Segmentation-free Traditional Mongolian OCR and its Public Dataset

Yunpeng Bai, Weiqi Wang, Hui Zhang[+],
Feilong Bao, Hongxi Wei, Guanglai Gao

National & Local Joint Engineering Research Center of

Intelligent Information Processing Technology for Mongolian

Inner Mongolia Key Laboratory of Mongolian Information Processing Technology

College of Computer Science, Inner Mongolia University

Hohhot, Inner Mongolia

Keywords*:* Mongolian; Optical Character Recognition; Segmentation-free; Dataset

Abstract

The cursive nature of traditional Mongolian script poses a challenge for accurate character segmentation, and the presence of internal spaces within words can result in incorrect word segmentation. Consequently, an segmentation-free approach to Mongolian Optical Character Recognition (OCR) is preferred as it recognizes the entire text line, overcoming the complexities associated with both character and word segmentation. In this study, we introduce an OCR model for recognizing whole-line Mongolian text. To facilitate the development and evaluation of this model, we have compiled a database encompassing 80,000 lines of Mongolian text and 10,000 scanned images of real-world Mongolian text lines, carefully annotated with their corresponding text content. Additionally, we leverage data synthesis and augmentation techniques to augment the dataset. Furthermore, we make this dataset publicly available, alongside the text utilized for generating synthetic images, data generation code, and OCR model training configurations, solely for research purposes.

[+]Corresponding author: Hui Zhang (Email: cszh@imu.edu.cn; ORCID: 0009-0001-5531-7115).

1. Introduction

The written language system, one of the most significant inventions in human history, stands as an indispensable tool for communication and documentation. Countless valuable documents have been printed on paper and subsequently scanned into digital images. However, the digitization of these documents poses a challenge, as they need to be converted into a format that computers can understand and manipulate. This is where Optical Character Recognition (OCR) technology comes into play. OCR converts scanned images of text into editable and searchable text formats, unlocking the vast semantic information contained within printed documents. Therefore, for approximately 150 years, from its inception to the present day, OCR or text recognition has continually attracted a significant number of researchers, fueling progress and innovations in this field [1].

The earliest OCR technology relied on pattern matching, involving a pixel-by-pixel comparison of the text image with pre-stored letterforms. While this approach worked well for text printed in stored fonts, its performance suffered when confronted with text in unmatched or unknown fonts. To address the challenges posed by variations in font style, size, color, and scanning quality, more robust pattern recognition methods were needed. These later systems extracted basic structural features, such as lines, loops, line directions, and intersections, to identify characters. This approach was more flexible and could handle a wider range of text variations compared to the earlier pixel-based methods. Modern OCR systems now leverage neural networks, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to automatically learn and extract increasingly complex and abstract features directly from raw text images. This eliminates the need for manual feature engineering and significantly improves the system's ability to handle diverse text variations, fonts, sizes, distortions, and scanning artifacts.

Neural networks-based OCR systems are data-driven, relying heavily on large-scale datasets to train accurate text recognizers. These datasets typically consist of paired examples: text images and their corresponding transcribed text, with the transcribed text serving as the ground truth to provide correct labels for the text images. Constructing such large-scale datasets is a tedious and costly task that involves multiple steps, including data collection, preprocessing, segmentation, annotation, and verification. Recent efforts have successfully led to the creation of large-scale datasets for major languages like Chinese and English. The availability of these datasets has significantly contributed to the advancement of OCR systems [2]-[5].

Minority language users often face challenges in accessing high-quality OCR services comparable to those available for mainstream languages. Many minority languages face a challenge in data collection due to their limited user base and market demand. As a result, only a handful of institutions have access to usable datasets, and an even smaller number make these datasets publicly available. Taking Mongolian as an example, the current OCR systems available are primarily designed for Cyrillic Mongolian [6], which is a derivative of Russian OCR development. This is because Cyrillic Mongolian shares most characters with the Russian alphabet, with only two additional characters. However, these systems are not optimized specifically for Cyrillic Mongolian, leading to often unsatisfactory recognition results. Traditional Mongolian script, on the other hand, carries deep historical and cultural significance for the Mongolian people. Dating back to the 13th century during the reign of Genghis Khan, this script has been the primary form of writing used for centuries in Mongolian literature, historical records, and religious texts. The Mongolian script is deeply intertwined with the cultural identity of the Mongolian people, preserving their history, literature, and spiritual heritage. In modern-day Mongolia, the script is still taught in schools and used in some official documents, though its usage has been significantly reduced due to the adoption of Cyrillic script during the 20th century. According to recent estimates, while the majority of Mongolians now use the Cyrillic alphabet, there is still a significant effort to revive and promote the traditional script, particularly for cultural and educational purposes. The importance of digitizing traditional Mongolian script cannot be overstated, as it plays a crucial role in preserving the cultural heritage of the Mongolian people. This project aims to bridge the gap between traditional Mongolian text and modern technology, enabling wider access to Mongolian historical and cultural knowledge. Classical or traditional Mongolian, despite their relatively large user base and abundance of textual material that needs to be recognized, face a significant challenge in the field of OCR. Even

