ЖАРАТЫЛЫСТАНУ ЖӘНЕ ТЕХНИКАЛЫҚ ҒЫЛЫМДАР

NATURAL AND TECHNICAL SCIENCES

IRSTI 50.01.85

Mukhtar Bimurat[1*], *Ardak Shalkarbaiuly*[2], *Akhmediyar Kazhymukhanuly*[3],
Aierke Myrzabayeva[4]
[1,2,4]SDU University, Kaskelen, Kazakhstan
[3]Narxoz University, Almaty, Kazakhstan
*e-mail: bimurat.mukhtar@gmail.com

SVTR MODEL FOR KAZAKH HANDWRITTEN TEXT RECOGNITION

Abstract: Handwritten Text Recognition (HTR) plays a crucial role in transforming historical and contemporary handwritten documents into digital formats, facilitating easier access, searchability, and analysis. The SVTR model, known for its state-of-the-art performance in scene text recognition (STR), stands out for its minimal resource use, and quick inference time. In this study, we apply the SVTR model to the Kazakh Offline Handwritten Text Dataset (KOHTD) to assess its capability in handwritten text recognition. Achieving a Character Error Rate (CER) of 4.59% and a Word Error Rate (WER) of 20%, our research establishes new accuracy benchmarks for the KOHTD. The findings underscore the SVTR model's high effectiveness in recognizing handwritten text.

Keywords: optical character recognition, handwritten text recognition, deep learning, KOHTD, SVTR

1. Introduction

The advancement of optical character recognition (OCR) technology has played a crucial role in facilitating the conversion of analog to digital information, especially when handling handwritten writings. Despite significant advancements, recognizing texts in less common languages, such as Kazakh, poses unique challenges due to limited datasets, language-specific characteristics, and lack of research. This study introduces a novel application of the SVTR model in the KOHTD [1] to address these challenges. By exploring this innovative approach, our research aims to push the boundaries of handwritten text recognition accuracy and efficiency, setting new standards for the KOHTD dataset.

1.1 Literature review

For offline handwritten text recognition, Convolutional Recurrent Neural Network (CRNN) models have been widely used due to their ability to process variable-length sequences and their effectiveness in learning spatial hierarchies of features [2]. These models combine convolutional layers for feature extraction with recurrent layers to model sequence dependencies, showing competitive accuracy on standard datasets.

Gated-CNN models, represented by the work on HTR-Flor and HTR-Flor++, offer an alternative to CRNNs by incorporating gated convolutional layers. These models achieve significant improvements over traditional CRNNs in recognition accuracy and computational efficiency, especially on datasets like IAM, demonstrating the potential of Gated-CNNs for offline HTR tasks [3].

The SVTR model revolutionizes scene text recognition by utilizing a Transformer-based architecture to streamline the recognition process, eliminating the need for sequential modeling and complex preprocessing. This method simplifies recognition by breaking down image text into character components and applying a hierarchical approach of component-level mixing, merging, and combining. Progressive overlapping patch embedding enhances feature extraction, while global and local mixing blocks provide a detailed analysis of character components, optimizing for accuracy and computational efficiency [4].

SVTR distinguishes itself with its architectural variants tailored for different computational needs, including SVTR-T (Tiny), SVTR-S (Small), SVTR-B (Base), and SVTR-L (Large). Among these, SVTR-L (Large) stands out for its highly competitive accuracy in English and its superior performance over existing methods in Chinese recognition tasks, coupled with faster processing speeds. On the other end, SVTR-T (Tiny) presents an effective, significantly smaller alternative that offers appealing speed at inference, demonstrating the model's versatility and efficiency across different scales of operation. This blend of high accuracy, speed, and flexibility underscores SVTR's potential to set a new standard in scene text recognition [13].

KOHTD is a large dataset created for research on handwritten text recognition in Kazakhstan. It includes about 922,010 symbols, over 140,335 split images, and 3,000 handwritten test papers. By offering a wealth of resources that encompass a range of common text recognition techniques for word and line recognition, including CTC-based and attention-based techniques, the dataset seeks to support research in HTR tasks [1].

