B.Comp. Dissertation

# Benchmarking and Improving OCR Systems for Southeast Asian Languages

By

Qiu Jiasheng, Jason

Department of Computer Science

School of Computing

National University of Singapore

2024/2025

B.Comp. Dissertation

# Benchmarking and Improving OCR Systems 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

Deliverables:

Report: 1 Volume

# Abstract

While Optical Character Recognition (OCR) has been widely studied for high-resource languages such as English and Chinese, the efficacy and limitations of OCR models on Southeast Asian (SEA) languages remain largely unexplored. This study aims to bridge this gap by assessing and improving the performance of OCR technologies on SEA languages. To achieve this objective, we propose a reusable pipeline to gather SEA-language text from Wikipedia and benchmark popular OCR tools.

Subject Descriptors:

H.3.3 Information Search and Retrieval

I.2.7 Natural Language Processing

I.2.10 Vision and Scene Understanding

Keywords:

Optical Character Recognition, Southeast Asian Languages

Implementation Software and Hardware:

Python, Tesseract, EasyOCR

## Acknowledgements

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 Figures

3.1 Pipeline for data collection from Wikipedia . . . . . . . . . . . . . . . . . . 6

# List of Tables

4.1 English Error Classification . . . . . 8
4.2 Indonesian Error Classification . . . . . 8
4.3 Vietnamese Error Classification . . . . . 9
4.4 Thai Error Classification . . . . . 9
A.1 Dataset of 98 Wikipedia articles . . . . . 14

# Table of Contents

**Title** i

**Abstract** ii

**Acknowledgements** iii

**List of Figures** iii

**List of Tables** iv

**Table of Contents** v

**1 Introduction** 1

**2 Related Work** 3

**3 Methodology** 4

3.1 Experiment Setup . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

3.1.1 OCR Systems . . . . . . . . . . . . . . . . . . . . . . . . . 4

3.1.2 Evaluation Metrics . . . . . . . . . . . . . . . . . . . . . . 5

3.1.3 Source of Data . . . . . . . . . . . . . . . . . . . . . . . . 5

3.1.4 Languages . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

3.2 Experiment 1: Benchmarking on Real Data . . . . . . . . . . . . . . 6

3.2.1 Data Collection . . . . . . . . . . . . . . . . . . . . . . . . 6

3.3 Experiment 2: Benchmarking on Synthetic Data . . . . . . . . . . . 7

3.3.1 Synthetic Data Generation . . . . . . . . . . . . . . . . . . . 7

3.4 Experiment 3: Fine-tuning . . . . . . . . . . . . . . . . . . . . . . . 7

**4 Results 8**

4.1 RQ1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

4.2 RQ2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

4.3 RQ3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

**5 Discussion 10**

**6 Conclusion 11**

**References 12**

**A Wikipedia Article Dataset 14**

# Chapter 1

# Introduction

Current research in Natural Language Processing (NLP) is heavily concentrated on 20 of the 7,000 languages in the world (Magueresse et al., 2020). In particular, Southeast Asia (SEA) is home to over 1,000 languages but remains a relatively under-researched region in NLP (Aji et al., 2023). Similar to most low-resource languages, a major challenge in developing NLP systems for SEA languages is the limited availability of datasets for the region's languages. Although many scanned documents and books in these low-resource languages are available online, the text within these files remains inaccessible due to formats like images and PDFs.

A solution to this problem is to use Optical Character Recognition (OCR) to extract the textual data. OCR is the process of identifying and converting text in an image into a computer-friendly text format. By extracting the text from these scanned documents, OCR can generate valuable datasets for low-resource languages. The created datasets can then be used for downstream NLP tasks, such as machine translation, training large language models, and named-entity recognition (Agarwal & Anastasopoulos, 2024; Ignat et al., 2022). Therefore, studying OCR performance on SEA languages is crucial to accelerating NLP research in the region.

While OCR has been widely studied for high-resource languages such as English and Chinese, the efficacy and limitations of OCR models on SEA languages remain largely unexplored. To address this gap, we propose a reusable pipeline to collect textual data in low-resource SEA languages from Wikipedia and benchmark popular open-source OCR tools on the collected data. The primary objective is to benchmark and improve the performance of OCR technologies on SEA languages, thereby contributing to the advancement of NLP applications in this linguistically diverse region. Specifically, this project seeks to answer the following research questions (RQs):

- **RQ1.** How do popular OCR tools perform on SEA scripts?
- **RQ2.** What specific linguistic and script-related challenges affect OCR accuracy on SEA languages?
- **RQ3.** What techniques and recommendations can enhance OCR accuracy on SEA languages?

# Chapter 2

# Related Work

# Chapter 3

# Methodology

## 3.1 Experiment Setup

### 3.1.1 OCR Systems

In our selection of OCR systems for benchmarking, we prioritized open-source solutions that support a diverse range of SEA languages, as this approach enhances accessibility and reusability for the proposed evaluation pipeline. Consequently, we selected to use Tesseract and EasyOCR.