though OCR systems may exist for traditional Mongolian, there is a notable lack of publicly available large-scale datasets suitable for training purposes.

To promote the development of Mongolian OCR models and new methods, we present our work on training Mongolian OCR systems and introduce the MnOCR① dataset. This dataset comprises 80,000 lines of traditional Mongolian text that have undergone meticulous proofreading to ensure accuracy and quality, and these texts are used to generate training data. Additionally, we include another 10,000 lines of text images extracted from various Mongolian books covering different topics. These images, along with their meticulously transcribed text, serve as the test set for evaluating the performance of OCR models. Furthermore, we have made accessible codes for generating synthetic training data as well as constructing Mongolian OCR systems utilizing the MnOCR dataset. These resources aim to facilitate further research and advancements in Mongolian OCR technology.

The rest of the paper is structured as follows. Section 2 presents an overview of related work in the field of Mongolian OCR and publicly accessible large-scale OCR datasets. In Section 3, we delve into the specific challenges encountered in printed Mongolian OCR. Section 4 details the methodology for synthesizing training data and outlines the process of training a Mongolian OCR system using the compiled dataset. Finally, Section 5 concludes the paper.

2. Related Work

2.1 Traditional Mongolian OCR

Mongolian is the language spoken by approximately 6 million Mongolian people. Amidst the wave of informatization, a significant amount of printed Mongolian literature still awaits digitization and further processing. Traditional Mongolian script exhibits a distinctive vertical cursive style, as depicted in Figure 1. This script is written vertically from top to bottom, progressing gradually from left to right in each column. The letters within a Mongolian word are connected, similar to Arabic script, forming a continuous stem. This characteristic sets Mongolian apart from other writing systems.

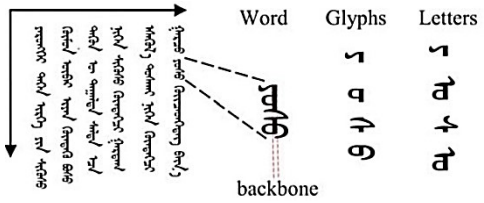

Figure 1: Mongolian Word Spelling[7]

Mongolian OCR research commenced in 2003 when Gao initially put forward a method specifically designed for this unique language and its script [8]. The methodology encompassed a preprocessing step, including skew correction [9] and layout analysis [10], to accurately segment Mongolian text into individual characters. Following this segmentation, traditional feature extraction and matching techniques were employed for character recognition. To further enhance recognition accuracy, in-depth research on feature selection were conducted [11][12]. Over time, more advanced and complex models emerged, such as BP neural networks [13] and Convolutional Neural Networks (CNNs) [14], which were integrated within the OCR framework.

All these methods rely on segmentation, which involves breaking words down into individual characters. Latin and Chinese scripts are relatively straightforward to segment since their characters are separate. However, the Mongolian script presents a unique challenge as its characters are interconnected, forming a continuous stem that makes it difficult to isolate them individually. Various approaches have been proposed to address this issue, including methods rooted in projection, stem analysis, and word contour tracking [13][15][16]. Nonetheless, segmentation remains error-prone, and these errors have the potential to propagate into recognition outcomes, ultimately compromising the overall accuracy of the results.

