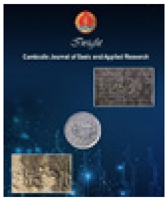
The Cambodia Journal of Basic and Applied Research 6(2) 2024

Print ISSN: 2790-3508 Online ISSN: 2790-3516

The Cambodia Journal of Basic and Applied Research

journal homepage: https://cjbar.rupp.edu.kh



Accuracy Improvement of Khmer Text Recognition by Correcting Post-recognized Characters

SRUN Sovila[1*], KEAN Tak[2], BUN Leap[3]

[1]Faculty of Engineering, Royal University of Phnom Penh (RUPP), Russian Federation Boulevard, Khan Toul Kork, Phnom Penh, Cambodia.
[2]Rector's Office, Royal University of Phnom Penh (RUPP), Russian Federation Boulevard, Khan Toul Kork, Phnom Penh, Cambodia.
[3]Faculty of Engineering, Royal University of Phnom Penh (RUPP), Russian Federation Boulevard, Khan Toul Kork, Phnom Penh, Cambodia.

ARTICLE INFO
Editorial responsibility: Prof. SOK Serey
Received: June 21, 2023
1st Revised: January 25, 2024
2nd Revised: September 23, 2024
Accepted: October 30 2024
Published online: December 30, 2024

Keywords:
Khmer OCR,
OCR accuracy,
Word misspelling,
Post-recognition.

សារគន្លឹះ:

- រដ្ឋធម្មនុញ្ញនៃព្រះរាជាណាចក្រកម្ពុជាបានកំណត់យកភាសាខ្មែរជាភាសាផ្លូវការ។ ហេតុនេះ ដំណើរការនៃការរៀបចំឯកសារជាភាសាខ្មែរក្នុងទម្រង់ឌីជីថលប្រកបដោយភាពត្រឹមត្រូវមានសារៈសំខាន់ណាស់សម្រាប់ការអភិវឌ្ឍជាតិ។ នាពេលបច្ចុប្បន្ន វិស័យសាធារណៈនិងឯកជនកំពុងជួបការលំបាកក្នុងការរៀបចំឯកសារជាភាសាខ្មែរនៅក្នុងទម្រង់ឌីជីថល ដោយសារបច្ចេកវិទ្យាស្គាល់តួអក្សរខ្មែរនៅមានកម្រិតនៅឡើយ។
- កម្មវិធីស្គាល់តួអក្សរខ្មែរ ដែលមានស្រាប់ បានបង្ហាញនូវឧបសគ្គជាច្រើនជាមួយតួអក្សរខ្មែរ ដោយមានកំហុសក្នុងការស្គាល់តួអក្សរខ្មែរ ដែលជាហេតុនាំឱ្យមានផលប៉ះពាល់ដល់ប្រសិទ្ធភាពនៃការរៀបចំឯកសារជាទម្រង់ឌីជីថល។ វិធីសាស្ត្រកែតម្រូវបន្ទាប់ពីដំណើរការទទួលស្គាល់តួអក្សរ បានធ្វើឱ្យភាពត្រឹមត្រូវនៃការធ្វើឌីជីថលអត្ថបទជាភាសាខ្មែរកាន់តែមានកម្រិតខ្ពស់រហូតដល់ ៩៣,៤% - ៩៦,៤%ដែលគួរឱ្យកត់សម្គាល់។
- ដំណោះស្រាយ ដែលបានរៀបចំឡើងនេះ អាចបញ្ចូលទៅក្នុងប្រព័ន្ធគ្រប់គ្រងឯកសារដែលមានស្រាប់ ដោយមិនចាំបាច់តម្រូវឱ្យមានការផ្លាស់ប្ដូរហេដ្ឋារចនាសម្ព័ន្ធច្រើនពេកនោះទេ។ ស្ថាប័នរដ្ឋ និងឯកជនអាចបញ្ចូលឯកសារជាភាសាខ្មែរក្នុងប្រព័ន្ធឌីជីថលបានប្រកបដោយប្រសិទ្ធភាពខ្ពស់ជាងមុន។
- រដ្ឋាភិបាលគួរផ្ដល់អាទិភាពដល់ការប្រើប្រាស់កម្មវិធីស្គាល់តួអក្សរដែលមានភាពប្រសើរជាងមុននេះ ដើម្បីជាដំណោះស្រាយនិងលើកស្ទួយដល់ការផ្ដល់សេវាសាធារណៈដែលប្រើប្រាស់ភាសាខ្មែរ។ ការវិនិយោគលើការអភិវឌ្ឍកម្មវិធីឌីជីថលភាសាខ្មែរនឹងគាំទ្រដល់គោលដៅកំណែទម្រង់ឌីជីថលរបស់កម្ពុជាក្នុងគ្រាដែលរដ្ឋាភិបាលបាននិងកំពុងខិតខំថែរក្សាបេតិកភណ្ឌភាសាជាតិរបស់ខ្លួន៕

