សាកលវិទ្យាល័យភូមិន្ទភ្នំពេញ

ROYAL UNIVERSITY OF PHNOM PENH

ការស្រាវជ្រាវវិធីសាស្ត្រថ្មី សម្រាប់កំណត់សម្គាល់អត្ថបទអក្សរខ្មែរទូទៅ នឹង បានប្រើប្រាស់ស្ថាបត្យកម្ម Craft ជាមួយនឹង TrOCR

A novel End-to-End approach for General Khmer Text Recognition using Craft with TrOCR Architecture

Mr. Vitou Soy

A Thesis

In Partial Fulfilment of the Requirement for the Degree of
Bachelor of Engineering in Information-Technology-Engineering

Examination committee:	Mr. Sokchea Kor (Advisor)
	Mr. Chanpiseth Chap (committee)
	Mrs. Daly Chea (committee)
	Dr. . . . .

June 2025

មូលន័យសង្ខេប

ក្នុងសហគមន៍បច្ចេកវិទ្យាព័ត៌មានសម័យថ្មី ការចាប់យកអត្ថបទចេញពីរូបភាព – [OCR] (Optical Character Recognition) ក្លាយជាបច្ចេកវិទ្យាសំខាន់មួយដែលត្រូវបានប្រើប្រាស់ យ៉ាងទូលំទូលាយ សម្រាប់បំលែងឯកសារសរសេរ ឬរូបភាពអក្សរឱ្យទៅជាអត្ថបទ អេឡិចត្រូនិច (digital text)។ ការអភិវឌ្ឍ OCR សម្រាប់ភាសាខ្មែរ តែងតែប្រឈមនឹងបញ្ហាជាច្រើន ដោយសារកង្វះនៃប្រភពទិន្នន័យ និងឯកសារសម្រាប់ train AI model។ ដើម្បីដោះស្រាយបញ្ហានេះ យើងបានបង្កើតទិន្នន័យសិប្បនិម្មិត (Synthetic Dataset) ដោយប្រើវិធីសាស្ត្របច្ចេកទេសកម្រិតខ្ពស់។

ក្នុងដំណើរការបង្កើតទិន្នន័យសិប្បនិម្មិត (Synthetic Dataset) រួមមាន៖

- វិធីសាស្ត្រក្នុងការប្រមូលអត្ថបទចេញពីអ៊ីនធឺណិត មានដូចខាងក្រោម (Scrape data) ៖
  - ដំណាក់កាលទីមួយ៖ យើងបានប្រមូលអត្ថបទចេញពី khsearch.com, Chuon-Nath-Dictionary, Alpha-Word, Google-Word, និងចុងក្រោយគឺ Huggingface.com ។
  - ដំណាក់កាលទីពីរ៖ យើងបានសម្អាត ទិន្នន័យទាំងអស់នោះ ឆ្លងកាត់ដំណើរការ ដូចជា លុបចោលតួរអក្សរណាដែលមិនសូវមាន វត្តមាននៅលើ រូបភាព ញឹកញាប់ និងបានលុបចោល តួរអក្សរណាដែល Fonts renders អត់ចេញ។
  - ដំណាក់កាលទីបី៖ ដំណាក់កាលមួយនេះ យើងបានធ្វើការ កាត់ប្រយោគទាំងអស់នោះ ជាពាក្យៗ ដោយប្រើប្រាស់ library ឈ្មោះ khmer-nltk
  - ដំណាក់កាលទីបួន៖ ចុងក្រោយ ក៏បានរៀបចំជា ប្រយោគដែល មានប្រវែង Random ពី ១ អក្សរ រហូតដល់ ១១០ អក្សរ ។
- បង្កើតរូបភាពដោយអនុវត្តតាមលក្ខខណ្ឌខាងក្រោម ៖
  - ផ្ទៃខាងក្រោយចៃដន្យ (Apply Different backgrounds)
  - បំពាក់ពុម្ពអក្សរផ្សេងៗគ្នា (Apply Different fonts)
  - Noise: `gaussian_noise`, `salt_pepper_noise`, `speckle_noise`, `blur`
  - បង្វិលអក្សរបន្តិច (random rotation text)
  - បញ្ចូល Margin Randomly (1, 5) pixels
- សរុបមកយើងបានបង្កើត Data ជាង ៤ លាន records សម្រាប់ train OCR model

Architecture OCR ត្រូវបានបែងចែកជា ២ ផ្នែក៖ Text Detection និង Text Recognition:

Text Detection: យើងប្រើម៉ូដែល CRAFT ដោយបានធ្វើការ Train ឡើងវិញដោយ បាន annotation ទៅលើ លើរូបភាពប្រហែល ៥០០ images និងសរុបចំនួន bounding box ជាង ១០,០០០ boxes។