① Visit http://mglip.com/corpus/corpus.html for the access information

With the rapid advancement of deep learning, there has been a growing application of sequence modeling methods in OCR (Optical Character Recognition) tasks. This development has led to an expansion in recognition units, evolving from single characters to words and, more recently, entire text lines. OCR tasks are viewed as sequence-to-sequence problems, where the input is a text image represented as a sequence of visual features, and the output is a corresponding sequence of characters. By formalizing the OCR task as a sequence-to-sequence problem and adopting an end-to-end approach, text recognition becomes a unified process. This integration consolidates multiple steps traditionally performed separately—such as segmentation, feature extraction, and feature matching—into a single, cohesive model. As a result, the final output is less likely to be adversely affected by errors in intermediate steps, such as segmentation errors. Furthermore, the powerful representation learning capabilities of deep neural networks not only free researchers from the exhausting work of designing and selecting hand-crafted features but also significantly improve recognition performance [17].

In 2017, Zhang introduced end-to-end methods to Mongolian OCR, successfully constructing a segmentation-free system [7]. Then, the end-to-end methods are upgraded and adapted in following works [18][19]. The recognition unit in these methods is the Mongolian word, eliminating the need for complex character segmentation and simplifying the process by segmenting text images into words. However, in Mongolian words, certain suffixes are written separately from the stem with a space, leading to their incorrect segmentation into two distinct words. This is the primary source of error in these methods. To mitigate this problem, we treat the entire text line as the recognition unit, eliminating the need for word segmentation altogether. This approach effectively reduces errors caused by mis-segmentation and significantly improves the overall accuracy of recognition.

2.2 Database for OCR

Most OCR methods are data-driven, making large-scale datasets crucial for training text recognizers. Major languages benefit from having a vast user base and widespread usage, facilitating the collection of resources. As a result, large-scale, publicly accessible datasets are often easily obtainable for these languages, particularly English and Chinese. Datasets like ICDAR 2003 (IC03) [20], ICDAR 2013 (IC13) [21], COCO-Text [22], Total-Text [23], and ICDAR 2017-RCTW [24], collected from real-world scenarios, have significantly contributed to the advancement of new models and methodologies. Additionally, synthetic datasets like Synth90k [25] and SynthText [26] have also played a vital role.

Databases suitable for OCR training in minor languages like Mongolian are exceedingly rare. To the best of our knowledge, the only printed Mongolian OCR database currently available is published alongside [7]. However, this database is limited to a word list printed in various fonts and lacks real-world test data, which is crucial for OCR performance evaluation. Interestingly, there are more handwriting-based Mongolian OCR databases, such as [27][28].

To tackle the challenge of data scarcity in Mongolian OCR research, we have introduced the MnOCR dataset in this study. This comprehensive dataset comprises 10,000 lines of text images, carefully scanned from a diverse range of Mongolian books covering various topics, along with precisely transcribed text. These data serve as a robust test set for evaluating OCR systems. Furthermore, we have utilized 80,000 lines of carefully proofread Mongolian text to generate synthetic training data. Through the incorporation of image augmentation techniques, these synthetic images effectively emulate real-world image characteristics, ultimately enhancing the recognizer's accuracy. Our experimental findings indicate that an OCR system trained on synthetic data can attain a performance comparable to that of a system trained on real-world datasets.

2.3 Discussion on Practical Applications and Future Directions

The findings from this study suggest that the MnOCR dataset and the proposed OCR model have the potential to significantly improve the accuracy of traditional Mongolian OCR systems. Our approach of eliminating word segmentation and treating entire text lines as the recognition unit has shown promising results, reducing errors caused by mis-segmentation. This advancement can be directly applied to the digitization of historical Mongolian texts, facilitating the preservation and accessibility of Mongolian cultural heritage.

In real-world applications, this model could be used to digitize various types of Mongolian literature,

including historical documents, religious texts, and educational materials, making them more accessible for both research and general use. Furthermore, it could be employed in government agencies and educational institutions for administrative purposes, such as the preservation of official records and the promotion of Mongolian script education. The implementation of this model could lead to more comprehensive and accurate digital archives of Mongolian historical texts, further contributing to the revitalization of the traditional script.

However, the current study does have limitations. One significant limitation is the relatively small size of the dataset, especially in comparison to large-scale datasets available for languages like English or Chinese. While the MnOCR dataset provides a solid foundation for training models, the availability of more diverse and real-world data would further enhance the model's performance and generalization ability. Additionally, our model is primarily focused on printed text and does not yet handle handwriting-based recognition, which remains a significant challenge in OCR.

Future research directions may include expanding the MnOCR dataset by incorporating more diverse sources, especially handwritten Mongolian texts and texts from different historical periods. Additionally, by leveraging large language models, integrating multilingual capabilities into the OCR system to enable recognition of languages such as Mongolian, Chinese, and Latin could significantly broaden its applicability. Research could also focus on improving the model's ability to handle degraded or low-quality texts, such as those found in ancient or damaged manuscripts.

