*International Journal on Data Science and Technology*
2025, Vol. 11, No. 2, pp. 27-34
https://doi.org/10.11648/j.ijdst.20251102.12

SciencePG
Science Publishing Group

Research Article

# RanPil: New Dataset and Benchmark for Offline Handwritten Korean Text Recognition

**Hyon-Gwang O, Myong-Chol Kim, Il-Nam Pak, Un-Hyok Choe, Chol-Jun O***

Institute of Mathematics, State Academy of Sciences, Pyongyang, Democratic People's Republic of Korea

## Abstract

In recent years, since deep learning technology have been applied to handwritten text recognition, the need for handwritten document image Datasets has been growing more and more. In particular, the development of the dataset is of great significance for improving performance of handwritten Korean text recognition because no dataset for handwritten Korean text recognition has been published. In this paper, we present the “RanPil”, a new training and performance evaluation dataset for handwritten Korean text recognition, which consists of a total of 8,600 pages of images (182,000 text lines and 4,300,000 characters) written by 1,804 people. We evaluate writing- diversity of handwritten document images, such as text line spacing, text line slope, character size, word spacing, and character compactness. In addition, we propose an MOS (Mean Opinion Score) evaluation method for the scrawl-level. Finally, we evaluate the performance of TrOCR based on vision encoder and decoder with a test dataset classified by the scrawl-levels.

## Keywords

Deep Learning Technology, Handwritten Korean Dataset, HTR (Handwritten Text Recognition)

## 1. Introduction

The HTR has been attracted many researchers for decades due to importance of its application. HTR presupposes the development of writing datasets, and the diversity of datasets plays a crucial role in improving recognition performance. Recently, many HTR datasets have been presented to enhance the recognition performance in this field. Generally, HTR datasets are divided into two groups, i.e., online and offline datasets. Typical offline HTR datasets include the English HTR dataset IAM [1], the French HTR dataset RIMES [2], the Arabic HTR dataset KHATT [3], the Chinese HTR datasets HIT-MW [4], SCUT-EPT [5], HCUT-HCCDOC [6], CASIA-HWDB 2.0~2.2 [7], and the German HTR dataset READ [14]. Also, online HTR datasets include the English HTR dataset IAM-OnDB [9] and the Japanese HTR dataset Kondate [8].

In recent years, the recognition performance has improved significantly, as deep learning techniques have been applied to HTR, and its performance depends on the size and diversity of the dataset. Currently, HTR has evolved from character-level recognition to text line, paragraph, or document-level recognition [13-16], and many HTR datasets supporting this recognition have been presented [1-7]. In order to achieve practically meaningful HTR performance, both the size of the dataset and writing diversity are the most important factors. Several national HTR datasets have been published worldwide, but no Korean HTR dataset has been

*Corresponding author: ocj1989@star-co.net.kp (Chol-Jun O)

**Received:** 7 April 2025; **Accepted:** 27 April 2025; **Published:** 20 June 2025

published. It is well known that the HTR of the Korean script, similar to Chinese script, is more difficult than other languages because of the large number of character classes, the variety of writing styles and the large number of similar characters.

In general, HTR can be divided into character segmentation based recognition and text based recognition. In character segmentation based recognition, the accuracy of character segmentation has a crucial influence on the recognition results. Text based recognition is a HTR method that recognizes text lines, paragraphs, or even pages, unlike character segmentation based recognition, does not require character segmentation and can use a language modeling. Therefore, text based recognition has become the mainstream HTR method [13].

In this paper, we present the "RanPil", being a benchmark for offline handwritten Korean text recognition. “RanPil” consists of a total of 8,600 pages of images and corresponding text label data, including 182,000 lines (4,300,000 characters) written by 1,804 writers. “RanPil” is divided into training and testing data, of which the training dataset contains 7,600 pages written by 1,696 writers and the testing dataset contains 1,000 pages written by 108 writers, respectively. The writers of the training and testing datasets were separated strictly.

We evaluate the writing diversity by using text line spacing, text line slope, character size and word spacing, and character compactness. Human writing characteristics (style of writing) are one of the most important factors that greatly affects recognition performance. Hence, we present a new evaluation index that reflects the scrawl-level and proposes a method to classify the dataset by MOS evaluation method. We also evaluate the performance of TrOCR [13] based on vision encoder and decoder in the “RanPil” dataset.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the proposed handwritten Korean document image dataset “RanPil” and presents its features as writing diversity, and also presents classification method of "RanPil" by MOS evaluation method. Lastly we evaluate recognition performance of TrOCR in “RanPil”. Section 4 shows the conclusion.