ResNet50+Transformer model was trained using the KOHTD and HKR databases for recognizing handwritten characters in the Kazakh language. The study found that the KOHTD database produced results with a CER of 9.46% and a WER of 20.18%, showcasing its utility in HTR research [5].

Another resource that offers a collection of forms with roughly 95% Russian and 5% Kazakh words/sentences for offline handwriting recognition is HKR for Handwritten Kazakh & Russian Database. With more than 1400 filled-out forms, almost 63000 words, and more than 715699 symbols, it provides researchers studying handwriting recognition with an extra dataset [6].

Employing the KOHTD in conjunction with the HKR database, the SimpleHTR [10] model was trained for the recognition of handwritten Kazakh characters [9]. The investigation produced state-of-the-art outcomes, demonstrating a notably low CER of 2.45% and a WER of 11.09%.

2. Main part

2.1 Framework and Model Selection

PaddleOCR is a comprehensive open-source OCR library developed by Baidu, designed to provide an all-in-one solution for text detection and recognition in images [7]. The framework integrates state-of-the-art algorithms for text detection, such as DB (Differentiable Binarization) [15] and EAST [14], with advanced text recognition models including SVTR.

SVTR (Scene Text Recognition with a Single Visual Model) represents a novel approach in the OCR domain, specifically designed to address the challenges of scene text recognition. Unlike traditional models that rely heavily on complex preprocessing steps and multiple components to detect and recognize text, SVTR utilizes a unified visual model based on the Transformer architecture, which excels at capturing the intricate dependencies in visual data.

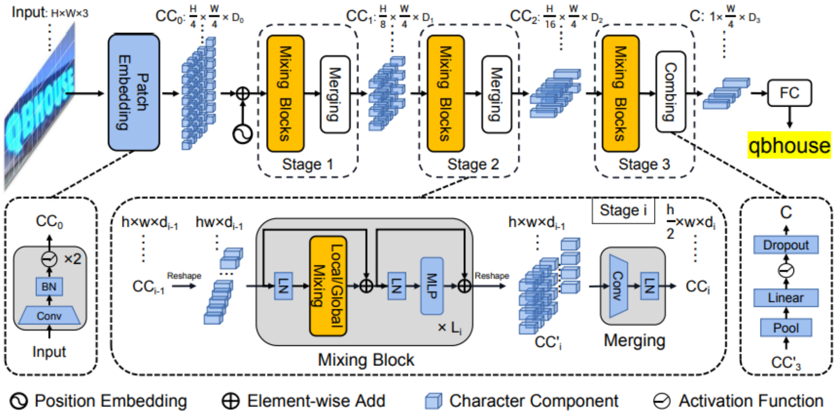


Figure 1. SVTR architecture [4]

The "Tiny" variant of the SVTR model, chosen for this project, is a lightweight version tailored for environments with limited computational resources or for applications requiring fast inference times without compromising accuracy. This makes SVTR-Tiny an ideal choice for the Kazakh Offline Handwritten Text Dataset, as it combines the efficiency and scalability needed for handwritten text recognition tasks.

Models	$[D_0, D_1, D_2]$	$[L_1, L_2, L_3]$	Heads	D_3	Permutation	Params (M)	FLOPs (G)
SVTR-T	[64,128,256]	[3,6,3]	[2,4,8]	192	$[L]_6[G]_6$	4.15	0.29
SVTR-S	[96,192,256]	[3,6,6]	[3,6,8]	192	$[L]_8[G]_7$	8.45	0.63
SVTR-B	[128,256,384]	[3,6,9]	[4,8,12]	256	$[L]_8[G]_{10}$	22.66	3.55
SVTR-L	[192,256,512]	[3,9,9]	[6,8,16]	384	$[L]_{10}[G]_{11}$	38.81	6.07

Table 1. Architecture specifications of SVTR variants [4]