3. Discussion on the Training Challenges of Traditional Mongolian OCR

All previous works on Mongolian OCR, such as [7], [10]-[15], [17]-[19], have stated that the Mongolian OCR task is challenging. The reasons listed include the vertical writing direction, character ligatures, large vocabulary, multi-font styles, and more. However, we believe that these reasons alone are not sufficient to fully explain the complexity of the task. Firstly, although Mongolian texts are originally written vertically, rotating them 90 degrees does not hinder reading, as many readers are accustomed to scanning from left to right and top to bottom. Additionally, numerous Mongolian word processing software currently do not offer support for vertical text arrangement. Consequently, the vertical writing direction in Mongolian OCR does not pose a significant challenge, as we can simply rotate the text by 90 degrees to adapt it to the well-established Latin and Chinese OCR frameworks. Secondly, Mongolian has ligatured characters, which traditionally pose a challenge for character segmentation. However, the segmentation-free method obviates the need to segment the text into individual characters, thereby reducing the complexity of this task. As a result, the ligatured Mongolian characters does not pose a significant challenge when using the segmentation-free approach. Thirdly, Mongolian has a large vocabulary, but our recognition method does not treat word as recognition unit. Instead, we recognit the text as a sequence of characters, with the individual character serving as the recognition unit, and the set of Mongolian characters is finite. As a result, the the large vocabulary size of Mongolian poses minimal challenges to our recognition system. Lastly, Mongolian, like Latin and Chinese, exhibits numerous font style variations. While these variations can be addressed with extensive training data for Latin and Chinese OCR systems, the same applies to Mongolian OCR. In fact, this challenge is not exclusive to Mongolian OCR alone.

We believe that the fundamental challenge in Mongolian OCR lies in the mismatch between glyph representation and text encoding. The traditional Mongolian character encoding scheme, standardized in both China's national standard system (GB) and Unicode, focuses on encoding the pronunciation of Mongolian words. While this facilitates the mapping of phonemes to letters, benefiting automatic speech recognition (ASR) tasks, it poses a significant hurdle for OCR.

As defined in GB 25914-2010 [29] and Unicode [30] standards, Table 1 depicts the Mongolian alphabet, consisting of 35 letters: 8 vowels and 27 consonants. Traditional Mongolian script is a context-sensitive writing system characterized by initial, medial, and final forms for each letter. Many letters also exhibit variant forms based on spelling and grammatical contexts, leading to a significantly larger number of glyph shapes compared to the number of letters. Table 2, adapted from [31], provides an example of these variant glyph shapes. In addition to the standard Mongolian letters, there exist several special characters that are essential for correctly shaping glyphs in Mongolian words. These special

characters include the Free Variation Selectors (specifically FVS1 with the Unicode code point U+180B, FVS2 with U+180C, and FVS3 with U+180D), the Mongolian Vowel Separator (MVS, identified by U+180E), and the Narrow No-Break Space (NNBSP, represented by U+202F). Furthermore, specific combinations of Mongolian letters with the aforementioned special characters result in the formation of a single ligature. As far as we are aware, there are over 200 such ligatures known to exist. For a comprehensive list of these ligatures, readers are encouraged to consult reference [32]. Therefore, traditional Mongolian script is considered a complex writing system.

Table 1: Mongolian Letters

Vowel									
Consonant									

Table 2: Glyph of Letter in Mongolian[32]

Mongolian letters		Isolate forms	Initial forms	Medial forms	Final form
Naming in Unicode	Code point in Unicode				
‘a’	U+1820				
‘o’	U+1823				
‘u’	U+1824				
‘oe’	U+1825				
‘ba’	U+182A				
‘ta’	U+1832				
‘da’	U+1833				

From Table 2, it is evident that a single coded letter can correspond to many distinct symbols, and conversely, a single symbol can be associated with multiple coded letters. This lack of a one-to-one correspondence between glyph shapes and text encoding poses a significant challenge for Mongolian typists, who struggle to accurately input the correct characters based only on the visual representation of the glyphs. Consequently, there is a widespread occurrence of misspelled Mongolian language texts. In this context, we define misspelling as words that visually appear correct in terms of their glyph representation but are encoded incorrectly. According to the pronunciation, Mongolian orthography, and Unicode encoding rules, each specific word has a unique correct encoding, and any deviation from this standard is considered a misspelling.