## 2. Related Works

In recent years, the ICDAR has published a number of new HTR datasets in different languages for handwritten text recognition research. Table 1 shows the multilingual datasets and their characteristics that are widely used in the field of HTR. As a dataset for handwritten English text recognition, the IAM dataset is an offline handwritten English dataset with 9,285 text lines (82,227 words) which by 400 writers have written from LOB (Lancaster-Oslo/Bergen) text corpus data. The IAM-OnDB dataset is also an online handwritten English dataset with 13,049 text lines (86,272 characters) written using an electronic pen by 221 people. The RIMES dataset is an offline French paragraph dataset with total of 1,600 paragraphs (12,111 text lines) contributed by 1,300 writers. The READ dataset is an offline handwritten German dataset with 1,962 pages of images and 10,550 text lines.

Many handwritten Chinese datasets have been published, typically HIT-MW, SCUT-EPT, SCUT-HCCDoc, and CASIA-HWDB 2.0-2.2. The HIT-MW dataset is an image dataset collected through e-mail for scientific research purposes, including 8,664 text lines (186,444 characters) written by 780 writers. In addition, the SCUT-EPT dataset is a Chinese dataset collected from an examination paper, including 50,000 text lines written by 2,986 writers. On the other hand, the SCUT-HCCDoc including text lines (1,515,801 characters) and 12,253 handwritten document images captured by the camera. This dataset is broadly classified into image-level diversity, text-level diversity, and character-level diversity. The CASIA-HWDB 2.0-2.2 dataset consists of 52,230 text lines written by 1,019 writers and is the most widely used Chinese dataset.

The Kondate dataset with total of 12,232 lines collected from 100 people as a dataset for online handwritten Japanese text recognition. In addition, the KHATT [3] dataset is an Arabic text dataset consisting of 1,000 handwritten forms written by 1,000 writers from different countries, which can be used for paragraph and line-level recognition tasks.

*Table 1. Multilingual handwritten datasets.*

| dataset name | writers | images | text lines | words | characters | classes | online/offline | language |
|---|---|---|---|---|---|---|---|---|
| IAM [1] | 400 | 1 539 | 9 285 | 89 896 | 86 227 | 81 | offline | English |
| IAM-OnDB [9] | 221 | - | 13 049 | - | 86 272 | 81 | online | English |
| RIMES [2] | 1 300 | 1 600 | 12 111 | 60 000 | - | 79 | offline | French |
| READ [14] | - | 1 962 | 10 550 | - | - | 93 | offline | German |
| KHATT [3] | 1 000 | - | - | - | - | 49 | offline | Arabic |
| Kondate [8] | 100 | - | 12 232 | - | 130 956 | 1 106 | online | Japanese |

| dataset name | writers | images | text lines | words | characters | classes | online/offline | language |
|---|---|---|---|---|---|---|---|---|
| HIT-MW [4] | 780 | - | 8 664 | - | 186 444 | 3 041 | offline | Chinese |
| CASIA-HWDB 2.0~2.2 [7] | 1 019 | - | 52 230 | - | 1 344 414 | 2 703 | offline | Chinese |
| SCUT-EPT [5] | 2 986 | - | 100 000 | - | 2 534 322 | 4 250 | offline | Chinese |
| HCUT-HCCDoc [6] | - |  | 116 000 |  | 1 150 000 | 6 109 | offline | Chinese |

# 3. Dataset for Offline Handwritten Korean Text Recognition – "RanPil"

## 3.1. Configuration Feature

To construct the dataset, 1,804 writers with different ages (102 persons in 10, 354 in 20, 591 in 30, 687 in 40, and 70 in 50) and occupations (102 students (5.7%), 354 university students (19.6%), 608 workers (33.7%), 357 office workers (19.8%) and 383 scientists (21.2%)) used bibliographic data written using different textual data (Novels (19.3%), diaries (4.1%), journals (21.8%), social science (25.6%), natural science (28.5%), and poems (0.7%)). From these bibliographic data, 200 dpi images were collected by scanner. The dataset consists of handwritten images and its corresponding labeled text (txt file) data, the image is segmented by a text line and each text line corresponds to a text data.