Text Recognition: យើងប្រើ TrOCR base model ចេញពី Microsoft (មាននៅក្នុង Hugging Face) ហើយបាន fine-tune ទៅលើ dataset ខ្មែរសិប្បនិម្មិត (Synthetic Dataset) ដើម្បីបង្កើនសមត្ថភាពក្នុងការសម្គាល់អក្សរខ្មែរ។

លទ្ធផលសិក្សាបានបង្ហាញថា OCR របស់ពួកយើងអាចសម្គាល់អត្ថបទចេញពីរូបភាព បានដោយភាពត្រឹមត្រូវលើសពី ៩០%។ ដូច្នេះ ការសិក្សានេះបង្ហាញអំពីសក្តានុពលនៃការបង្កើត dataset និងការប្រើម៉ូដែលជំនាន់ថ្មីដើម្បីអភិវឌ្ឍន៍ OCR ភាសាខ្មែរឱ្យមានប្រសិទ្ធភាពកាន់តែខ្ពស់។ វាមានសមត្ថភាព អាចចាប់យកអត្ថបទមិនត្រឹមតែពាក្យខ្លីៗ ប៉ុណ្ណោះទេ តែវាក៏អាចធ្វើការចាប់យក ដូចជា មួយតួអក្សរដោយមួយតួអក្សរ, ពាក្យដោយពាក្យ, ប្រយោគដោយប្រយោគ រហូតដល់ មួយប្រយោគវែង ១១០ តួអក្សរថែមទៀតផង ។ ហើយលើសពីនោះទៀត វាក៏អាចធ្វើការកំណត់សម្គាល់ទៅលើ ពីរ ភាសាចម្បង ទាំងភាសាខ្មែរ និងភាសាអង់គ្លេស ។

Abstract

In the modern era of information technology, Optical Character Recognition (OCR) has emerged as a crucial technology for converting printed or handwritten text from images into digital form. However, the development of OCR systems for the Khmer language presents significant challenges, primarily due to the lack of large-scale annotated datasets. To address this limitation, we constructed a high-quality synthetic dataset using an advanced data generation pipeline. Our Khmer OCR system consists of two core components:

- Text Collection: We gathered Khmer text data from various online sources, including khsearch.com, Chuon-Nath Dictionary, Alpha-Word, Google-Word, and Hugging Face.
- Data Cleaning: We processed and cleaned the collected text by removing uncommon characters, symbols that are rarely rendered correctly by fonts, and excessive whitespace.
- Text Segmentation: Sentences were tokenized into words using the khmer-nltk library, and then reconstructed into randomized sentence lengths ranging from 1 to 110 characters.
- Image Generation: We rendered text into synthetic images by:
  - Applying random backgrounds and a variety of Khmer fonts
  - Adding diverse noise types such as Gaussian noise, salt-and-pepper noise, speckle noise, and blur
  - Introducing slight random rotations and random margins (1–5 pixels)
- As a result, we generated over 4 million high-quality synthetic image-text pairs to train the OCR model.

Our Khmer OCR system consists of two core components:

- **Text Detection:** We fine-tuned the CRAFT (Character Region Awareness for Text Detection) model using 500 manually annotated images, totaling over 10,000 bounding boxes.
- **Text Recognition:** We fine-tuned Microsoft's TrOCR base model (available on Hugging Face) on our synthetic Khmer dataset to improve its ability to recognize Khmer text.

The evaluation results demonstrate that our system achieves a recognition accuracy exceeding 90%. These findings highlight the effectiveness of combining synthetic data generation with modern transformer-based architectures to significantly advance Khmer OCR capabilities. Notably, the system can accurately recognize a wide range of text—from single characters and individual words to full sentences of up to 110 characters—and supports both Khmer and English languages.

SUPERVISOR's RESEARCH SUPERVISION STATEMENT

Name of program: Khmer Studies
Name of candidate: Vitou Soy

Title of research report: A novel End-to-End approach for General Khmer Text Recognition using Craft with TrOCR Architecture

This is to certify that the research carried out for the above titled master's research report was completed by the above named candidate under my direct supervision. This thesis material has not been used for any other degree. The candidate has demonstrated strong research capabilities and independence in developing novel approaches for Khmer text recognition. The research methodology, implementation, and results are original contributions to the field of Khmer OCR technology. I have provided guidance and oversight throughout the research process while allowing the candidate to explore innovative solutions.

Supervisor's name: Sokchea Kor
Supervisor's signature:……………………
Date……………………………………………

CANDIDATE'S STATEMENT

TO WHOM IT MAY CONCERN

This is to certify that the dissertation that I, Vitou Soy, hereby present, entitled ”Advancing Khmer Optical Character Recognition: A Synthetic Data-Driven Approach,” for the degree of Bachelor of Engineering in Information Technology at the Royal University of Phnom Penh, is entirely my own work. Furthermore, it has not been used to fulfill the requirements of any other qualification, in the whole or in part, at this or any other University or equivalent institution. The research methodology, implementation, and findings represent original con- tributions to the field of Khmer OCR technology, particularly in developing novel approaches for synthetic data generation and transformer-based text recognition. Through this work, I have demonstrated strong research capabilities and independence in addressing the critical challenges of Khmer text digitization and recognition.