According to a case study [33], theoretically, there are 1728 possible Mongolian spellings for the word "ᠶᠡᠬᠡᠢᠯᠡᠮᠡᠵᠢ" (meaning: ethnic group), all of which could produce the visually correct glyph appearance when printed. However, only one of these spellings is considered linguistically accurate. Using a corpus consisting of 76 million Mongolian words sourced from the internet, the word was found to appear 102,532 times. Of these occurrences, only 24,708 were spelled correctly, indicating that over 75% of instances were misspelled.

If we attempt to train an OCR model to recognize Mongolian words using data from the previous case study, we encounter a significant challenge. In the training set, there are 102,532 word images that look likes the word "ᠶᠡᠬᠡᠢᠯᠡᠮᠡᠵᠢ". However, each of these images could theoretically correspond to any one of 1728 different spellings. The presence of such many potential targets creates confusion for the learning algorithm, causing it to vacillate between different spellings and making it difficult for the training process to converge or converge to incorrect model parameters. This issue is a classic example of "label noise", where inconsistencies in the training labels hinder the model's ability to learn effectively. To address this problem, removing the label noise by ensuring accurate and consistent labeling is the only viable solution.

Previous research on Mongolian OCR has avoided utilizing real-world texts riddled with spelling errors for model training. For instance, [17] relied on a word list sourced from an authoritative Mongolian dictionary. Other studies [7],[10]-[15] circumvented the issue of Mongolian encoding label noise by defining their training targets based on glyph-specific encodings, rather than predicting the actual Mongolian word encoding. If we must follow the same careful selection process for training data as in previous works, it will become challenging to efficiently scale up our dataset and develop a recognition system with improved performance. In this work, we utilize our previous research on spelling correction [33] to clean the raw text corpus and reduce label noise in the training data. This purification process enables us to generate synthetic text images from these refined texts, significantly expanding the scale of our training dataset at a relatively low cost.

4. Experiments and Analysis

4.1 Data Collection

We have collected authentic scanned printed Mongolian text line data sourced from a diverse range of over 200 Mongolian books covering various themes. Subsequently, we hired 20 Mongolian typists for a careful labeling process that lasted over a year, ensuring that each line of text was accurately labeled with its corresponding text. Due to the extensive workload, we were unable to conduct a comprehensive secondary check on all transcriptions. Although we were unable to perform a comprehensive secondary check on all transcriptions due to the extensive workload, sampling checks identified numerous errors in the initial transcripts. Therefore, we must apply additional refinement and validation steps to ensure the accuracy of our data. The refinement process involves several steps. Firstly, based on the confidence scores provided by a pretrained OCR system, an automatic filtering program was utilized to select 100,000 text lines that were labeled as highly likely to be correct for further refinement. Subsequently, each text line undergoes meticulous checking by two inspectors who are instructed not to correct any errors. Instead, if either inspector believes that a text line does not match the corresponding transcription, it is promptly deleted from the database. This rigorous approach ensures that only highly accurate text lines are retained. In the end, approximately 90,000 text lines remain. Notably, this process reveals that the sentence error rate (SER) of human Mongolian typists is approximately 10% or higher, with an average of 1 to 2 errors per page. This suggests that even for humans, Mongolian text recognition is not a trivial task.

The database is randomly divided into two distinct sets: 80,000 text lines for training purposes and 10,000 for testing. Both the test set and the training set transcriptions from the MnOCR database are made publicly accessible, with the test set including both text line images and their corresponding transcriptions. Additionally, all transcriptions undergo automatic spelling correction to ensure their accuracy.

4.2 Synthesis Text Image Generation

We generate synthetic text images in a two-step process. First, we render the Mongolian text onto a blank canvas, allowing for customization of fonts to produce single-line text images. Secondly, we employ the "straug" toolkit② to enhance these text images with various effects, including noise, blur, and others. Since the generated Mongolian text lines often have a high aspect ratio and delicate strokes, certain effects from the "straug" library may not yield realistic outcomes. Therefore, we carefully select a subset of effects that produce visually acceptable results. Additionally, to mimic the aliasing artifacts commonly seen in real-world images, we introduce scaling variations into our set of candidate effects. Each time an image is enhanced, one of these effects is randomly chosen. The code for our data generation pipeline is publicly available③.

