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 evaluating OCR technologies for SEA languages and exploring script-specific challenges. We propose a pipeline to collect textual data from Wikipedia and benchmark open-source OCR tools. Additionally, we demonstrate the potential of fine-tuning existing models on SEA languages, aiming to expand OCR capabilities for these languages.

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

List of Tables

Table of Contents

Chapter 1

Introduction

Current research in Natural Language Processing (NLP) is heavily concentrated on only 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). A similar trend can be observed in Optical Character Recognition (OCR) research, where the focus is predominantly on high-resource languages (Salehudin et al., 2023; R. Smith, 2007), leaving many SEA languages underserved.

OCR, the process of converting textual images into machine-readable formats, offers significant potential for languages with limited datasets. While many scanned documents and books in these low-resource languages are available online, the text within them often remains inaccessible due to formats like images and PDFs. By extracting the text from these documents, OCR can generate valuable datasets for low-resource languages, which can then be used for downstream NLP tasks, such as machine translation and named-entity recognition (Agarwal & Anastasopoulos, 2024; Ignat et al., 2022). Therefore, studying OCR performance on SEA languages is crucial to accelerating NLP research and technology development 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, this study presents a pipeline to collect textual data from Wikipedia and benchmark several open-source OCR tools on the collected data.

Additionally, we explore the potential of fine-tuning existing models to improve OCR performance on SEA languages. The primary objective is to evaluate and enhance the performance of OCR technologies on SEA languages, thereby advancing 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 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

2.1 Overview of OCR Systems

To benchmark OCR performance programmatically, there are two broad categories of OCR systems: open-source and commercial.

Open-source OCR systems are characterized by their accessibility, allowing users to view, modify, and distribute the source code freely. This transparency enables developers to customize the tools to meet specific needs. Furthermore, open-source systems typically incur no licensing fees, making them cost-effective options for research purposes. For instance, Tesseract (R. W. Smith, 2013) is a popular open-source OCR engine mentioned consistently in related studies (Hegghammer, 2022; Ignat et al., 2022).

Commercial OCR systems are typically accessed through paid services via application programming interfaces (APIs). Notable examples of these systems include Amazon Textract, Google Document AI, and Google Vision API. Generally, commercial off-the-shelf OCR tools tend to perform better than their open-source counterparts (Hegghammer, 2022; Ignat et al., 2022). However, the proprietary nature of these "black box" models limits their utility for research purposes, such as fine-tuning and customization. Additionally, the associated costs of these services contribute to fewer studies focused on commercial systems in comparison to the more widely researched open-source tools.

2.2 OCR on Low-resource Languages

Applying OCR yields a plain text prediction, which is then compared with the ground truth data to assess the tool's accuracy and performance. Recent studies have demonstrated that OCR systems tend to perform better on artificially generated data than on real-world data (Ignat et al., 2022). This observation suggests that synthetic datasets may not fully capture the complexities of authentic documents, which often feature issues like imperfect text alignment, varied fonts, and complex layouts. Furthermore, the addition of synthetic noise significantly raises error rates, especially for open-source systems, which appear more susceptible to noise interference than their commercial counterparts (Hegghammer, 2022).

When comparing performance on different scripts, OCR tools generally achieve higher accuracy on scripts written in Latin alphabets (Hegghammer, 2022; Ignat et al., 2022). This disparity in performance partly stems from market incentives that prioritize the development of English-language OCR systems, resulting in more extensive training data and refinement for Latin-based scripts. Ornate scripts, such as those with complex diacritics or unique letter shapes, present additional challenges and tend to yield lower OCR accuracy.

In terms of benchmarking OCR on SEA languages, the most related study is the recent work by Ignat et al. (2022). They grouped 60 low-resource languages by region and script, including SEA languages like Khmer, Lao, Burmese, Thai, and Vietnamese. Their research revealed that while OCR models perform well on artificial SEA-language data, accuracy drops significantly on real-world data. This discrepancy underscores the need for more real-world training data to improve OCR outcomes for SEA languages.

