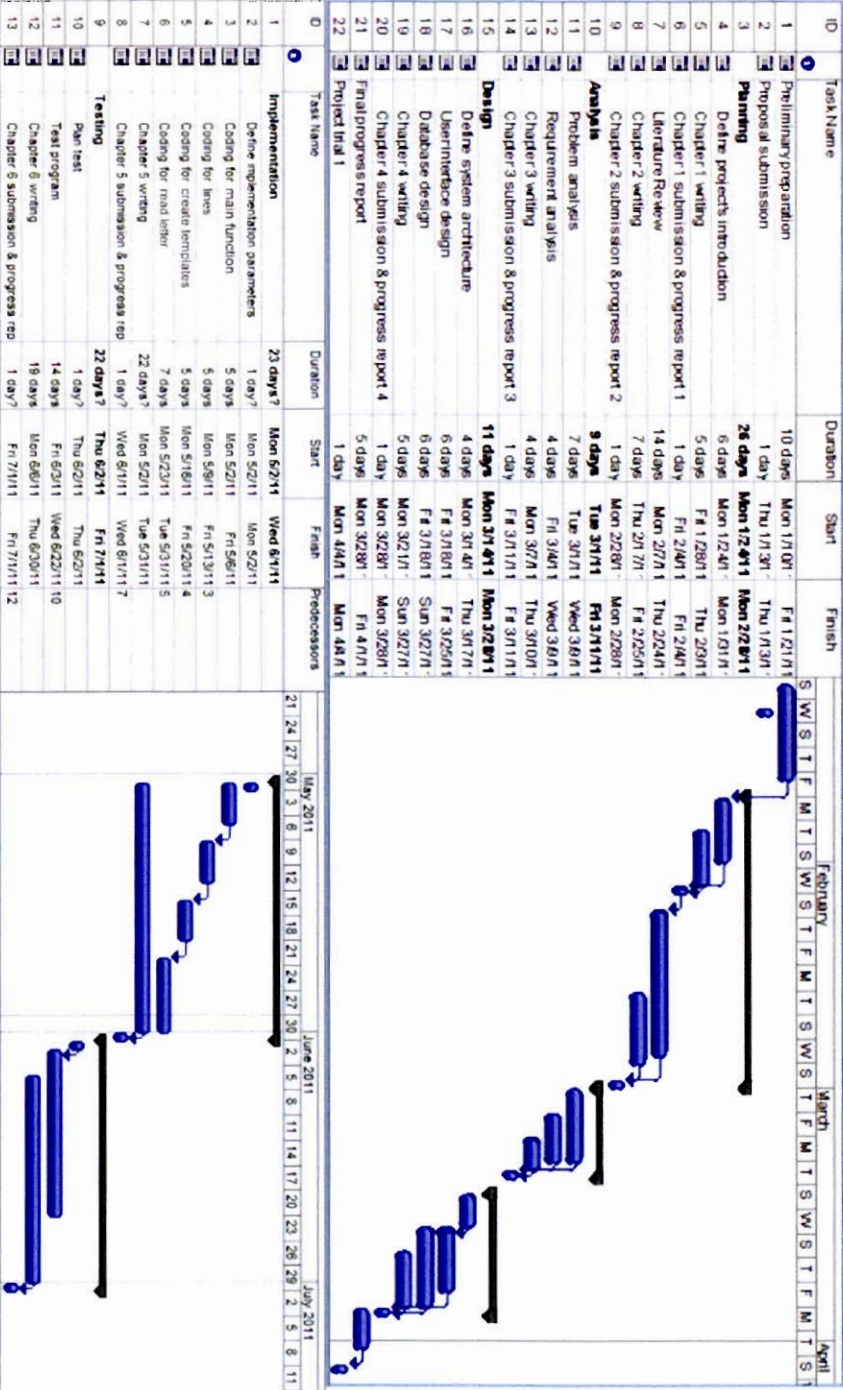


Figure 2.1: Project Schedule & Milestone

2.6 Conclusion

The Optical Character recognition (OCR) is not a new program that have been invented, there a lot of program that can translate the image that contains character (text) to computerized text have been invent nowadays.

The development of the Optical Character recognition (OCR) has started since 1929 but until nowadays there is still no perfect software or program than invented to translate 100% correct from the scanned image. The Optical Character recognition (OCR) is still under develop to find the most accurate way in scanned image translation.

In the next chapter, analysis phase of the Optical Character recognition (OCR) development will be elaborate to get a clear explanation about description about system's requirements.

CHAPTER III

ANALYSIS

3.1 Introduction

The development of the Optical Character Recognition (OCR) depends to its system requirement. To proceed with this system requirement, the analysis study about the system should be done first. Problem analysis is the first step that should we take in analysis study phase which is where we obtain or describe the problem then produce the problem statement.

When the results (problem statement) have been defined, we will proceed with next step that is requirement analysis. Requirement analysis defines the requirement of the system that included with Data requirement, Functional requirement and Non-functional requirement.