4.3 End-to-end OCR Model

In this work, we employ the convolutional recurrent neural network (CRNN) architecture to construct the OCR system, as cited in [34]. The CRNN is capable of recognizing text sequences of

varying lengths in an end-to-end manner, eliminating the need to segment the input text line into individual characters. It facilitates solving sequence learning problems by converting the visual feature sequences extracted from text images into corresponding character sequences.

The CRNN network is composed of three distinct parts: the CNN (convolutional layer), the RNN (recurrent layer), and the transcription layer with CTC loss. The CNN utilizes deep convolutional neural networks to extract features from the input image, resulting in feature maps. The RNN, specifically employing BiLSTM (Bi-directional Long Short-Term Memory), predicts the feature sequence, learns each feature vector within the sequence, and outputs the predicted label distribution. Finally, the transcription layer, leveraging CTC (Connectionist Temporal Classification) loss, converts the series of label distributions obtained from the recurrent layer into the final label sequence.

We implemented the CRNN using the PaddleOCR toolkit, and the configuration of the OCR model is detailed in Table 3. To evaluate the performance of the model, we employed the Character Error Rate (CER) and Word Error Rate (WER) as evaluation metrics.

Table 3: Model Setting

Algorithm	CRNN
Backbone	MobileNetV3
Neck	SequenceEncoder
Encoder Type	RNN
Hiden Size	96
Loss Type	CTCLoss

4.4 Comparison

We utilize the identical architecture as depicted in Table 3 for all comparison models, with variations only in the training data employed. Specifically:

- The “Real” model: Trained using 80,000 real scanned data samples.
- The “Synth” model: Trained using 80,000 synthetic text images.
- The “Aug” model: Trained using 80,000 synthetic text images that have been augmented.
- The “Synth+Aug” model: Trained using a combination of 80,000 synthetic text images and an additional 80,000 augmented synthetic text images, totaling 160,000 training samples.

The comparison results presented in Table 4 demonstrate that using synthetic data alone is insufficient to achieve an optimal OCR system due to the mismatch between clean printed text and scanned text. However, the introduction of data augmentation techniques helps to bridge this gap by

② https://github.com/roatienza/straug

③ https://github.com/IMU-M-Lab/MnOCR-baseline

providing variations that more closely resemble real-world scanned text. Furthermore, the performance of the OCR system trained on synthetic data can be further enhanced with additional training samples. It is noteworthy that the text used to generate synthetic data does not necessarily have to originate from the real training set. Instead, it can be sourced from a variety of texts available on the Internet. This approach enables us to significantly expand the scale of training data while keeping costs low. By leveraging synthetic data generation and augmentation techniques, we can effectively improve the performance of OCR systems without relying only on expensive and time-consuming manual data annotation processes.

Table 4: Experimental Results

Data Type	CER	WER
Real	0.81%	4.03%
Synth	3.85%	17.25%
Aug	2.57%	12.93%
Aug+Synth	1.95%	8.86%

We observed that there are also some misspell in the OCR results. To address this issue, we employed an automatic spelling correction to refine the recognition results. As reported in Table 5 we evaluated the CER and WER of the various comparison models after applying the spelling correction. The results indicate that there are improvements across all models when the spelling correction is applied. This suggests that the OCR models, while trying to replicate the spelling patterns of Mongolian words like human typists, may sometimes produce visually correct but encoding-wise incorrect outputs. The spelling correction helps to correct these encoding errors, leading to more accurate OCR results.

Table 5: Experimental Results with Spell Correction

Data Type	CER	WER
Real	0.66%	3.06%
Synth	3.56%	14.54%
Aug	2.26%	10.28%
Aug+Synth	1.73%	7.12%

5. Conclusion

In this study, we approached the OCR task as a sequence-to-sequence problem, focusing on building a segmentation-free Mongolian OCR model. This model is capable of recognizing entire text lines without the need for prior segmentation into words or characters. To achieve this, we leveraged the power of data-driven deep learning methods.

To meet the data requirements of deep learning models, we employed data synthesis and augmentation techniques. Our experimental results suggest that the use of synthetic data has the potential to achieve comparable recognition accuracy to that of costly real data collection methods. This finding is significant as it opens up the possibility of using synthetic data as a viable alternative to real data in OCR tasks, especially when real data is scarce or difficult to obtain.