*Corresponding author: Faculty of Engineering, Royal University of Phnom Penh (RUPP), Russian Federation Boulevard, Khan Toul Kork, Phnom Penh, Cambodia.
E-mail addresses: srun.sovila@rupp.edu.kh (SRUN Sovila)
DOI: https://doi.org/10.61945/cjbar.2024.6.2.05

Key Messages

- The Constitution of Cambodia establishes Khmer as the official language, making accurate digital processing of Khmer documents crucial for national development. Limitations hinder current digital transformation efforts in both public and private sectors in Khmer text recognition technology.
- Existing Optical Character Recognition (OCR) tools show significant limitations with Khmer script, with common character recognition errors affecting document processing efficiency. Our post-processing correction method improves Khmer OCR accuracy from 93.4 to 96.4%, representing a significant advancement in Khmer text digitization.
- The proposed solution can be integrated into existing document management systems without requiring extensive infrastructure changes. Government agencies and private organizations can achieve higher efficiency in document digitization while maintaining Khmer language integrity.
- Government institutions should prioritize the adoption of improved Khmer OCR systems to enhance public service delivery. Investment in Khmer language digital tools will support Cambodia's digital transformation goals while preserving its linguistic heritage.

Background

The conversion of an image of handwritten or typewritten text into machine-encoded text, which is called optical character recognition (OCR) (Cheriet et al., 2007), is widely used in fields of office automation and information retrieval such as capturing data from invoices, bank statements, health record papers, and legacy documents (Singh et al., 2012). Unlike English, Japanese, Chinese, and others, the OCR for the Khmer language is still limited. Neither the availability is so wide, nor the accuracy is so high. It is because of the complexities of the Khmer writing system and the lack of research related to the Khmer OCR. There is no word delimiter in Khmer text, and some characters are written cursively. Furthermore, in one line, there can be five levels of character (1 baseline level, two superscript levels, and two subscript levels). There have been some researches so far focused on the Khmer OCR. The researchers have improved some parts to make the whole Khmer OCR better. Sok (2014) describes the use of a support vector machine (SVM) based classification method on Khmer Printed Character-set Recognition (PCR) in a bitmap document. The paper proposed one new method, SVM, for the Khmer character classification system by using 3 different SVM kernels (Gaussian, Polynomial and Linear Kernel) on data training and recognition to find out the best kernel for the Khmer language. Lengleng & Ahmed (2015) presented the complete OCR system for the Khmer language by demonstrating four main processes of the system such as pre-processing, segmentation, recognition, and mapping. However, the accuracy of an OCR system could also be increased after all of these stages are finished, which is called the post-recognition stage, by correcting the wrong characters in the output text. In this paper, we proposed a method to correct the wrong Khmer characters after being recognized by an OCR system based on the whole words, and we also conducted an experiment to measure the accuracy.