In summary, there exists a gap in benchmarking OCR specifically for SEA languages, largely due to the lack of real-world training data. This project addresses this gap by proposing a reusable pipeline for benchmarking OCR performance on real-world SEA-language data sourced from Wikipedia.

2.3 Using Synthetic Data for OCR Evaluation

To evaluate OCR performance accurately, a collection of textual data in the form of images or PDFs paired with reliable ground truth is needed. Similar to most NLP tasks, data scarcity poses a major obstacle to advancing OCR technology in low-resource languages, where the limited availability of annotated textual data restricts both model training and evaluation. Although an abundance of scanned documents in these low-resource languages exists online, they lack the ground truth required for evaluation. While plain text in these languages is often available separately, it typically exists in text-based formats rather than images or PDFs, limiting its direct usefulness as ground truth for OCR.

To bridge this gap, many studies rely on artificial images and PDFs generated from plain text to create usable evaluation data. For instance, Ignat et al. (2022) artificially created PDFs from the Flores 101 dataset, which consists of text data from Wikipedia in 101 languages. Generalizing this concept of a document creation pipeline further, Gupte et al. (2021) published an open-source Python package for generating document images from plain text, including several document styling templates. Using these methods, high-quality low-resource language data paired with ground truth can be generated from text-based formats on a large scale.

A common trend in using artificial data is the augmentation of noise to simulate real-

world conditions, which often contain complex layouts, stains, and scribbles (Heggham-mer, 2022). Directly using noise-free data, i.e., single-column text in a clear font, limits OCR processors' usefulness on real-life scanned documents. Thus, noise augmentation is often applied to artificial data. Some popular techniques include changing font style, size, color, letter spacing, and adding Gaussian blur, bleed-through, and salt-and-pepper noise (Gupte et al., 2021; Ignat et al., 2022).

2.4 Fine-tuning OCR Systems

Chapter 3

Methodology

To answer the research questions, this study conducts three experiments to benchmark and improve OCR performance on SEA languages.

3.1 Experiment Setup

3.1.1 Languages

In this study, we chose to benchmark on English, Indonesian, Vietnamese, and Thai. English serves as a baseline comparison due to its extensive OCR research and established tool support. Meanwhile, Indonesian, Vietnamese, and Thai were selected as a representative subset of SEA languages for several reasons.

Firstly, these three languages encompass a range of script types: Latin scripts for Indonesian, Latin scripts with diacritics for Vietnamese, and Brahmic scripts for Thai. By covering these scripts, we capture a broad spectrum of orthographic features, from diacritics to tone marks and from Latin-based scripts to complex character shapes. This allows us to examine how these unique linguistic features impact OCR performance. Furthermore, many other SEA languages, including Malay, Filipino, and Cebuano, use modified Latin scripts, while languages like Khmer, Burmese, and Javanese use Brahmic scripts. Thus, findings from this study can be applied to other languages with similar script types, accelerating OCR research in the region.

Table 3.1: Benchmarked Languages

	Speaker Population	Script Type	Example
English	1.5 billion	Latin	Good morning
Indonesian	252 million	Latin	Selamat pagi
Vietnamese	97 million	Latin with diacritics	Chào buổi sáng
Thai	71 million	Brahmic	**สวัสดีตอนเช้า**

Note: Speaker population data from Wikipedia (2025).

Secondly, the wide usage of these languages makes it feasible to obtain textual data. The high number of speakers, active online communities, and abundant digital content ensure sufficient resources for OCR benchmarking. Their prominence in SEA further highlights their relevance, as improving OCR for these languages benefits a large portion of the region's population.

While this study covers only a small fraction of the languages spoken in SEA, the selection of these languages provides a strong starting point, as they cover popular script types and offer abundant online data for benchmarking.

3.1.2 Data Source

To collect textual data, this study uses Wikipedia due to its accessibility and multilingual scope. Wikipedia articles can be converted into images via screenshots, simulating real-world OCR scenarios. The platform also offers a convenient Application Programming Interface (API) that allows retrieval of plain text from most articles, serving as a reliable reference for evaluating OCR accuracy and generating synthetic documents. Moreover, the availability of large corpora in various SEA languages, including Thai, Vietnamese, Indonesian, Tamil, and Burmese, makes Wikipedia suitable for this study's language needs ("List of Wikipedias", 2024).