The dataset consists of 8,600 pages of images, including 182,000 text lines (4,300,000 characters and 1,200,000 words), and the label data contains 1,198 Korean characters, 42 English characters, 10 digits and 51 characters.

*Table 2. Configuration Feature of "RanPil".*

| character type |  |  |  |  | characters | words | text lines | images |
|---|---|---|---|---|---|---|---|---|
| classes | Korean | English | digits | symbols |  |  |  |  |
| 1 301 | 1 198 | 42 | 10 | 51 | 4.3M | 1.2M | 182K | 8.6K |

## 3.2. Writing Diversity

Since writing characteristics vary from person to person, it is necessary to fully reflect the writing characteristics of many people in order to become a universal dataset for HTR. We express this writing feature as a writing diversity. Writing diversity includes text line spacing, text line slope, word spacing and character size, and character compactness (number of characters per unit length). When assessing diversity, image data are used in units of centimeters, a measure in the paper document because the number of pixels varies on the resolution of an acquisition device.

### 3.2.1. Text Line Spacing Diversity

In the handwritten document images, the text line spacing varies from writer to writer and the distribution characteristics of these text line spacing affect the HTR. As shown in Figure 2-a), the text line spacing in the dataset varies from 0.1 cm to 0.9 cm approximately, and the overlapping characteristics of the text lines are shown in about 7% of the data.

### 3.2.2. Text Line Slope Diversity

Text line slope is also a key factor affecting HTR. Figure 2-b) shows the slope distribution characteristics of the text lines in the dataset "RanPil". As can be seen from the figure, the 3% of the data in the dataset showed strong slope characteristics of the writing environment and in the writing characteristics.

### 3.2.3. Character Size Diversity

A character size is important both in the segmentation and recognition of handwritten document image. As shown in Figure 2-c) and d), the width and height of the character are intensively distributed between 0.3 cm and 0.9 cm.

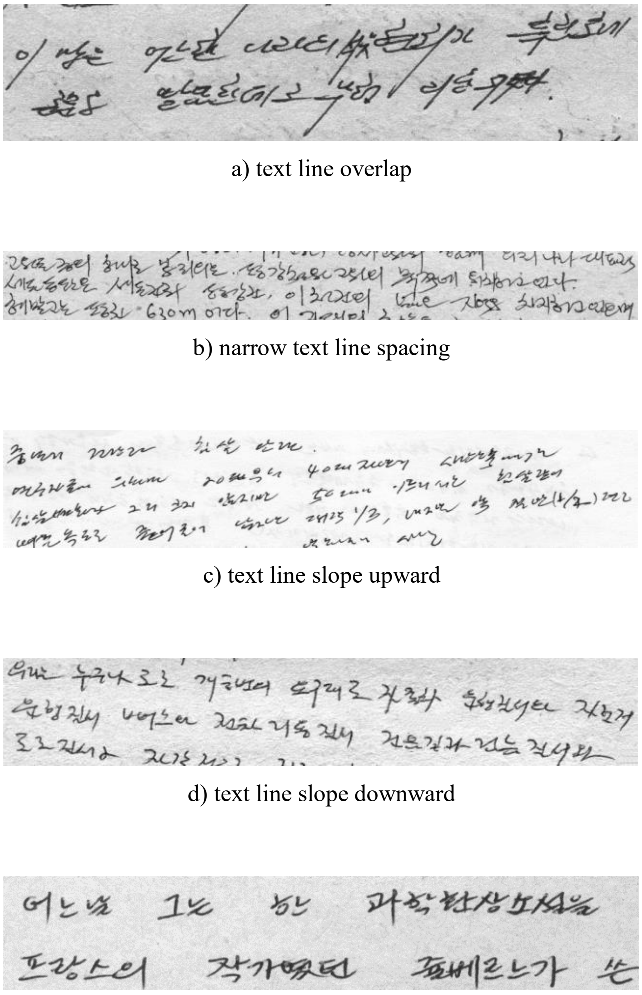
### 3.2.4. Word Spacing Diversity

A word spacing also varies from writer to writer and affect HTR. Figure 2-e) shows the word spacing distribution in the

"RanPil". As can be seen from the figure, the word spacing is very rare for words greater than 1 cm, and word spacing is usually from 0.4 cm to 0.5 cm.