No reference to, or quotation from, this document may be made without the written approval of the author.

Name of Candidate: Vitou Soy

Signed by the candidate:

Date: ....

Name of Supervisor: Mr. Sokchea Kor

Countersigned by the Supervisor: ....

Date: ....

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Mr. Sokchea Kor, for his guidance and expertise. His feedback and support throughout the research process have been instrumental in shaping this work. This research was inspired by Dr. Rina Buoy's contributions to the field. I appreciate the Royal University of Phnom Penh management for establishing this program within the Faculty of Engineering. I would also like to acknowledge khsearch.com, Chuon-Nath Dictionary, Alpha-Word, Google-Word, and Hugging Face for providing essential datasets.

I am profoundly thankful to the entire Faculty of Engineering community for their exceptional support and contributions to my academic journey. The knowledge, resources, and supportive environment they provided have been crucial to my success. The faculty's unwavering commitment to excellence and their dedication to nurturing future engineers have created an atmosphere that truly fosters innovation and learning. The state-of-the-art facilities and cutting-edge technology available have enabled me to conduct advanced research in optical character recognition with unprecedented precision. The collaborative spirit among faculty members, researchers, and students has fostered an environment of intellectual growth and continuous innovation. Through numerous workshops, seminars, and technical discussions, I have gained deep insights into the field of computer vision and machine learning. The faculty's strong industry connections and emphasis on practical applications have ensured that my research remains relevant and impactful. Their guidance in implementing transformer-based architectures and synthetic data generation techniques has been particularly valuable. The mentorship provided by senior researchers and the opportunity to participate in various research projects have significantly enhanced my technical capabilities and research methodology.

Finally, I would like to thank my friends for their encouragement throughout my studies. Their support and belief in my capabilities have helped make this journey meaningful.

TABLE OF CONTENTS

Preliminary Pages

Appendices

List of Tables

List of Figures

LIST OF ABBREVIATIONS

OCR: Optical Character Recognition

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

LSTM: Long Short-Term Memory

GRU: Gated Recurrent Unit

Transformer: Transformer Model

BERT: Bidirectional Encoder Representations from Transformers

TrOCR: Transformer OCR

ViT: Vision Transformer

ViT-OCR: ViT OCR

ViT-OCR-S: ViT OCR Small

ViT-OCR-B: ViT OCR Base

ViT-OCR-L: ViT OCR Large

ViT-OCR-H: ViT OCR Huge

Chapter 1

Introduction

This chapter presents the main components of this research on Khmer optical character recognition (OCR). It begins with background information on OCR technology and its importance for the Khmer language, followed by identifying the key challenges and research gaps in current Khmer OCR systems. The chapter then outlines the study's objectives and research questions focused on improving Khmer text recognition through synthetic data generation and deep learning approaches. The rationale highlights the significance of developing better OCR tools for preserving and digitizing Khmer texts. Finally, it describes the scope and limitations of the study, along with an overview of the thesis structure.

1.1 Background to the Study

Optical Character Recognition (OCR) technology has become increasingly important in Cambodia's digital transformation journey. As a nation with a rich literary and cultural heritage spanning over a millennium, Cambodia possesses countless historical documents, manuscripts, and texts written in the Khmer script. These materials include ancient palm leaf manuscripts, historical records, educational materials, and government documents that hold significant cultural and practical value.

The Khmer script, which has been in use since the 7th century, presents unique challenges for OCR systems due to its complex writing system. Unlike Latin-based scripts, Khmer is an abugida writing system with intricate character combinations, subscripts, diacritics, and contextual forms. Traditional OCR solutions, which were primarily developed for Latin-based scripts, often struggle with these complexities.

A particularly pressing challenge is the digitization of Khmer educational materials, especially textbooks from grade 1 to grade 12. Many of these essential learning resources exist only in physical form, with their original digital files lost or never created. This creates significant barriers for educators and students who need digital access to these materials for modern learning environments. The lack of digital versions makes it difficult to update, reproduce, or widely distribute these educational resources efficiently.

While some attempts have been made to develop Khmer OCR solutions, most existing systems have limited accuracy and struggle with real-world variations in text appearance, fonts, and document quality. The scarcity of large-scale training datasets for Khmer text recognition has further hampered progress in this field. This situation has created a pressing need for innovative approaches to improve Khmer OCR technology, especially for recovering and digitizing educational materials that are crucial for Cambodia's education system.

In recent years, there has been growing recognition of the need to digitize Khmer texts for preservation, accessibility, and practical applications. Libraries, museums, and educational institutions across Cambodia are increasingly seeking efficient ways to convert physical documents into searchable digital formats. However, the lack of robust Khmer OCR systems has been a significant bottleneck in these digitization efforts, particularly affecting the education sector where digital versions of textbooks are desperately needed.