Furthermore, all resources used in building our OCR model are publicly accessible for research purposes. We believe that making these resources available to the research community will facilitate further advancements in Mongolian OCR and related fields. We hope that our work will serve as a valuable resource for researchers working on Mongolian OCR and inspire future studies in this area.

The broader implications of this work extend beyond Mongolian OCR and can be applied to other minority languages facing similar challenges in digital text recognition. Many minority languages, like Mongolian, face significant hurdles in developing OCR systems due to a lack of large-scale datasets, limited research, and fewer commercial incentives. The approach we have developed, which combines sequence-to-sequence learning with data synthesis and augmentation, can be adapted to other minority languages with unique scripts. For example, languages such as Tibetan, Uyghur, and various indigenous languages could benefit from similar methods, enabling better preservation and accessibility of their cultural and historical texts.

Moreover, the success of synthetic data in OCR tasks suggests that similar techniques could be applied to these languages, which often suffer from a lack of real-world training data. By leveraging synthetic datasets, it becomes possible to develop robust OCR systems for languages with limited resources, ultimately contributing to the digital preservation and revitalization of minority languages.

Author Contributions

Guanglai Gao (Email: csggl@imu.edu.cn) has proposed the research problems. Hui Zhang (Email: cszh@imu.edu.cn) has designed the research framework, wrote and revised the manuscript. Feilong Bao (Email: csfeilong@imu.edu.cn) has proposed and implemented the data collection plan. Hongxi Wei (Email: cswhx@imu.edu.cn) has designed and implemented the OCR program for data filtering. Yunpeng Bai (Email: byp@mail.imu.edu.cn) has conducted the experiment, analyzed the results, and contributed to writing and revising the manuscript. Weiqi Wang (Email: 32209024@mail.imu.edu.cn) has also conducted the experiment, analyzed the results, and participated in writing and revising the manuscript.

Acknowledgements

This research was supported in part by the National Natural Science Foundation project (No.62066033), Inner Mongolia Natural Science Foundation Outstanding Youth Fund project (No.2022JQ05), Inner Mongolia Autonomous Region Science and Technology Plan project (No.2021GG0158), Hohhot City University-Institute Collaborative Innovation project, and Inner Mongolia University Young Scientific and Technological Talent Cultivation project (No.21221505).