### 3.2.5. Character Compactness Diversity

Figure 2-f) shows the character compactness in the "RanPil". As can be seen from the figure, within 1 cm usually two or three characters are included.



a) text line overlap

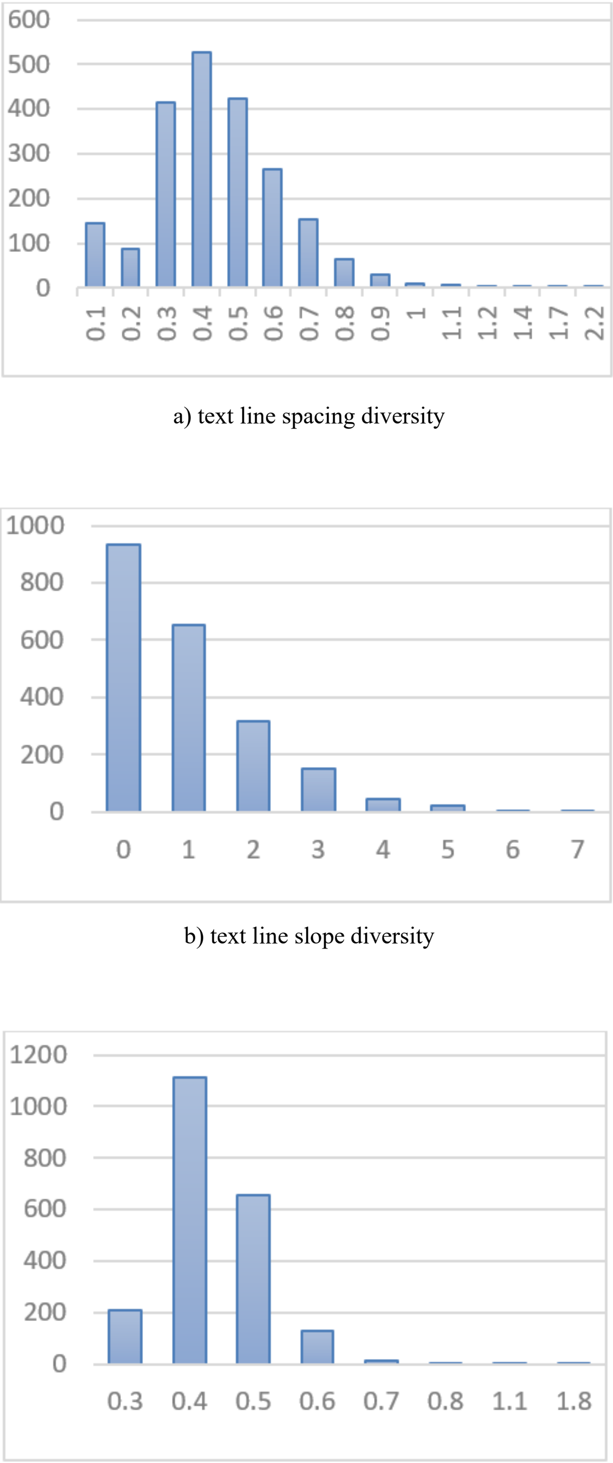
b) narrow text line spacing

c) text line slope upward

d) text line slope downward

e) wide word spacing

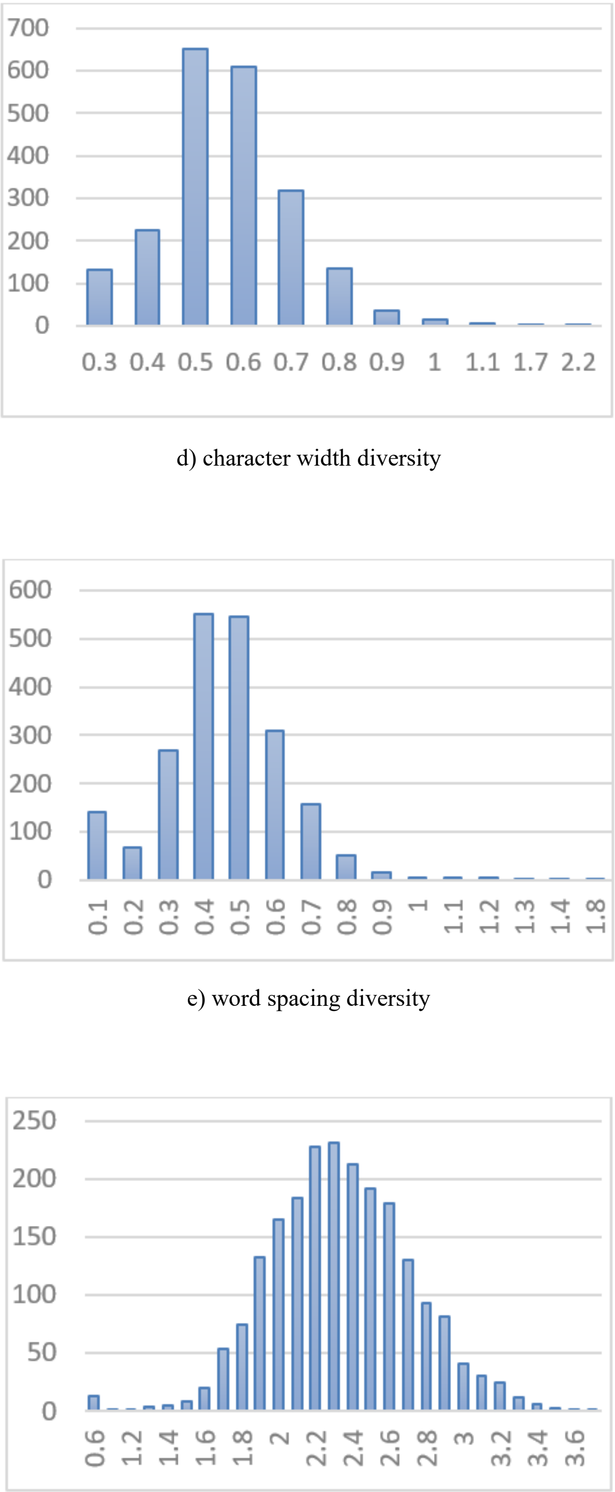
*Figure 1. Shows different examples of writing diversity.*



a) text line spacing diversity

b) text line slope diversity

c) character height diversity



d) character width diversity

e) word spacing diversity

f) character compactness diversity

***Figure 2.*** *Analysis of writing diversity in "RanPil".*

## 3.3. Classification of the "RanPil" by MOS Evaluation Method

In the above, the writing diversity that appears in the handwritten text is defined from the topological characteristics of the elements (text lines, words, and characters) that constitute the handwritten text. However, the most important factor that has a crucial impact on handwritten text recognition performance is the style of writing. To be a meaningful benchmark dataset, the style of writing of people must be fully reflected. In human writing, it is difficult to specify a style of writing as clearly as in printed characters, because of the omission of strokes or the concatenation of adjacent strokes or characters. Therefore, there is no other way but to classify the style of writing of people according to the scrawrl-level.