Table 1.1: Current State of Khmer Textbook Digitization in Cambodia's Education System

Education Level	Subject Areas	Format Availability	Notes
Grade 1–6	All core subjects	Mostly physical only	Many original digital files missing
Grade 7–9	Math, Science, Khmer	Some digital scans	Scanned PDFs, not text-searchable
Grade 10–12	All major subjects	Few digitized	Hard to find editable versions

Beyond the education sector, the need for robust Khmer OCR technology extends to numerous other critical applications across different domains:

- **AI and Language Models:** Digitizing Khmer books and documents from libraries would enable training of large language models on Cambodian content, making AI systems more culturally aware and capable of processing Khmer language queries and knowledge.
- **Digital Libraries:** Converting physical books into searchable digital formats would dramatically improve access to knowledge, allowing readers to instantly search across thousands of Khmer texts and enabling advanced research capabilities.
- **Cultural Heritage Preservation:** Thousands of ancient palm leaf manuscripts and historical documents in temples and museums require digitization for preservation and scholarly access, while making this knowledge accessible to AI systems for cultural understanding.
- **Government Records:** Vast archives of administrative documents, legal records, and civil registries need conversion into machine-readable formats, enabling automated processing and AI-assisted analysis of public records.
- **Healthcare Systems:** Medical records and health documentation could be digitized to train specialized medical AI models that understand Khmer medical terminology and practices.
- **Business Intelligence:** Companies could extract insights from digitized Khmer business documents using AI analysis, while making their archives searchable and processable by modern business systems.
- **Media Archives:** Converting newspapers and magazines into machine-readable text would allow AI systems to analyze decades of cultural and historical information, identifying trends and patterns in Cambodia's social development.
- **Research and Academia:** Digitized academic papers and research materials could feed into knowledge bases for AI systems, making Cambodian research more accessible globally while enabling advanced cross-referencing and analysis.

These applications highlight how Khmer OCR technology could not only preserve and digitize texts, but also make them machine-readable for AI systems and language models. This would create a powerful feedback loop where better digitization enables smarter AI systems, which in turn can help process and analyze more Khmer content, ultimately making Cambodia's rich textual heritage more accessible and useful in the digital age.

1.2 Problem Statement

Optical Character Recognition (OCR) for the Khmer language presents a unique set of challenges that significantly hinder the development of accurate and robust recognition systems. Unlike Latin-based scripts, Khmer is an abugida writing system, where each character represents a consonant-vowel unit and includes complex combinations of base characters, subscripts, superscripts, and diacritics. This structural complexity introduces difficulties at both the text detection stage and the text recognition stage.

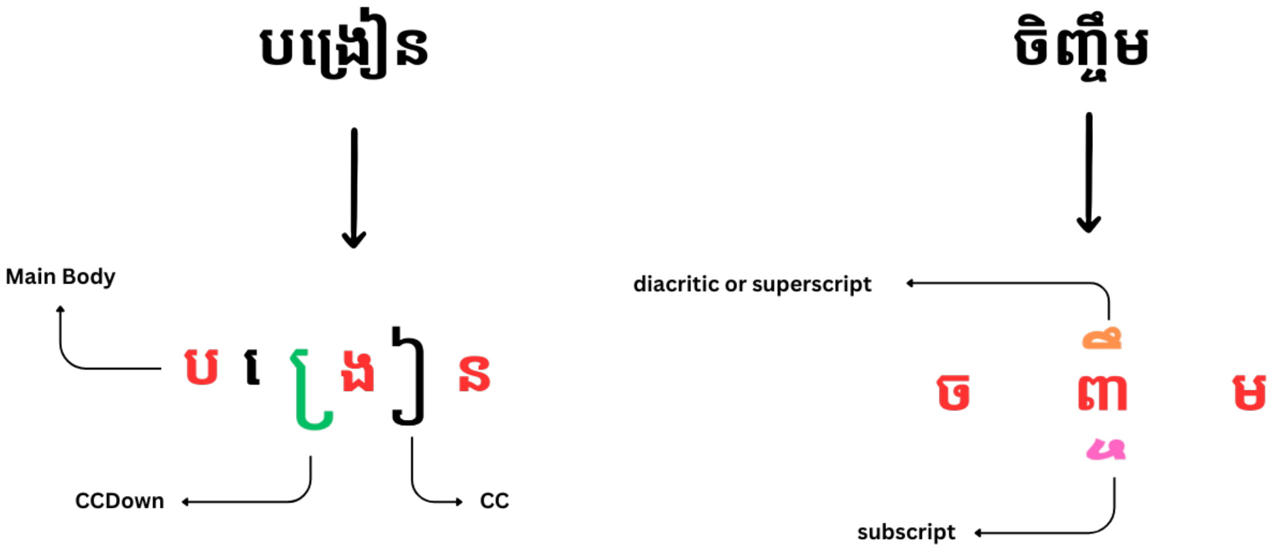

Figure 1.1: Example of Khmer text format showing the complexity of character combinations and diacritics

One of the fundamental obstacles is the lack of clear word boundaries in Khmer writing. In contrast to Latin-based languages, where spaces are consistently used to separate words, spaces in Khmer are used infrequently and inconsistently. This makes it extremely difficult to segment text accurately into word-level units for training sequence-to-sequence OCR models such as TrOCR. The absence of reliable word boundaries reduces recognition accuracy and complicates tasks like error correction, search indexing, and language modeling.

