

B.Comp. Dissertation CA Report

Benchmarking and Improving OCR System for Southeast Asian Languages

By
Qiu Jiasheng, Jason

Department of Computer Science
School of Computing
National University of Singapore

2024/2025

B.Comp. Dissertation CA Report

Benchmarking and Improving OCR System for Southeast Asian Languages

By

Qiu Jiasheng, Jason

Department of Computer Science
School of Computing
National University of Singapore

2024/2025

Project ID: H0792230

Supervisor: A/P Min-Yen Kan

Advisor: Tongyao Zhu

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Subject Descriptors:

I.2.7 Natural Language Processing

Keywords:

Optical Character Recognition

Implementation Software and Hardware:

Python, Tesseract, EasyOCR

Acknowledgement

I would like to thank my supervisor, A/P Kan Min-Yen, and my advisor, Tongyao Zhu, for their invaluable guidance and mentorship. Their encouragement and constructive guidance have been a significant source of inspiration throughout the project.

Table of Contents

Abstract	ii
Acknowledgement	iii
Table of Contents	iii
1 Introduction	1
2 Related Work	2
3 Methodology	3
3.1 Data Collection	3
3.2 Benchmarking OCR Tools	3
4 Future Work	4

Introduction

Related Work

Methodology

3.1 Data Collection

3.2 Benchmarking OCR Tools

Future Work