In this section, we present a method to evaluate scrawl-level by MOS evaluation method and use it to classify the "RanPil". The degree of understanding of writing varies on the level of human knowledge, but it is largely dependent on the scrawl. In fact, different HTR methods also have a large difference in recognition rates depending on the scrawl, so if we evaluate it and classify the dataset, we can evaluate the performance of the recognition methods by classes according to the scrawl.

To the best of our knowledge, no method has been studied to evaluate the scrawl in a handwritten dataset.

### 3.3.1. Definition of the Scrawl-level in Handwritten Character

When a person writes, there are many ambiguous cases because of the use of contiguous characters to be used interconnected or omitted strokes of characters. Thus, when you look at a handwritten text, you usually understand the handwritten characters at the character level, and you can identify the characters by considering words and sentences according to the difficulty of writing. Considering the concrete cases occurring in such writing, we introduce the following definition of the scrawl-level. This definition presupposes the case when an individual sees a character.

A) Character-level handwritten character

It is a handwritten character that can be recognized correctly with itself.

B) Word-level handwritten character

It is a handwritten character that is vague in meaning at the character level but can be recognized correctly at the word- level.

C) Sentence-level handwritten character

It is a handwritten character that cannot be recognized at the character-level or at the word-level, but can be understood exactly at the sentence-level.

D)Ambiguous handwritten character

It means a handwritten character that is ambiguous in meaning even at the sentence- level.

Figure 3 is an example image classified according to the scrawl-level.