2.2 Dataset

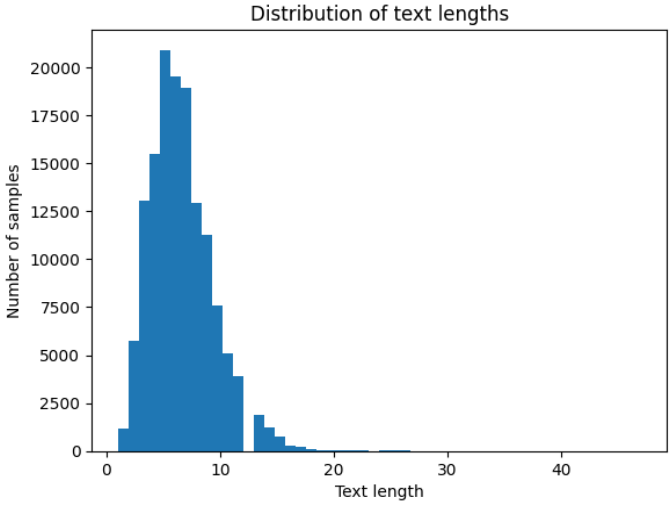
The dataset used in this study is the KOHTD, which is made up of samples of handwritten Kazakh text. Using the original distribution supplied by KOHTD, this dataset is randomly split into 70% training, 15% validation, and 15% test sets. This division makes it possible to conduct a thorough training procedure and a thorough assessment of how the model performs in comparison to the original version.

2.3 Data Analysis

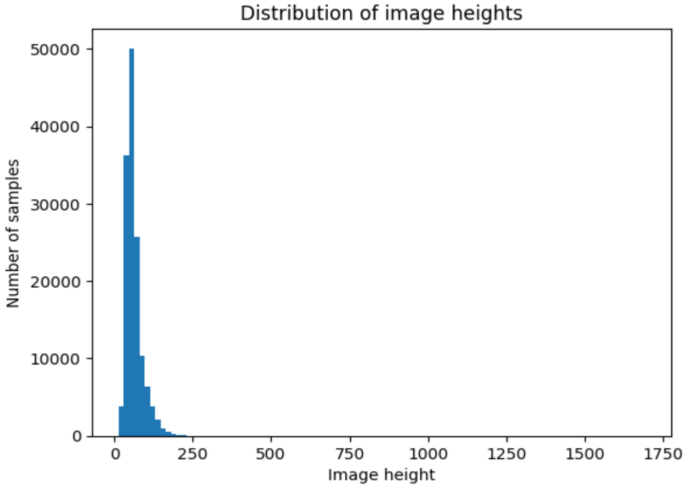
In the course of our research, we conducted an Exploratory Data Analysis (EDA) on the KOHTD dataset. This EDA served as a foundational element for our study, enabling us to fine-tune the hyperparameters of the SVTR model by following official recommendations [8]. The initial dataset contained a total of 140,355 images; however, we excluded 11 images with distorted labels (those containing the newline character '\n'), to parse the dataset using the PaddleOCR framework.

The EDA provided us with valuable statistics, which are summarized in Table 1 below. These descriptive statistics encompass several dimensions: the number of characters per image, image height (px), image width (px), and the aspect ratio (height/width) of each image.

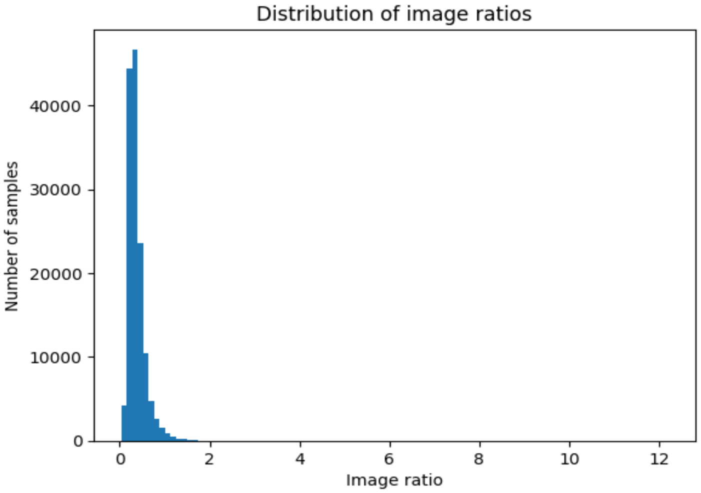
The histogram of text lengths reveals a skewed distribution, with the majority of the text lengths falling below 10 characters. This suggests that shorter texts are more prevalent in the dataset. A significant drop is observed as the text length increases, indicating that longer sequences of characters are less common.