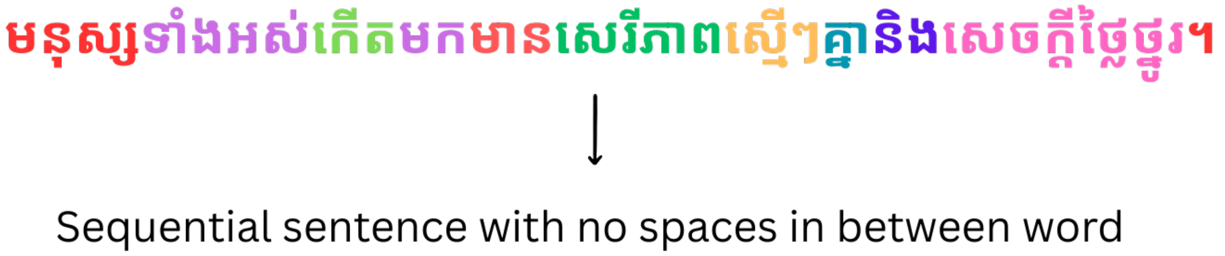
មនុស្សទាំងអស់កើតមកមានសេរីភាពស្មើៗគ្នានិងសេចក្ដីថ្លៃថ្នូរ។
Sequential sentence with no spaces in between word

Figure 1.2: Example of sequential Khmer text showing how characters combine to form syllables and words

A critical barrier to advancing Khmer OCR is the scarcity of annotated training datasets. There is a severe lack of high-quality, large-scale datasets that provide paired image-text data with bounding boxes, character-level annotations, or transcription lines tailored to Khmer script. This data scarcity limits the potential for supervised learning approaches and transfer learning, which are essential for training modern deep OCR models like TrOCR.

Additionally, font and style variability further degrade recognition performance. Khmer documents in the real world are printed in diverse typefaces and stylistic variations (e.g., Khmer

OS, Nokora, and Hanuman), with differences in stroke thickness, spacing, and decoration. The lack of standardization across documents and poor documentation of these fonts means that OCR models trained on one style often fail to generalize to others. This problem is exacerbated in noisy or low-resolution scans of textbooks and historical texts.

Taken together, these challenges create a significant barrier to digitizing Khmer documents using CRAFT + TrOCR pipelines. The absence of word delimiters, the visual complexity of character stacking, cross-script confusion, data scarcity, and font inconsistency all contribute to the low accuracy and poor reliability of existing OCR solutions for Khmer. Addressing these issues requires the development of customized preprocessing, augmentation, and model training strategies—as well as targeted data collection and annotation efforts—to make Khmer OCR viable for real-world applications, especially in the context of educational digitization and cultural preservation.

1.3 Aim and Objectives of the Study

The primary aim of this research is to develop an improved optical character recognition (OCR) system specifically designed for the Khmer language, addressing the unique challenges of Khmer script while achieving high accuracy and reliability in real-world applications.

The specific objectives of this study are:

1. To analyze and quantify the key challenges in Khmer OCR, including character stacking, absence of word boundaries, and font variations

2. To develop enhanced preprocessing techniques that better handle the complex visual structure of Khmer text, particularly focusing on character segmentation and diacritic preservation

3. To create and curate a comprehensive annotated dataset of Khmer text images suitable for training modern OCR models

4. To design and implement customized data augmentation strategies that account for real-world variations in Khmer text appearance

5. To adapt and optimize the CRAFT text detection and TrOCR recognition models for improved performance on Khmer script

6. To evaluate the developed system's performance across different document types, fonts, and quality levels

7. To establish best practices and guidelines for Khmer OCR system development and deployment

Through achieving these objectives, this research aims to significantly advance the state of Khmer OCR technology and enable more effective digitization of Cambodian textual heritage.

1.4 Research Questions

This research aims to address the following key questions:

1. How can text detection and recognition models be effectively adapted to handle the unique characteristics of Khmer script, particularly the stacking of characters and presence of diacritics?

2. What preprocessing and augmentation techniques are most effective for improving OCR accuracy on Khmer text documents with varying fonts, styles, and quality levels?

3. How can the lack of word boundaries in Khmer text be addressed to improve recognition accuracy and enable better post-processing?