Literature Review

Several recent works on the pre-processing and recognition stages of the Khmer OCR were carried out. A multi-feature extraction method was proposed by Vanna & Wataru (2011), which used Scale Invariant feature transform and fourier transform descriptors. The experiment of this method with 1,104 words taken from 2 newspaper documents showed that the accuracy as of 97.4%. An approach to line segmentation for the Khmer OCR was conducted by Sok et al. (2012). The author proposed a method to segment printed text document images into text lines based on the topological assumption that for each text line, there is a path or white gap from left to right, which separates those text lines. The proposed method achieved 95 and 96% of line segmentation for Khmer typewritten and handwritten documents, respectively. In the same year, the use of edge detection and template matching for Khmer printed character segmentation and recognition was proposed by Setha (2012). With this method, the author used only one font named Khmer OS Content with a specific size in the experiment.

The result of the work was around 98% of the accuracy. Since the Khmer language does not have a word delimiter, word segmentation is also one of the challenges. Bi & Taing (2014) proposed a method for Khmer word segmentation based on bi-directional maximal matching. The result of this research was 98.13% accuracy for 1,110,809 Khmer characters, which was about 160,000 words. In the post-recognition stage of the OCR, many researchers focused on error correction, but none focused on the Khmer OCR. Bassil & Alwani (2012) proposed a new post-recognition method for OCR error correction based on the "Did you mean" spelling suggestion feature of Google's online search web engine. The algorithm starts by chopping the OCR text into several tokens of words and sends it as a search query to Google's search engine. In case the query contains a

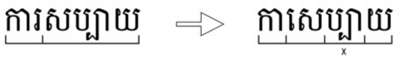


Figure 1: Irregular order problem

misspelled word, Google will suggest a possible correction via its "Did you mean" feature. The experiment revealed an error detection and correction improvement of 690% for English text and 403% for Arabic text. In other words, 6.9 times more English errors and 4 times more Arabic errors were detected and corrected. Thus, most of the researchers stated above tried to increase the accuracy of an OCR only in the pre-processing and recognition stage. Since actually it can also be increased in the post-processing stage, the paper proposes a method to increase the accuracy of the Khmer OCR after it was recognized.

Methods

This paper proposes a method to correct the wrong recognized Khmer characters. Primarily, the method consists of 2 main modules: error detection and error correction.

Error Detection

In the Khmer language, it is a major challenge to detect error words inside the text because of the complexities of the writing system. The text is written continuously without word delimiter, and some characters stand in irregular order positions. It means that the way of writing is in contrast with the way of spelling.

This rule can cause problems when this kind of character is recognized wrongly. The problem is that it does not only cause its own word error but also causes a neighbor's error.

In Figure 1, the consonant "រ" belongs to the first word. But when it was recognized wrongly to the vowel "េ", it was moved to the second word according to the special ordering rule of Khmer characters. So, it causes the second word error.

In another case, a Khmer word consists of one or more consonantal clusters (Hok, 2005). When a character of a cluster is recognized wrongly, it can break one original word into multiple words.

In Figure 2, the consonant "ម" was recognized wrongly as "ប". This breaks the original word into 3 words, where the middle one is considered an error.

Therefore, it is a bad way to detect only a single error word and try to correct it. In this method, we will detect both a single error word and its appropriate neighbors together as we call it a suspect error area (SEA). The module is divided into three steps: word segmentation, SEA detection, and SEA minimization.

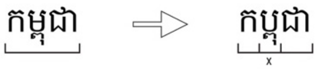


Figure 2: Multi-cluster word error

Table 1: Word indicated numbers

Indicated No	*Description*	*Example*
-1	Error word	រដ្ឋ
0	No previous and next word	កម្ពុជរដ្ឋ
1	Has next word	រដ្ឋា => រដ្ឋាភិបាល
2	Has previous work	រដ្ឋសភា => វិមានរដ្ឋសភា
3	Has previous and next word	រដ្ឋ => ពលរដ្ឋខ្មែរ

Word Segmentation

We will use the algorithm proposed by Bi & Taing (2014) using maximal matching, but we will modify some indicated numbers of each word as shown in Table 1.