The distribution of image heights is highly concentrated at the lower end of the scale, showing a steep decline as the height increases. This indicates that most images are relatively short in height, with very few images possessing extreme height values.



An analysis of image widths demonstrates a peak at the lower end, followed by a gradual tailing off, which signifies that while a wide range of image widths is present, the dataset predominantly consists of images with narrower widths.



The image aspect ratios exhibit a steep peak at lower values, suggesting that most images have a similar, narrower aspect ratio. There is a long tail, however, extending towards higher ratios, which denotes that there are still a considerable number of images with more elongated shapes.

2.4 Training Process

The training was conducted on an RTX-3060 GPU with 6 GB of memory using CUDA. The SVTR-Tiny model was trained using the PaddleOCR framework. The training process involved several steps, including pre-processing of the dataset to match the PaddleOCR framework format. Setting appropriate hyperparameters, and iteratively training the model to minimize the

CER and WER. The primary objective was to enhance the model's ability to accurately recognize handwritten text in the Kazakh language.

We have trained 100 epochs. The optimizer of choice was Adam, with beta1 and beta2 set to 0.9 and 0.999, respectively. The learning rate was managed by a Cosine scheduler starting at 0.0001 with a warm-up period of one epoch, and regularization was applied via L2 regularization with a factor of 3.0e-05.

For the architecture, we adopted the recognition model type with the SVTR algorithm, utilizing a MobileNetV1Enhance backbone with a scale of 0.5 and average pooling as the last pool type. The head configuration comprised a MultiHead arrangement with both CTCHead and SARHead components. The loss function was a composite MultiLoss, incorporating both CTCLoss and SARLoss components.

2.5 Evaluation Metrics

To assess the effectiveness of the SVTR-Tiny model applied to the KOHTD dataset, we utilized two main evaluation metrics: Character Error Rate (CER) and Word Error Rate (WER).

One popular statistic for assessing the effectiveness of optical character recognition and speech recognition systems is the CER. The total number of character-level errors divided by the total number of characters in the reference is how it is defined. Character-level mistakes are divided into three categories: deletions (D), insertions (I), and substitutions (S). The following is the formula to calculate CER:

$$CER = \frac{S + I + D}{N}$$

where N is the reference text's total character count.

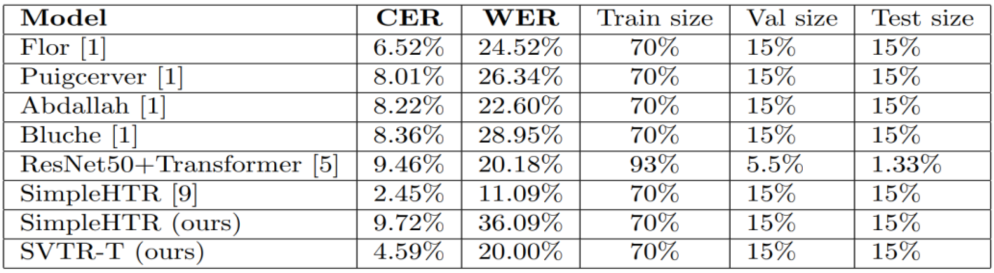
In a similar vein, speech recognition systems' word-level performance is assessed using the WER. It is the ratio of word mistakes to total words in the source. Additions, deletions, and substitutions are also considered word mistakes. The WER can be written as follows:

$$WER = \frac{S_w + I_w + D_w}{N_w}$$

In this formula, S_w, I_w, and D_w correspond to the number of words in the reference, and reflect the number of word swaps, insertions, and deletions, respectively. However, as most photos only contain one word, we evaluated WER as the number of erroneously identified images divided by the total number of images.