4. What are the minimum dataset requirements and optimal annotation strategies for training robust Khmer OCR models?

5. How do different architectural modifications to CRAFT and TrOCR impact recognition performance on Khmer script?

6. What evaluation metrics and benchmarks should be established to meaningfully assess Khmer OCR system performance?

1.5 Rationale of the Study

This research is motivated by several compelling factors. First, there is an urgent need to digitize and preserve Cambodia's vast textual heritage, including historical documents, educational materials, and cultural artifacts. Without effective OCR technology for Khmer script, this digitization process remains labor-intensive and prone to errors.

Second, the current limitations of OCR systems for Khmer significantly hinder educational and academic initiatives in Cambodia. Many educational institutions struggle to convert physical textbooks and learning materials into digital formats, impacting accessibility and modernization efforts in education.

Third, the unique challenges posed by Khmer script—from character stacking to the absence of word boundaries—present an opportunity to advance the field of OCR technology as a whole. Solutions developed for Khmer may benefit other scripts with similar characteristics.

Finally, improving Khmer OCR technology aligns with broader digital transformation goals in Cambodia, supporting efforts to preserve cultural heritage while enabling more efficient information processing and accessibility in various sectors.

1.6 Limitations and Scope

While this research aims to advance Khmer OCR technology significantly, it is important to acknowledge certain limitations and define the scope of the study:

1. The research focuses specifically on printed Khmer text and does not address handwritten text recognition, which presents additional challenges requiring separate investigation.

2. The study primarily considers modern Khmer fonts and typography, with limited coverage of historical or decorative text styles.

3. While the system aims to handle various document quality levels, extremely degraded or damaged documents may fall outside the scope of reliable recognition.

4. The research concentrates on pure Khmer text and may not fully address documents containing mixed scripts or languages.

5. The study focuses on optical character recognition and does not extend to higher-level natural language processing tasks such as semantic analysis or machine translation.

6. Resource constraints may limit the size and diversity of the training dataset, though efforts will be made to ensure sufficient representation of common use cases.

These limitations help maintain a focused research scope while acknowledging areas that may require future investigation.

1.7 Structure of the Thesis

This thesis is organized into the following chapters:

1. **Introduction**: Presents the research background, objectives, research questions, rationale, and scope of the study.

2. **Literature Review**: Reviews existing OCR technologies, challenges in Khmer script recognition, and relevant deep learning approaches.

3. **Methodology**: Details the proposed approach, including dataset preparation, model architecture, and training procedures.

4. **Implementation**: Describes the technical implementation, including preprocessing techniques, model modifications, and system integration.

5. **Results and Analysis**: Presents experimental results, performance analysis, and comparative evaluation with existing solutions.

6. **Conclusion**: Summarizes key findings, contributions, and suggests directions for future research.

Each chapter builds upon the previous ones to present a comprehensive study of Khmer OCR development.

Chapter 2

Literature Review

2.1 Overview of Optical Character Recognition (OCR)

This section provides a comprehensive overview of Optical Character Recognition technology, including its fundamental principles, key components, and evolution over time. The discussion covers both traditional and modern approaches to OCR, establishing the foundation for understanding current challenges and opportunities.

2.2 Challenges in Khmer OCR

This section examines the specific challenges faced in developing OCR systems for the Khmer script. Khmer writing presents unique structural complexities such as stacked characters and inconsistent spacing, making segmentation and recognition highly challenging [1]. It analyzes the unique characteristics of Khmer writing, including its complex character combinations, subscripts, and diacritics, which pose significant challenges for accurate recognition.

2.3 Synthetic Data for Low-Resource Languages

The role of synthetic data generation in addressing the data scarcity problem for low-resource languages is explored here. This section reviews existing approaches to synthetic data creation and their effectiveness in training OCR systems.

2.4 OCR Datasets and Benchmarks

A comprehensive review of existing OCR datasets and evaluation benchmarks is presented, with particular focus on those relevant to Southeast Asian scripts and low-resource languages. The limitations of current datasets are also discussed.

2.5 Deep Learning Models for Text Recognition

2.5.1 CNN-based Methods

This subsection examines the application of Convolutional Neural Networks in OCR, including architectures specifically designed for text recognition tasks and their performance characteristics.

2.5.2 Transformer-based Architectures

The emergence and impact of transformer-based models in OCR are analyzed, with particular attention to recent developments and their advantages over traditional approaches.

2.6 Summary of Research Gaps

This section synthesizes the key findings from the literature review and identifies critical gaps in current research, particularly in the context of Khmer OCR and low-resource language processing. These gaps form the basis for the research objectives addressed in this thesis.

Chapter 3

Dataset Construction

3.1 Text Source Collection

This section describes the process of collecting Khmer text data from various sources to create a comprehensive dataset for OCR training and evaluation.

3.1.1 Khmer Websites and Dictionaries

We gathered text samples from popular Khmer news websites, online dictionaries, and digital libraries to ensure diverse vocabulary coverage and writing styles.

3.1.2 Online NLP Resources and Tools

Additional text data was collected using available Khmer NLP tools and resources, including pre-existing corpora and language processing utilities.