Tesseract[1] is an established OCR engine, recognized as one of the top performers in the 1995 UNLV Test (Rice et al., 1995). It utilizes an underlying Long Short-Term Memory (LSTM) model. EasyOCR[2] is a modern OCR framework that integrates a text detection model based on the Character Region Awareness for Text (CRAFT) algorithm with a recognition model utilizing a Convolutional Recurrent Neural Network (CRNN). Both Tesseract and EasyOCR provide robust support for English, Indonesian, Vietnamese, and Thai, making them suitable candidates for our benchmarking study.

[1] https://github.com/tesseract-ocr/tesseract
[2] https://github.com/JaidedAI/EasyOCR

### 3.1.2 Evaluation Metrics

$$CER = \frac{I + D + S}{N} \tag{3.1}$$

Similar to most OCR benchmark studies, we utilize Character Error Rate (CER) and Word Error Rate (WER) as our evaluation metrics (Hegghammer, 2022; Ignat et al., 2022). CER measures the accuracy of character recognition and is calculated using the Levenshtein or edit distance, which represents the minimum number of single-character insertions (I), deletions (D), and substitutions (S) required to transform one word into another. As shown in Equation 3.1, CER is defined as the edit distance between the OCR-predicted text and ground truth text, divided by the total number of characters in the ground truth text (N). A lower CER value indicates higher accuracy, with 0 representing perfect recognition. Notably, CER can exceed 1, particularly when there are a significant number of insertions. WER serves as the word-based counterpart to CER.

### 3.1.3 Source of Data

We chose to use Wikipedia as our text corpus for several reasons. Firstly, Wikipedia articles can be easily converted into images via screenshots, making them suitable for OCR applications. The platform also offers a convenient source of ground truth through its APIs that provide plain text for most articles. Secondly, Wikipedia hosts a large corpus in several popular SEA languages, including Thai, Vietnamese, Indonesian, Tamil, and Burmese, supporting our language needs ("List of Wikipedias", 2024). Lastly, Wikipedia articles contain visual elements like images and tables that are common in modern real-world documents.

### 3.1.4 Languages

From the languages available on Wikipedia, we selected English, Indonesian, Vietnamese, and Thai text. English serves as a baseline for sanity checks and bug fixing. The remaining SEA languages were chosen to capture diverse script characteristics. Indonesian represents Latin-based scripts, Vietnamese represents Latin scripts with diacritics, and Thai represents non-Latin scripts.

## 3.2 Experiment 1: Benchmarking on Real Data

### 3.2.1 Data Collection

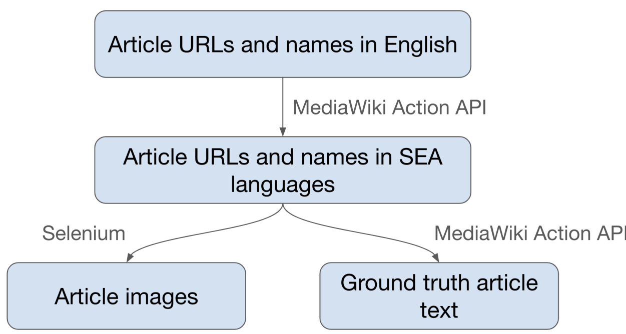


Figure 3.1: Pipeline for data collection from Wikipedia

From the dataset of 100 Wikipedia articles, we collected article images and ground truth article text in our selected languages using Python, Selenium[3], and the MediaWiki Action API[4]. Figure 3.1 illustrates the overall pipeline for data collection. The detailed steps are as follows:

[3]Selenium is a framework for automating web browsers, commonly used for web scraping by programmatically interacting with websites.

[4]The MediaWiki Action API allows access to wiki page operation features such as search and retrieval.

1. Manually compile the dataset's article names and URLs in English.

2. Fetch the article names and URLs in Thai, Vietnamese, and Indonesian from the MediaWiki Action API.

3. Download the article PDFs in all languages using Selenium.

4. Convert the article PDFs into PNG images, where each image represents one page in the PDF.

5. Download the ground truth article text into TXT files from the MediaWiki Action API.

## 3.3 Experiment 2: Benchmarking on Synthetic Data

### 3.3.1 Synthetic Data Generation

## 3.4 Experiment 3: Fine-tuning

# Chapter 4

# Results

## 4.1 RQ1

## 4.2 RQ2

Table 4.1: English Error Classification

| | Count | EasyOCR % Missed | Tesseract % Missed | GOT % Missed |
|---|---|---|---|---|
| Arabic digit | 38,324 | 0.7% | 1.9% | 0.3% |
| Latin letter | 1,546,964 | 1.3% | 1.8% | 0.4% |
| Latin letter w/ diacritic | 424 | 100.0% | 53.1% | 14.6% |
| Punctuation | 53,403 | 28.4% | 2.3% | 3.4% |
| Whitespace | 317,587 | 4.9% | 4.3% | 3.6% |
| Other | 3,298 | 82.8% | 68.5% | 76.9% |