SEA Detection

We will detect a suspect error area by checking neighbors of an error word. To the left side, neighbors are included in SEA if they have the next word (indicated number 1) or both the previous and next word (indicated number 3). To the right side, neighbors are included if they have the previous work (indicated number 2) or both the previous and next word (indicated number 3).

Let's say sentence S consists of a list of words w:

$$S = w_0 \ldots w_k \ldots w_n$$

Where w_k is an error word (indicated number -1), then the suspect error area SEA is:

$$SEA = w_i \ldots w_k \ldots w_m$$

Where:
- $i \geq 0$ and $m \leq n$
- w_i has indicated number 1, or 3 for i = 0
- $w_{i+1}, \ldots, w_{k-1}$ has indicated number -1 or 3
- $w_{k+1}, \ldots, w_{m-1}$ has indicated number -1 or 3
- w_m has indicated number 2, or 3 for i = n.

SEA Minimization

Since the error words from an OCR is a kind of machine error, most of the same error words would come from the same original words. Therefore, after SEAs are generated, we can minimize them to perform faster correction.

For suspect error area SEA and SEA'

$$SEA = w_0 \ldots w_n$$
$$SEA' = w'_i \ldots w'_k$$

where i ≥ 0 and k ≤ n

If - $w'_i \ldots w'_k = w_i \ldots w_k$

- $w_{0}, \ldots, w_{i-1}$ and $w_{k+1}, \ldots, w_n$ are all correct word (indicated number > -1).

Then we could minimize SEA to:

$$SEA = w_i \ldots w_k$$

Error Correction

In this module, we propose a semiautomatic method to correct the errors by generating suggested words of each error word, and asking user to choose the correct one. The method is divided into 3 steps: generating a candidate word, generating suggested words, and selecting a correct word.

Generating a Candidate Word

Since one SEA can consist of one or more words, and one word can consist of one or more clusters, we need to generate it into a single candidate word to find its suggestions. In (Bi & Taing, 2014), a candidate segmentation was generated by combining accumulated clusters and follow the maximal matching rule. Similarly, our algorithm also uses this concept to build a candidate word.

Generating suggested words

Damerau (1964) stated that 80% of all spelling errors are the result of four rules: (1) transposition of two letters; (2) one letter extra; (3) one letter missing; and (4) one letter wrong. With these kinds of errors, (Peterson, 1980) addressed the basic algorithm by (1) interchanging one character to another; (2) deleting one character; (3) inserting one character; and (4) substituting one character.

However, since the error words are the output from OCR, we can focus only on the wrong character rule. The algorithm is that, from a candidate word, one candidate character is selected to be substituted with others which have similar font shape defined in a character similarity matrix (Figure 3). Then if the word exists in the dictionary, it is considered as one suggestion. The order of the suggested words depends on the order of the replaced characters from the matrix.

ក	គ	ត	ភ
ខ	ឧ	ឩ	
គ	ត	ក	
ឃ	ឈ		
.			
.			
.			
្ក	្គ	្ត	្ភ

Figure 3: Character similarity matrix

The character similarity matrix, in Figure 3, stores all single Khmer characters with its correspond similar ones. The first column lists all Khmer candidate characters sorted by how often OCR recognizes them wrongly. The next columns list characters which have a similar font shape to the candidate character in the first column. The matrix is initially defined by human looking on how one character looks similar to others. Therefore, it would not be so fit with specific OCR system. For example, one OCR system often recognizes wrongly character “ក” as “គ” while another recognizes “ក” as “ត” the most.

To improve this, the algorithm will keep the matrix updated continuously by investigating when user selects one of the suggestions.