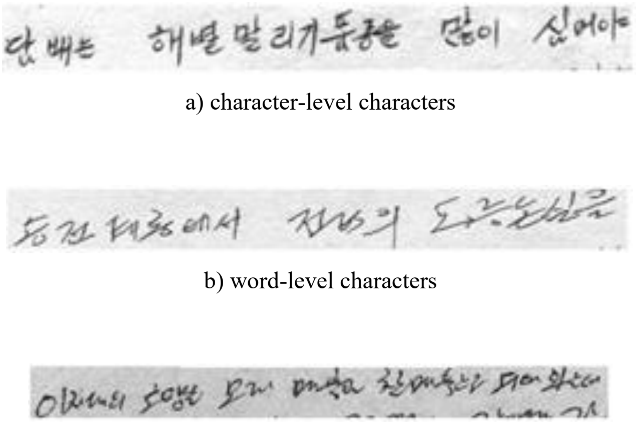
### 3.3.2. The scrawl-assessment in Handwritten Character by MOS Evaluation Method

The MOS evaluation method is applied to the individual's assessment of the scrawl-level of handwritten characters as follows:

$$SC_i = arg\max_{sc_{i,j}}\{P(\mathrm{sc}_{i,j})\}$$

where $\mathrm{sc}_{ij}$ denotes the level evaluated by the j-th member for the character $c_i$, $\mathrm{P}(sc_{i,j})$ is the frequency of $\mathrm{sc}_{ij}$, and $SC_i$ is the scrawl- assessment score for the character $c_i$.

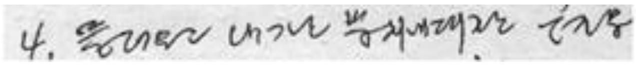
Since the number of levels of the scrawl-assessment is 4, we perform MOS evaluation on five people to eliminate ambiguity.



a) character-level characters

b) word-level characters

c) sentence-level characters



d) ambiguous handwritten character

*Figure 3. Examples of handwritten images according to the scrawl-level.*

### 3.3.3. The Scrawl-assessment Vector

To classify a dataset according to the scrawl-level, an assessment must be made on an individual handwritten document image, and this assessment is not possible only on an individual handwritten character, but also on a vector of evaluations.

Since there are different levels of characters in the handwritten document image, the scrawl-level in the handwritten document image must be evaluated by summing up to what degree each class of characters is. Thus, the scrawl-assessment of a handwritten document image is expressed as the following 4-dimensional evaluation vector:

$$\mathrm{S}_I = (P(SC_c), P(SC_w),\ P(SC_s), P(SC_{ns})) \qquad (1)$$

$$P(SC_c) = \frac{N_{cc}}{N_c}, P(SC_w) = \frac{N_{wc}}{N_c}, P(SC_s) = \frac{N_{sc}}{N_c},\ P(SC_{ns}) = \frac{N_{nsc}}{N_c} \qquad (2)$$

where

$SC_c$: character-level handwritten character

$SC_w$: word-level handwritten character

$SC_s$: sentence-level handwritten character

$SC_{ns}$: ambiguous handwritten character

$P(SC_c)$: the proportion of character-level handwritten characters to all handwritten characters in a handwritten document image

$P(SC_w)$: the proportion of word-level handwritten characters to all handwritten characters in a handwritten document image

$P(SC_s)$: the proportion of sentence-level handwritten characters to all handwritten characters in a handwritten document image

$P(SC_{ns})$: the proportion of ambiguous handwritten characters to all handwritten characters in a handwritten document image

$N_c$: total number of characters

$N_{cc}$: number of character-level

$N_{wc}$: number of word-level

$N_{sc}$: number of sentence-level

$N_{nsc}$: number of ambiguous characters

The scrawl-assessment vector represents the frequency of each class, so it evaluates the scrawl-level of the handwritten document image into the highest frequency class.

### 3.3.4. Classification and Recognition Performance Evaluation of "RanPil" Based the Scrawl-assessment

We have classified the test data of the "RanPil" by the method described above. As shown in Table 3, the test data of "RanPil" contains more data of category 2 and category 3 than the others. This shows that there are many word- level or sentence-level handwritten characters when people write.

Figure 4 shows an example images and label data for each category of test dataset. As can be seen, the letters are clearly understood in the category 1, but it is not the case in the category 2 and 3, the ambiguity is gradually increasing, and in the category 4, no comprehension is possible without a linguistic representation.

For the recognition performance evaluation, we used the TrOCR recognition method proposed in [13]. TrOCR is a recognition model with a combination of encoder and decoder, of which the encoder is ViT, a vision encoder proposed in [15], extracting image features from the text line image, and the decoder is a RoBERTa language model pre-trained on a text corpus.

