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, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer 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.

List of Tables

4.1 Character Error Rate 4
4.2 Word Error Rate 4

Table of Contents

Abstract **ii**

Acknowledgement **iii**

List of Tables **iii**

Table of Contents **iv**

1 Introduction **1**

2 Related Work **2**

2.1 OCR on Low-Resource Languages 2

2.2 Benchmarking OCR 2

3 Methodology **3**

3.1 Data Collection 3

3.2 Benchmarking OCR Tools 3

4 Results **4**

5 Future Work **5**

References **6**

Introduction

Optical Character Recognition (OCR) is the process of detecting and converting text in a image into a computer-friendly text format (Santos, 2019).

This project aims to answer the following research questions (RQs):

- **RQ1.** How do popular OCR tools perform on Southeast Asian scripts?
- **RQ2.** What specific linguistic and script-related challenges affect OCR accuracy on Southeast Asian languages?
- **RQ3.** How does the choice of OCR tool impact accuracy on Southeast Asian scripts?

Related Work

2.1 OCR on Low-Resource Languages

(Ignat et al., 2022)

2.2 Benchmarking OCR

Methodology

3.1 Data Collection

3.2 Benchmarking OCR Tools

Results

	EasyOCR	Tesseract
English	0.17	0.20
Indonesian	0.20	0.18
Vietnamese	0.30	0.39
Thai	0.26	0.51

Table 4.1: Character Error Rate

	EasyOCR	Tesseract
English	0.25	0.29
Indonesian	0.27	0.33
Vietnamese	0.31	0.42
Thai	1.68	1.77

Table 4.2: Word Error Rate

Future Work

References

Ignat, O., Maillard, J., Chaudhary, V., & Guzmán, F. (2022). OCR improves machine translation for low-resource languages. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the association for computational linguistics: Acl 2022* (pp. 1164–1174). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.92