Table 2: Error words from the OCR

Type of error	*1-character error*	*2-character error*	*Total*
Transposition error	0	18	18
Extra-character error	5	0	5
Missing-character error	15	0	15
Wrong-character error	97	0	97
Other error	31		31
Total	117	18	166

Table 3: Error words after corrected

Type of error	*1-character error*	*2-character error*	*Total*
Transposition error	0	18	18
Extra-character error	5	0	5
Missing-character error	15	0	15
Wrong-character error	20	0	20
Other error	31		31
Total	40	18	89

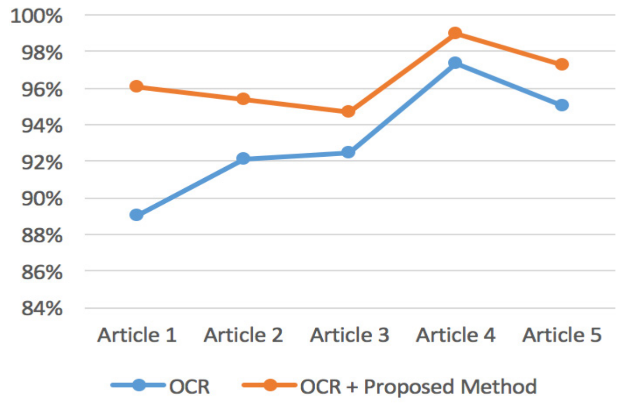


Figure 4: OCR accuracies comparison

Selecting a word from suggestions

Since the algorithm is semiautomatic, the user choose a word from the suggestions. However, some error words may not be found for their suggestions. In this case, the algorithm will ask the user to input the correct word, and this word will be stored for finding error and generating suggestions in the next correcting tasks.

Finding and Results

Experiments are conducted on 5 articles from 5 online newspapers which totally contain 2,496 words and 10,604 characters. We use a Tesseract OCR system proposed by Bunthom (2015) to recognize. Table 2 shows the statistics of error words produced by the OCR system. There is total 166 error words where most of the errors caused by OCR recognized wrong characters (58.43%), and each error word caused by only one wrong character.

With this statistic, it is suitable to apply our method to correct the errors. By using our method, we got the result as shown in Table 3. For the wrong-character error words, the proposed method corrected 77 among 97 equals to 79.38%. This result does not include words which manually input by user.

This correction improved the accuracy of the selected OCR system as shown in Figure 4. As an average, the OCR accuracy was improved from 93.35% to 96.43%.

Conclusion and Policy Implication

This paper presented a new post-processing method to detect and correct the wrong recognized Khmer characters from an OCR system. The method starts by defining suspect error areas and corrects those using character similarity matrix. Since the most error words from an OCR system caused by wrong recognized characters, the algorithm is suitable to improve the accuracy. However, from the experiments, we got only 79.38% for correcting the wrong-character error words. The problem is that, while recognizing, some combined characters were recognized into only one character and vice versa. Therefore, the proposed method cannot find suggested words since the character similarity matrix stores only single characters not combined ones. As further research, the proposed algorithm can be improved by extending the character similarity matrix to support one-to-many character similarity and vice versa. The expected OCR accuracy would be higher since most of the wrong-character error words will be corrected.

Acknowledgement

First of all, we would like to express our sincere gratitude to Southern Federal University, Taganrog Institute of Technology for providing the research environment and valuable support throughout this study. We are also deeply grateful to the Royal University of Phnom Penh (RUPP) for their institutional support and necessary facilities that contributed to the completion of this research.

Declaration of Competing Interest

The authors have no competing interests to declare. This research was conducted independently without any commercial, financial, or personal relationships that could have influenced the work presented in this paper. No funding was received from any organization that could have influenced the research outcomes. The research methodology, data analysis, and results presented in this manuscript reflect the authors' independent academic work. All authors have read and approved the final, published version of the manuscript. The authors confirm that there are no known conflicts of interest associated with this publication.

Credit Authorship Contribution Statement

KEAN Tak: Conceptualization, research design, data collection and analysis, writing - original draft of article, reviewing and editing. BUN Leap: Data analysis, methodology, reviewing and editing. SRUN Sovila: Supervision, methodology, reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

Raw data were collected at the Royal University of Phnom Penh (RUPP) during the study period. The datasets used and analyzed in this study are available from the corresponding author upon reasonable request.