3.2 Text Cleaning and Preprocessing

Raw text data underwent several preprocessing steps to ensure quality and consistency for synthetic image generation.

3.2.1 Removal of Invalid Characters and Whitespace

We implemented filtering mechanisms to remove invalid Unicode characters, normalize whitespace, and handle special characters that could affect OCR performance.

3.2.2 Unicode Normalization

All text was normalized to ensure consistent Unicode representation of Khmer characters and their combinations.

3.3 Sentence Segmentation and Reconstruction

The cleaned text was processed to create meaningful sentence units suitable for OCR training.

3.3.1 Tokenization Using khmer-nltk

We utilized the khmer-nltk library to perform accurate tokenization of Khmer text while preserving linguistic properties.

3.3.2 Sentence Length Variation

Sentences were segmented and reconstructed to create samples with varying lengths, ensuring the dataset represents real-world text diversity.

3.4 Image Generation Pipeline

A robust pipeline was developed to convert processed text into synthetic training images.

3.4.1 Font and Background Selection

Multiple Khmer fonts and background variations were incorporated to create diverse and realistic training samples.

3.4.2 Noise Injection Techniques

Various types of noise and distortions were systematically added to simulate real-world document conditions.

3.4.3 Image Rotation and Margin Augmentation

Geometric transformations and margin variations were applied to improve model robustness to different text orientations and layouts.

3.5 Dataset Statistics and Format

This section presents detailed statistics about the generated dataset, including size, character distribution, and format specifications.

3.6 Comparison with Existing Datasets

A comparative analysis of our dataset with existing Khmer OCR datasets, highlighting improvements and unique characteristics.

Chapter 4

Experiments

4.1 Experimental Environment and Tools

This section describes the hardware, software, and tools used to conduct the OCR experiments.

4.2 Model Architecture and Configuration

Details of the model architectures and configurations used in the OCR system.

4.2.1 CRAFT for Text Detection

Description of the CRAFT model architecture and its configuration for Khmer text detection.

4.2.2 TrOCR for Text Recognition

Details of the TrOCR transformer-based model used for Khmer text recognition.

4.3 Training Methodology

The training approach and methodology used to develop the OCR models.

4.3.1 Fine-tuning CRAFT on Annotated Images

Process and parameters for fine-tuning the CRAFT model on annotated Khmer document images.

4.3.2 Fine-tuning TrOCR on Synthetic Dataset

Details of fine-tuning TrOCR using the synthetic Khmer dataset.

4.4 Evaluation Metrics

Metrics used to evaluate the performance of the OCR system.

4.4.1 Detection Metrics (Precision, Recall)

Text detection evaluation using precision and recall metrics.

4.4.2 Recognition Metrics (Accuracy, CER, WER)

Text recognition evaluation using character error rate, word error rate, and accuracy metrics.

4.5 Baseline and Benchmark Comparison

Comparison of our system's performance against existing baselines and benchmarks.

Chapter 5

Results and Analysis

5.1 Text Detection Results

5.1.1 Quantitative Metrics

This subsection presents the quantitative evaluation of the text detection model, including precision, recall, and F1-score metrics across different test scenarios.

5.1.2 Qualitative Visual Examples

Visual examples demonstrating the text detection performance on various document types, highlighting both successful cases and challenging scenarios.

5.2 Text Recognition Results

5.2.1 Accuracy on Character, Word, Sentence Levels

Detailed analysis of recognition accuracy at different granularities - character level, word level, and complete sentence level.

5.2.2 Performance Across Sentence Lengths

Analysis of how recognition performance varies with different sentence lengths and complexity levels.

5.3 Khmer vs. English Performance

Comparative analysis of the system's performance on Khmer text versus English text, highlighting key differences and challenges.

5.4 Error Analysis and Failure Cases

Systematic analysis of common error patterns and challenging cases that lead to recognition failures.

5.5 System Robustness and Generalization

Evaluation of the system's ability to generalize across different document types, fonts, and image quality conditions.

Chapter 6

Discussion

6.1 Effectiveness of Synthetic Data

This section analyzes the impact and effectiveness of using synthetic data for training the OCR system, including benefits and limitations observed.

6.2 Strengths and Limitations of the OCR System

A critical examination of the system's capabilities and areas for improvement, based on experimental results.

6.3 Research Challenges and Lessons Learned

Discussion of key technical and methodological challenges encountered during the research, and important lessons learned.

6.4 Comparison with Related Works

Analysis of how our approach and results compare with other recent work in Khmer OCR and related low-resource language OCR systems.

6.5 Impact on Khmer NLP and OCR Research

Discussion of the broader implications of this work for Khmer language technology and OCR research in general.

Chapter 7

Conclusion and Future Work

7.1 Summary of Contributions

This section summarizes the key contributions of this research to Khmer OCR and language technology.

7.2 Key Findings

A synthesis of the main experimental results and insights gained through this research.

7.3 Limitations

Discussion of current limitations and constraints of the developed OCR system and methodology.