We trained the training data of the "RanPil" in the TrOCR model and then evaluated the recognition performance in the test data by category.

The performance index is the character error rate (CER). As can be seen in Table 3, CER is 5.3 for category 1 which is the lowest and is 28.5 for category 4 which is the highest. And CER is 7.2, 16.8 for category 2 and 3 respectively. The average CER for the entire test data is 10.93.

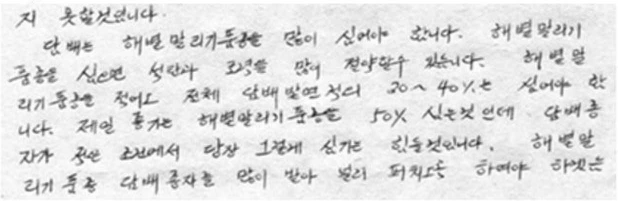
*Table 3. Classification of "RanPil" test data and recognition performance evaluation according to the scrawl-assessment.*

| | Category 1 | Category 2 | Category 3 | Category 4 | Total |
|---|---|---|---|---|---|
| Images | 160 | 486 | 300 | 54 | 1000 |
| CER (%) | 5.3 | 7.2 | 16.8 | 28.5 | 10.93 |

## 4. Conclusions

In this paper, we propose a new benchmark dataset for handwritten Korean text recognition "RanPil", including 8,600 pages of the handwritten document images (182,000 lines and 4,300,000 characters) written by 1,804 people and analyze the writing diversity with the topological characteristics of the writing elements (lines, words, letters) that constitute the handwritten text. Writing diversity in the dataset was evaluated by text lien spacing, text line slope, character size and word spacing, and character compactness. The writing characteristics of people (style of writing) have the greatest impact on recognition performance. We propose a new method of evaluation of the scrawl and then consider the dataset classification method. Finally, the recognition performance in the proposed dataset was evaluated by categories, using TrOCR recognition method. As a future work, we will investigate more realistic benchmark dataset and recognition method for handwritten Korean document recognition.
Image

Image



Label

지 못할것입니다.

담배는 해별말리기폼종을 많이 심어야 합니다. 해별말리기

폼종을 심으면 석탄과 로력을 많이 절약할수 있습니다. 해별말

리기폼종은 적어도 전체 담배발면적의 30~40%는 심어야 합

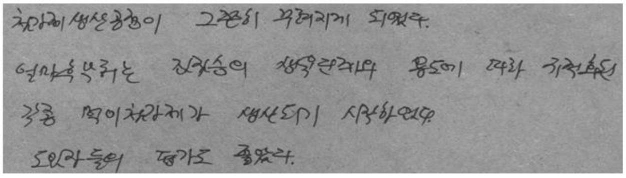
니다. 제일 좋기는 해별말리기폼종을 50%심는것인데 담배종

자가 적은 조건에서 당장 그렇게 심기는 힘들것입니다. 해별말

리기폼종 담배종자를 많이 받아 널리 퍼치도록 하여야 하겠습

a) Example of category 1

Image



Label

첨가제생산공정이 그쯘히 꾸려지게 되였다.

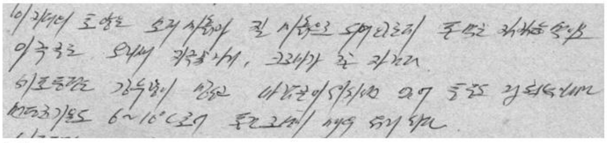
얼마후버터는 집짐승의 생육단계와 용도에 따라 규격화된

각종 먹이첨가제가 생산되기 시작하였다.

도입자들의 평가도 좋았다.

b) Example of category 2

Image



Label

이 지대의 토양은 모래메흙과 질메흙으로 되여있는데 풀먹는 집짐승먹이로

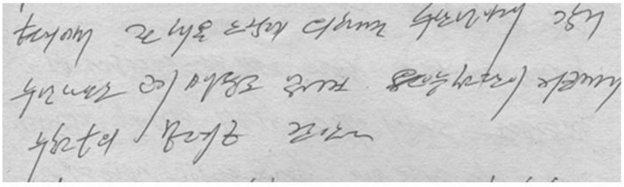
아주 좋은 오리새 자주꽃자리, 크로바가 잘 자란다.