2.6 Results

The SVTR-T model showcased impressive results in recognizing handwritten text, achieving a CER of 4.59% and a WER of 20.00%. This performance positions the SVTR-T as a formidable option, surpassing several other evaluated models in accuracy and effectiveness for handwritten Kazakh text recognition. The compared models include Flor, Puigcerver, Abdallah, Bluche, and ResNet50+Transformer, with the SVTR-T demonstrating superior capabilities.



Model	**CER**	**WER**	Train size	Val size	Test size
Flor [1]	6.52%	24.52%	70%	15%	15%
Puigcerver [1]	8.01%	26.34%	70%	15%	15%
Abdallah [1]	8.22%	22.60%	70%	15%	15%
Bluche [1]	8.36%	28.95%	70%	15%	15%
ResNet50+Transformer [5]	9.46%	20.18%	93%	5.5%	1.33%
SimpleHTR [9]	2.45%	11.09%	70%	15%	15%
SimpleHTR (ours)	9.72%	36.09%	70%	15%	15%
SVTR-T (ours)	4.59%	20.00%	70%	15%	15%

Table 3: Recognition Error Rates by Model

Notably, the SimpleHTR model [9] was an exception, outperforming others with a lower CER of 2.45% and a WER of 11.09%. These results, however, were challenging to replicate in our experiments. Despite our efforts to reproduce the SimpleHTR's state-of-the-art results using the same training, validation, and testing sizes, our replication attempts resulted in significantly higher error rates, with a CER of 9.72% and a WER of 36.09%. Table 3 presents a comprehensive comparison, including our attempts to replicate the SimpleHTR model.

3. Conclusion

In this study, we explored the application of the SVTR-Tiny model within the PaddleOCR framework for the KOHTD, aiming to improve OCR technologies for Kazakh text. The results indicate progress in handwritten text recognition, especially for languages like Kazakh that have not been extensively studied in OCR research. Our efforts to replicate the findings of previous studies, particularly the SimpleHTR model's lower error rates, faced challenges, leading to higher error rates in our attempts. However, the performance of the SVTR-Tiny model, designed for environments with limited computational resources, demonstrates the capability of transformer-based models in HTR problems and holds promise for the KOHTD.

To further enhance the results demonstrated by the SVTR-T model in handwritten text recognition, exploring larger variants of the SVTR model presents a promising direction [4]. Larger models, such as SVTR-Small, SVTR-Base, and SVTR-Large, with increased computational capacities and more complex architectures, have the potential to significantly improve the CER and WER by capturing more nuanced features of handwritten text. Additionally, future efforts could focus on optimizing model parameters, exploring advanced training techniques, and increasing dataset diversity to address underrepresented languages more effectively.

References

1 Toiganbayeva, Nazgul, et al. "KOHTD: Kazakh offline handwritten text dataset." Signal Processing: Image Communication 108 (2022): 116827.

2 Geetha, R., T. Thilagam, and T. Padmavathy. "Effective offline handwritten text recognition model based on a sequence-to-sequence approach with CNN–RNN networks." Neural Computing and Applications (2021): 1-12.

3 Arthur Flor de Sousa Neto et al. "HTR-Flor: A Deep Learning System for Offline Handwritten Text Recognition." 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) (2020): 54-61. https://doi.org/10.1109/SIBGRAPI51738.2020.00016.

4 Yongkun Du et al. "SVTR: Scene Text Recognition with a Single Visual Model." (2022): 884-890. https://doi.org/10.48550/arXiv.2205.00159.

5 Y. Amirgaliyev et al. "ResNet50+Transformer: kazakh offline handwritten text recognition." Bulletin of the National Engineering Academy of the Republic of Kazakhstan (2022). https://doi.org/10.47533/2020.1606-146x.150.

6 Nurseitov, Daniyar, et al. "Handwritten Kazakh and Russian (HKR) database for text recognition." Multimedia Tools and Applications 80 (2021): 33075-33097.

7 Li, Chenxia, et al. "PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system." arXiv preprint arXiv:2206.03001 (2022).

8 "High-precision Chinese Scene Text Recognition Model SVTR." AI Studio, Baidu, https://aistudio.baidu.com/projectdetail/5073182?contributionType=1. Accessed 1 Feb. 2024.