Table 4.2: Indonesian Error Classification

| | Count | EasyOCR % Missed | Tesseract % Missed | GOT % Missed |
|---|---|---|---|---|
| Arabic digit | 24,947 | 0.4% | 1.8% | 0.2% |
| Latin letter | 1,208,707 | 0.5% | 1.8% | 0.4% |
| Latin letter w/ diacritic | 262 | 5.3% | 100.0% | 15.3% |
| Punctuation | 37,788 | 22.1% | 3.1% | 0.8% |
| Whitespace | 207,556 | 4.8% | 5.1% | 4.1% |
| Other | 2,468 | 72.2% | 80.5% | 43.2% |

Table 4.3: Vietnamese Error Classification

| | Count | EasyOCR % Missed | Tesseract % Missed |
|---|---|---|---|
| Arabic digit | 31,473 | 1.1% | 2.2% |
| Latin letter | 916,667 | 8.5% | 1.8% |
| Latin letter w/ diacritic | 292,686 | 14.8% | 1.8% |
| Punctuation | 40,420 | 24.6% | 2.2% |
| Whitespace | 367,936 | 10.9% | 5.3% |
| Other | 35,767 | 12.5% | 7.7% |

Table 4.4: Thai Error Classification

| | Count | EasyOCR % Missed | Tesseract % Missed |
|---|---|---|---|
| Arabic digit | 22,580 | 0.9% | 6.7% |
| Latin letter | 36,174 | 100.0% | 100.0% |
| Latin letter w/ diacritic | 96 | 100.0% | 100.0% |
| Thai letter | 617,699 | 0.4% | 3.1% |
| Thai diacritic | 90,620 | 3.7% | 3.6% |
| Punctuation | 13,669 | 6.4% | 8.4% |
| Thai punctuation | 901 | 78.8% | 3.9% |
| Whitespace | 58,164 | 37.5% | 37.2% |
| Other | 306,647 | 2.2% | 7.1% |

## 4.3 RQ3

# Chapter 5

# Discussion

# Chapter 6

# Conclusion

# References


Agarwal, M., & Anastasopoulos, A. (2024). A concise survey of OCR for low-resource languages. In M. Mager, A. Ebrahimi, S. Rijhwani, A. Oncevay, L. Chiruzzo, R. Pugh, & K. von der Wense (Eds.), *Proceedings of the 4th workshop on natural language processing for indigenous languages of the americas (americasnlp 2024)* (pp. 88–102). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.americasnlp-1.10

Aji, A. F., Forde, J. Z., Loo, A. M., Sutawika, L., Wang, S., Winata, G. I., Yong, Z.-X., Zhang, R., Doğruöz, A. S., Tan, Y. L., & Cruz, J. C. B. (2023). Current status of NLP in South East Asia with insights from multilingualism and language diversity. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 8–13. https://aclanthology.org/2023.ijcnlp-tutorials.2

Hegghammer, T. (2022). OCR with Tesseract, Amazon Textract, and Google Document AI: A benchmarking experiment. *Journal of Computational Social Science*, *5*, 861–882. https://doi.org/https://doi.org/10.1007/s42001-021-00149-1

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

List of wikipedias. (2024). https://en.wikipedia.org/wiki/List_of_Wikipedias

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. *CoRR*, *abs/2006.07264*. https://arxiv.org/abs/2006.07264

Rice, S., Jenkins, F., & Nartker, T. (1995). *The fourth annual test of OCR accuracy* (tech. rep.). Information Science Research Institute.

# Appendix A

# Wikipedia Article Dataset

| Category | Articles |
|---|---|
| People | Elizabeth II, Barack Obama, Michael Jackson, Elon Musk, Lady Gaga, Adolf Hitler, Eminem, Lionel Messi, Justin Bieber, Freddie Mercury, Kim Kardashian, Johnny Depp, Steve Jobs, Dwayne Johnson, Michael Jordan, Taylor Swift, Stephen Hawking, Kanye West, Donald Trump |
| Present countries | United States, India, United Kingdom, Canada, Australia, China, Russia, Japan, Germany, France, Singapore, Israel, Pakistan, Philippines, Brazil, Italy, Netherlands, New Zealand, Ukraine, Spain |
| Cities | New York City, London, Hong Kong, Los Angeles, Dubai, Washington, D.C., Paris, Chicago, Mumbai, San Francisco, Rome, Monaco, Toronto, Tokyo, Philadelphia, Machu Picchu, Jerusalem, Amsterdam, Boston |
| Life | Cat, Dog, Animal, Lion, Coronavirus, Tiger, Human, Dinosaur, Elephant, Virus, Horse, Photosynthesis, Evolution, Apple, Bird, Mammal, Potato, Polar bear, Shark, Snake |
| Buildings and structures | Taj Mahal, Burj Khalifa, Statue of Liberty, Great Wall of China, Eiffel Tower, Berlin Wall, Stonehenge, Mount Rushmore, Colosseum, Auschwitz concentration camp, Great Pyramid of Giza, One World Trade Center, Empire State Building, White House, Petra, Large Hadron Collider, Hagia Sophia, Golden Gate Bridge, Panama Canal, Angkor Wat |

Table A.1: Dataset of 98 Wikipedia articles