References

[1] H. F. Schantz, History of OCR, Optical Character Recognition. History of OCR, Optical Character Recognition, 1982.
[2] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in Proceed ings of the 21st international conference on pattern recognition (ICPR2012). IEEE, 2012, pp. 3304–3308.
[3] C. Yao, X. Bai, and W. Liu, "A unified framework for multior iented text detection and recognition," Image Processing IEEE Transactions on, vol. 23, no. 11, pp. 4737–4749, 2014.
[4] G. Nagy, "Twenty years of document image analysis in pami," IEEE Transactions on Pattern Analysis and Machine Intelli gence, vol. 22, no. 1, pp. 38–62, 2000.
[5] Y. Zhou, S. Liu, Y. Zhang, Y. Wang, and W. Lin, "Perspective scene text recognition with feature compression and ranking," in Asian Conference on Computer Vision. Springer, 2014, pp. 181–195.
[6] N. Natsagdorj, "Mongolian handwriting character recognition based on convolutional neural network (cnn)," in 6th Interna tional Conference on Information Engineering for Mechanics and Materials. Atlantis Press, 2016, pp. 580–584.
[7] Z. Hui, H. Wei, F. Bao, and G. Gao, "Segmentation-free printed traditional mongolian ocr using sequence to sequence with at tention model," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 585 590.
[8] G. Gao, W. Li, H. Hou, and Z. Li, "Multi-agent based recog nition system of printed mongolian characters," in Proceedings of the International Conference on Active Media Technology, 2003, p. 376–381.
[9] H. Wei and G. Gao, "A skew detection method of mongo lian documents images," Journal Of Inner Mongolia Univer sity(Natural Science Edition), vol. 38, no. 5, pp. 586–590, 2007.
[10] ——, "A method of layout analysis for mongolian document images based on connected components," Journal Of Inner Mongolia University(Natural Science Edition), vol. 38, no. 5, p. 5, 2007.
[11] Z. Li and G. Gao, "Extraction of features of mongolian printed character recognition," Microcomputer Development, vol. 13, no. 11, pp. 117–119, 2003.
[12] H. Wei and G. Gao, "Feature selection of mongolian characters in the recognition of printed mongolian characters," Journal of Inner Mongolia University (Natural Science Edition), vol. 37, no. 6, pp. 694–697, 2006.
[13] ——, "Machine-printed traditional mongolian characters recog nition using BP neural networks," in 2009 International Confer ence on Computational Intelligence and Software Engineering. IEEE, 2009, pp. 1–7.
[14] H. Hu, H. Wei, and Z. Liu, "The cnn based machine-printed traditional mongolian characters recognition," in 2017 36th Chinese Control Conference (CCC). IEEE, 2017, pp. 3937 3941.
[15] W.Li, G. Gao, H. Hou, andZ.Li, "Adesign and implementation of element segmentationin the recognition of printed mongolian characters," Journal Of Inner Mongolia University(Natural Sci ence Edition), vol. 34, no. 3, pp. 357–360, 2003.
[16] W.Li, G. Gao, H. Hou, andZ.Li, "Adesign and implementation of element segmentationin the recognition of printed mongolian characters," Journal Of Inner Mongolia University(Natural Sci ence Edition), vol. 34, no. 3, pp. 357–360, 2003.
[17] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, "Text recognition in the wild: A survey," ACM Computing Surveys (CSUR), vol. 54, no. 2, pp. 1–35, 2021.
[18] W. Wang, H. Wei, and H. Zhang, "End-to-end model based on bidirectional lstm and ctc for segmentation-free traditional mongolian recognition," in 2019 Chinese Control Conference (CCC), 2019, pp. 8723–8727.
[19] H. Wei, C. Liu, H. Zhang, F. Bao, and G. Gao, "End-to-end model for offline handwritten mongolian word recognition," in Natural Language Processing and Chinese Computing, 2019, pp. 220–230.
[20] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto et al., "Icdar 2003 robust reading competitions: entries, results, and future directions," International Journal of Document Analysis and Recognition (IJDAR), vol. 7, no. 2, pp. 105–122, 2005.
[21] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. De Las Heras, "Icdar 2013 robust reading competition," in 2013 12th international conference on document analysis and recognition. IEEE, 2013, pp. 1484–1493.
[22] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Be longie, "Coco-text: Dataset and benchmark for text de tection and recognition in natural images," arXiv preprint arXiv:1601.07140, 2016.
[23] C.-K. Ch'ng, C. S. Chan, and C.-L. Liu, "Total-text: toward ori entation robustness in scene text detection," International Jour nal on Document Analysis and Recognition (IJDAR), vol. 23, no. 1, pp. 31–52, 2020.
[24] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, "Icdar2017 competition on reading chinese text in the wild (rctw-17)," in 2017 14th iapr international conference on document analysis and recognition (ICDAR), vol. 1. IEEE, 2017, pp. 1429–1434.
[25] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," arXiv preprint arXiv:1406.2227, 2014.
[26] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2315–2324.

[27] L.-L. Ma, J. Liu, and J. Wu, “A new database for online hand written mongolian word recognition,” in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 1131 1136.
[28] D. Fan, G. Gao, and H. Wu, “MHWMongolianoffline handwrit ten dataset and its application,” Journal of Chinese Information Processing, no. 1, pp. 89–95, 2018.
[29] GB25914-2010: Information technology of traditional Mongo lian nominal characters, presentation characters and control characters using the rules, 2010.
[30] The Unicode Consortium, “The Unicode Standard, Version 14.0.0, (Mountain View, CA: The Unicode Consortium, 2021. ISBN 978-1-936213-29-0),” Standard, 2021. [Online]. Available: https://www.unicode.org/versions/Unicode14.0.0/
[31] F. K. A. M. Biligsaikhan Batjargal, Garmaabazar Khaltarkhuu, ing: Use of unicode in opentype fonts.” International Journal on Asian Language Processing, vol. 21, no. 1, pp. 23–43, 2011.
[32] K. Whistler, “Unicode Technical Repor#54: UNICODE MONGOLIAN 12.1 SNAPSHOT,” Tech. Rep., 2020. [Online]. Available: https://www.unicode.org/reports/tr54/
[33] M. Lu, F. Bao, G. Gao, W. Wang, and H. Zhang, “An auto matic spelling correction method for classical mongolian,” in International Conference on Knowledge Science, Engineering and Management. Springer, 2019, pp. 201–214.
[34] B. Shi, B. Xiang, and Y. Cong, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298– 2304, 2016.