7.4 Future Research Directions

Exploration of potential future work and research opportunities building on this foundation.

7.5 Final Remarks

Concluding thoughts on the significance and implications of this research for Khmer language technology.

Chapter 8

Practical Applications

8.1 Use in Document Digitization

This section explores how the developed OCR system can be applied to digitize Khmer documents, including historical texts, government records, and business documents.

8.2 OCR for Education and Cultural Preservation

Discussion of applications in educational settings and the role of OCR in preserving Khmer cultural heritage through digital archiving.

8.3 Deployment Considerations

Analysis of technical and practical considerations for deploying the OCR system in real-world environments, including scalability and performance requirements.

8.4 Opportunities for Government and Enterprise Use

Examination of potential applications in government agencies and private enterprises, including workflow automation and document management systems.

Bibliography

[1] Muaz, Ahmed and LengIeng, Ing. *Khmer Optical Character Recognition (OCR)*. 2015. DOI: 10.13140/RG.2.1.2393.3926.

Appendices

Appendix A: Sample Annotated Images

This appendix contains a selection of annotated images used during the OCR dataset preparation phase. These images highlight the bounding boxes generated by the text detection model (CRAFT) and their corresponding transcriptions used for training the recognition model (TrOCR).

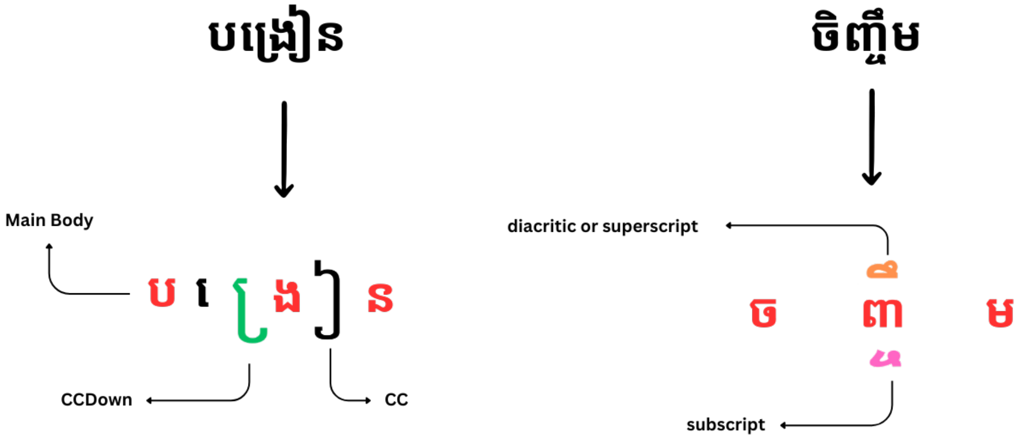

Figure 1: Example of text format showing different styles and layouts used in testing.

Appendix B: List of Fonts Used

This appendix lists the Khmer and Latin fonts used during synthetic data generation and model evaluation. Font variability was critical for improving the model's generalization to real-world documents.

Appendix C: Code Snippets and Training Configuration

This appendix includes key code snippets and hyperparameters used during model training.

Example TrOCR Training Configuration

```
# Sample training configuration
model_args = {
    "model_name": "microsoft/trocr-base-stage1",
    "learning_rate": 5e-5,
    "warmup_steps": 500,
    "max_steps": 10000,
    "batch_size": 16,
    "max_length": 256
}

trainer = Trainer(
    model=model,
    args=TrainingArguments(**model_args),
    train_dataset=train_dataset,
    eval_dataset=val_dataset
)
```

Example CRAFT Detection Parameters

- Text confidence threshold: 0.7
- Link confidence threshold: 0.4
- Input resolution: 1280x720
- Post-processing NMS threshold: 0.2

Appendix D: Additional Evaluation Examples

This appendix includes additional OCR results to showcase the model's behavior on varied layouts, font types, and Khmer-English mixed inputs.

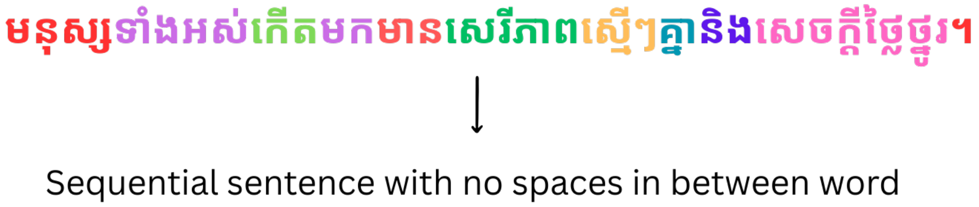
មនុស្សទាំងអស់កើតមកមានសេរីភាពស្មើៗគ្នានិងសេចក្ដីថ្លៃថ្នូរ។
Sequential sentence with no spaces in between word

Figure 2: Example showing sequential text processing and recognition capabilities.