3.1.3 OCR Systems

In our selection of OCR systems for benchmarking, we prioritize open-source solutions that support a diverse range of SEA languages, promoting accessibility and reusability for the proposed evaluation pipeline. Additionally, we aim to include models with different underlying architectures, enabling a more comprehensive assessment of their performance across different languages. Consequently, we selected EasyOCR, Tesseract, and GOT, each representing distinct modeling approaches to OCR.

EasyOCR[1] 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).

Tesseract[2] 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.

Both EasyOCR and Tesseract provide robust support for English, Indonesian, Vietnamese, and Thai, making them suitable candidates for our benchmarking study.

3.1.4 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

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

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 when there is a significant number of insertions. WER serves as the word-based counterpart to CER.

3.2 Experiment 1: Benchmarking on Real Data

3.2.1 Data Collection

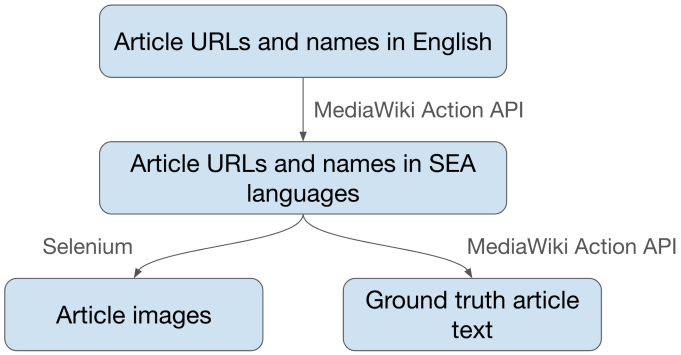

Figure 3.1: Pipeline for data collection from Wikipedia

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 for Vietnamese and Thai

Chapter 4

Results and Analysis

In this chapter, we analyze and evaluate the results of the experiments to provide answers to the research questions.

4.1 RQ1: How do popular OCR tools perform on SEA scripts?

4.1.1 OCR Accuracy

4.1.2 Runtime

Table 4.1: Average OCR runtime per page (seconds)

	EasyOCR	Tesseract	GOT
English	3.23	11.68	24.35
Indonesian	2.92	13.19	31.44
Vietnamese	3.91	11.80	-
Thai	2.32	16.76	-

Table 4.2: Error classification by character type for English articles

	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 4.3: Error classification by character type for Indonesian articles

	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 4.4: Error classification by character type for Vietnamese articles

	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 4.5: Error classification by character type for Thai articles

	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%

4.2 RQ2: What script-related challenges affect OCR accuracy on SEA languages?

4.3 RQ3: What techniques and recommendations can enhance OCR accuracy on SEA languages?

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

Gupte, A., Romanov, A., Mantravadi, S., Banda, D., Liu, J., Khan, R., Meenal, L. R., Han, B., & Srinivasan, S. (2021). Lights, camera, action! A framework to improve NLP accuracy over OCR documents. *CoRR*, *abs/2108.02899*. https://arxiv.org/abs/2108.02899

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 languages by total number of speakers. (2025). https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

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.

Salehudin, M., Basah, S., Yazid, H., Basaruddin, K., Safar, M., Som, M. M., & Sidek, K. (2023). Analysis of optical character recognition using easyocr under image degradation. *Journal of Physics: Conference Series*, *2641*(1), 012001. https://doi.org/10.1088/1742-6596/2641/1/012001

Smith, R. (2007). An overview of the tesseract ocr engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, *2*, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Smith, R. W. (2013). History of the Tesseract OCR engine: what worked and what didn't. In R. Zanibbi & B. Coüasnon (Eds.), *Document recognition and retrieval xx* (p. 865802). SPIE. https://doi.org/10.1117/12.2010051

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