세포등판은 강수량이 많고 바람골에 위치하고있어 통풍도 잘 될 뿐아니라

년평균기온도 6~10°C로서 폴판조성에 매우 유리하다.

c) Example of category 3

Image



Label

현지에서 전해온 소식에 의하면 수산성아래 각지

수산사업소 고기배들로 무어진 원양선단이 지금까지

수천t의 물고기를 잡았다

d) Example of category 4

*Figure 4. Catrgory-specific examples of the "RanPil" test data.*

## Abbreviations

| | |
|---|---|
| MOS | Mean Opinion Score |
| HTR | Handwritten Text Recognition |

## Acknowledgments

The authors would like to express their sincere thanks to the editor and the anonymous referees for their valuable comments concerning the paper.

## Author Contributions

**Hyon-Gwang O**: Project administration, Conceptualization, Resources

**Myong-Chol Kim**: Conceptualization, Methodology, Writing – original draft

**Il-Nam Pak:** Investigation, Methodology

**Un-Hyok Choe:** Software, Validation

**Chol-Jun O:** Writing – review & editing

## Funding

The authors declare that no funds, grants or other supports were received during the perpetration of this research.

## Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Ethics Approval and Consent to Participate

Not applicable.

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

[1] U. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition", International Journal on Document Analysis and Recognition, IJDAR, vol. 5, no. 1, pp. 39–46, 2002.

[2] E. Grosicki, M. Carr´e, J. M. Brodin, and E. Geoffrois, "RIMES evaluation campaign for handwritten mail processing", International Conference on Document Analysis and Recognition. ICDAR, pp. 941–945, 2009.

[3] S. A. Mahmoud et al., "KHATT: Arabic offline handwritten text database," in Proc. Int. Conf. Frontiers Handwriting Recognit. (ICFHR), pp. 449–454, 2012.

[4] T. Su, T. Zhang, D. Guan, "Hit-mw dataset for offline Chinese handwritten text recognition", in: Proceedings of International Workshop on Frontiers in Hand- writing Recognition (IWFHR), Citeseer, 2006.

[5] Y. Zhu, Z. Xie, L. Jin, X. Chen, Y. Huang, and M. Zhang, "Scut-ept: New dataset and benchmark for offline Chinese text recognition in examination paper", IEEE. Access, vol. 7, pp. 370–382, 2018.

[6] H. Zhang, L. Liang, L. Jin, "SCUT-HCCDoc: A new benchmark dataset of handwritten Chinese text in unconstrained camera-captured documents", Pattern Recognition, vol. 108, 2020.

[7] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "CASIA online and offline Chinese handwriting databases," in Proc. Int. Conf. Document Anal. Recognit. (ICDAR), pp. 37–41, 2011.

[8] T. Matsushita and M. Nakagawa, "A database of on-line handwritten mixed objects named 'Kondate,'" in Proc. 14th Int. Conf. Frontiers Handwriting Recognit. (ICFHR), pp. 369–374, 2014.

[9] M. Liwicki and H. Bunke, "IAM-OnDB—An on-line English sentence database acquired from handwritten text on a whiteboard," in Proc. 8thInt. Conf. Document Anal. Recognit., pp. 956–961, 2005.

[10] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, Dinei Florencio, C. Zhang, Z. Li, and Furu Wei. (2021). Trocr: Transformer-based optical character recognition with pre-trained models. [Online]. Available: https://arxiv.org/abs/2109.10282

[11] Denis Coquenet, Clément Chatelain, and Thierry Paquet. (2021)., End-to-end Handwritten Paragraph Text Recognition Using a Vertical Attention Network, [Online]. Available: https://arxiv.org/abs/2012.03868v2

[12] Denis Coquenet et al. (2023). Faster DAN: Multi-target Queries with Document Positional Encoding for End-to-end Handwritten Document Recognition, [Online]. Available: https://arxiv.org/abs/2301.10593v1

[13] Masato Fujitake. (2023). DTrOCR: Decoder-only Transformer for Optical Character Recognition, [Online]. Available: https://arxiv.org/abs/2308.15996v1

[14] Sanchez, J. A., V. Romero, A. H. Toselli, and E. Vidal, "ICFHR2016 competition on handwritten text recognition on the READ dataset", In Proceedings of the International Conference on Frontiers in Handwriting Recognition. ICFHR, pp. 630-635, 2016.

[15] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, [Online]. Available: https://arxiv.org/abs/2010.11929v2

[16] Liu, Y. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach, [Online]. Available: https://arxiv.org/abs/1907.11692