Funding Declaration

No funding was received to assist with the preparation of this manuscript.

Author's Biography

SRUN Sovila received his Bachelor's degree in Information Science and Computer Engineering from Taganrog State University of Radio Engineering (TSURE), Russian Federation in 2005, followed by an Engineering degree in Computers and Automated Systems Software from TSURE in 2006. He earned his PhD in Basic Theory of Informatics from Taganrog Institute of Technology, Southern Federal University, Russian Federation in 2010. Currently, Dr. Sovila serves as the Director of the National Incubation Center of Cambodia (NICC) and Head of Information Technology Engineering Department at the Faculty of Engineering, Royal University of Phnom Penh (RUPP). He is also the coordinator of Master's and PhD programs in Information Technology Engineering.

KEAN Tak is currently a Ph.D. candidate at King Mongkut's

University of Technology Thonburi (KMUTT), Thailand. He received his Master of Science in Computer Science (2006) and Bachelor of Science in Mathematics (1998) from the Royal University of Phnom Penh (RUPP), Cambodia. He currently serves as Vice-Rector of RUPP, appointed by the Prime Minister of Cambodia in 2020, where he oversees ICT and Digital Technology, Industrial Linkages, Student Affairs, and Youth Development. His previous roles include Vice-Dean of the Faculty of Engineering (2016–2020) and Head of the Department of IT Engineering (2013–2016). He has extensive experience in project management, leading various international projects including the Higher Education Improvement Project, ICT Infrastructure Development Project, and multiple Erasmus+ Capacity Building initiatives. His professional development includes specialized training in ICT integration and institutional governance from Stockholm University, Sweden (2014–2015) and Osnabrück University, Germany (2018/2019).

BUN Leap received his Master's degree in Information Technology Engineering from the Royal University of Phnom Penh (RUPP), Cambodia. Currently, he is a lecturer at the Department of Information Technology Engineering, Faculty of Engineering, Royal University of Phnom Penh (RUPP).

References

Bassil, Y., & Alwani, M. (2012). Ocr post-processing error correction algorithm using google online spelling suggestion.

Bi, N., & Taing, N. (2014). Khmer word segmentation based on bi-directional maximal matching for plaintext and microsoft word document. In Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific (pp. 1-9). IEEE.

Bunthom T. (2015). A Study on the Effectiveness of Khmer Touching and Non-Touching Printing Character Recognition using Tesseract OCR Engine. Master Research Report, Royal University of Phnom Penh.

Cheriet, M., Kharma, N., Liu, C. L., & Suen, C. Y. (2007). Introduction: Character Recognition. Evolution And Development, Character Recognition Systems: A Guide For Students And Practitioner, 1-4.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3), 171-176.

Hok, P. (2005). Khmer Spell Checker (Doctoral dissertation, MS thesis, Australian National University, Canberra, Australia.

Lengleng, A. M. (2015). Khmer Optical Character Recognition (OCR).

Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors. Communications of the ACM, 23(12), 676-687.

Setha I. (2012). The Use of Edge Detection and Template Matching for Khmer Printed Character Segmentation and Recognition. Master Thesis, Royal University of Phnom Penh.

Shinde, A. A., & Chougule, D. G. (2012). Text pre-processing and text segmentation for OCR. International Journal of Computer Science Engineering and Technology, 2(1), 810-812.

Singh, A., Bacchuwar, K., & Bhasin, A. (2012). A survey of OCR applications. International Journal of Machine Learning and Computing, 2(3), 314.

Sok, D., Srun, S., Heng, P., Rim, B., Saovorak, K., & Phal, D. (2012). Line Segmentation for Khmer Printing Text Document.

Sok, P., & Taing, N. (2014, December). Support vector machine (SVM) based classifier for khmer printed character-set recognition. In Signal and information processing association annual summit and conference.

Vanna, K., & Wataru, K. (2011). A Proposed Multi-Feature Extraction Method for Khmer OCR,. Graduate School of Global Information and Telecommunication Studies, WASEDA University.