9 Тойганбаева, Н. А., et al. "ҚАЗАҚ ТІЛІНДЕГІ ҚОЛЖАЗБА МӘТІНДЕРІНІҢ МӘЛІМЕТТЕР ҚОРЫН ҚҰРУ МЕН ТАНУ." Вестник Ауэс 3.62 (2023).

10 Scheidl, Harald. “Handwritten Text Recognition in Historical Documents.” (2018).

11 Diaz, Daniel Hernandez, et al. "Rethinking text line recognition models." arXiv preprint arXiv:2104.07787 (2021).

12 Kalken, M. "HANDWRITTEN OPTICAL CHARACTER RECOGNITION: IMPLEMENTATION FOR KAZAKH LANGUAGE."

13 Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, Yu-Gang Jiang. "SVTR: Scene Text Recognition with a Single Visual Model." Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022), pp. 884-890. [Online].

14 Zhou, Xinyu, et al. "East: an efficient and accurate scene text detector." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.

15 Liao, Minghui, et al. "Real-time scene text detection with differentiable binarization." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 07. 2020.

Мұхтар Бимұрат[1], Ардак Шалқарбайұлы[2], Ахмедияр Қажымұханұлы[3], Айерке Мырзабаева[4]
[1,2,4]SDU University, Қаскелең, Қазақстан
[3]Narxoz University, Алматы, Қазақсан
*e-mail: bimurat.mukhtar@gmail.com

ҚАЗАҚША ҚОЛЖАЗБА МӘТІНІН ТАНУҒА АРНАЛҒАН SVTR МОДЕЛІ

Аңдатпа. Қолжазба мәтінін тану (HTR) қолжазба тарихи және қазіргі заманғы құжаттарды сандық форматтарға түрлендіруде, қолжетімділікті, іздеуді және талдауды жеңілдетуде шешуші рөл атқарады. Көрініс мәтінін танудағы (STR) ең заманауи өнімділігімен танымал SVTR моделі ресурсты минималды пайдалануымен және өңдеу жылдамдығымен ерекшеленеді. Бұл зерттеуде біз қолжазба мәтінін тану мүмкіндігін бағалау үшін SVTR моделін қазақша офлайн қолжазба мәтіндік деректер жиынына (KOHTD) қолданамыз. 4,59% таңба қателігінің жылдамдығына (CER) және 20% сөз қатесінің жылдамдығына (WER) қол жеткізу арқылы біздің зерттеуіміз KOHTD үшін дәлдіктің жаңа көрсеткіштерін белгілейді. Біздің зерттеуіміз SVTR моделінің қолмен жазылған мәтінді танудағы жоғары тиімділігін көрсетеді.

Түйін сөздер: оптикалық тану, қолжазбаны тану, KOHTD, терең оқыту, нейрондық желілер, KOHTD, SVTR

Мұхтар Бимұрат[1], Ардак Шалқарбайұлы[2], Ахмедияр Қажымұханұлы[3], Айерке Мырзабаева[4]
[1,2,4]SDU University, Каскелен, Казахстан
[3]Narxoz University, Алматы, Қазахсан
*e-mail: bimurat.mukhtar@gmail.com

МОДЕЛЬ SVTR ДЛЯ РАСПОЗНАВАНИЯ КАЗАХСКОГО РУКОПИСНОГО ТЕКСТА

Аннотация. Распознавание рукописного текста (HTR) играет решающую роль в преобразовании исторических и современных рукописных документов в цифровые форматы, облегчая доступ, возможность поиска и анализа. Модель SVTR, известная своей современной производительностью в распознавании текста сцены, отличается минимальным использованием ресурсов и быстрым временем отработки. В этом исследовании мы применяем модель SVTR к казахскому набору данных рукописного текста (KOHTD), чтобы оценить его возможности в распознавании рукописного текста. Достигнув коэффициента ошибок в символах (CER) 4,59% и коэффициента ошибок в словах (WER) 20%, наше исследование устанавливает новые стандарты точности для KOHTD. Результаты подчеркивают высокую эффективность модели SVTR в распознавании рукописного текста.

Ключевые слова: оптическое распознавание, распознавание рукописного текста, глубокое обучение, KOHTD, SVTR