SATBAYEV UNIVERSITY

Computing & Engineering

Volume 2 (2024), Issue 4, 1-7

https://doi.org/10.51301/ce.2024.i4.01

Development of Handwritten Text Recognition system for the Kazakh Language

A. Razaque*, B. Makezhanuly, O. Alimseitov, Zh. Kalpeyeva, A. Ayapbergenova

Satbayev University, Almaty, Kazakhstan

**Corresponding author: a-razaque@onu.edu*

Abstract. The low digitalization of the Kazakh language is a problem that affects bureaucracy efficiency, the accessibility of literature, and education in the Kazakh language. This research introduces a modern approach to handwritten text recognition (HTR) for the Kaz akh language. It optimizes document flow and text mining, increases accessibility to Kazakh literature and historical resources, helps teachers in students' essay scoring, and judges in decision -making. This solution optimizes operational processes in business, education, and government services. The state -of-the-art algorithms are integrated to achie ve improved accuracy and performance of text translation. HTR for the Kazakh language uses effective machine learning (ML) methods to create an HTR system specifically tuned for the Kazakh script. The se leverage features of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), image augmentation, transfer learning, and classic ML methods. HTR is implemented using Python programming language, O penCV, PyTorch, and Scikit- learn libraries. The system was trained on a large dataset of Kazakh handwritten text with different topics.

Keywords: *handwritten text recognition, machine learning, kazakh language, deep learning, convolutional neural networks, recurrent neural networks, character error rate, word error rate.*

1. Introduction

Low digitalization is the biggest issue in the Kazakh language these days. There are different historical reasons for this, like a low number of speakers, colonial past, and distance from technological centers.

During the Soviet Union epoch, the main language in Kazakhstan was Russian. As showed Popova, M. & De Bot, K. (2020) it negatively affected the popularity of the Kazakh language [1]. All documents, including professional and cultural literature in that period, were written in Russian. A low number of materials written in Kazakh is why the Kazakh language is considered a low-resource language. From Tatineni, S. (2020) research we know that low- resource languages nowadays struggle with innovations, because of a lack of datasets and financial outcomes from its digitalization [2].

As a result of the low digitalization of the Kazakh language, it is not very popular in science and culture in Kazakhstan. Historical materials written in Kazakh belong to the early Soviet period and earlier. Such literature is not open for people, because it exists only in handwritten form. Document flow in Kazakhstan is slow because some documents need to be duplicated in handwritten and digital forms. As showed Turganbayeva, A., & Tukeyev, U. (2020, March), cultural and linguistic barriers often hinder collaboration between linguists, technologists, and native speakers, exacerbating the problemof digital exclusion [3].

Low digitalization reduces the quality of translation services, intelligent assistants, and speech recognition systems in the Kazakh language, hindering integration into global digital platforms. Additionally, as showed Bogdanchikov, A. et al. (2022) it complicates working with large datasets necessary for scientific and commercial tasks [4].

These problems can be solved by HTR. It translates text written by hand into digital form.

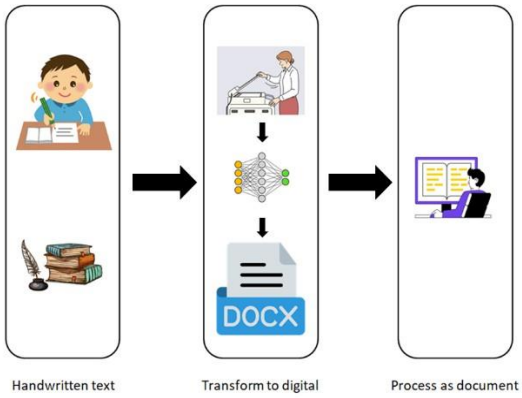

Figure 1. Process of handwritten text to digital document transformation

A. Research Novelty and Contribution

The main contributions of this paper can be summarized as follows. A deep-learning-based model is used to translate handwritten text to digital form. It seeks to reduce the time needed to process documents, like applications, statements, and essays, optimizing operational processes in business, education, and government services. The state-of-the-art algorithms are integrated to achieve improved accuracy and performance of text translation. The first algorithm is a CNN to encode the image. The second algorithm is a transformer algo-

rithm for finding the best matching word for an image. The integration accomplishes better effectiveness. HTR Model for Kazakh Language: The proposed deep-learning- based model is used for handwritten Kazakh text efficient recognition. It works through analysis of the handwritten characters and word structure with advanced machine- learning techniques to convert them into digital text. Considering the limitations of traditional Optical Character Recognition (OCR) methods, this model is designed for fast and accurate text recognition, aiming to enhance the overall efficiency and responsiveness of document processing systems in business, education, and government services. The model design utilizes state-of-the-art algorithms, integrating CNN for image encoding and RNN for sequence learning. This integration improves accuracy and performance, allowing the system to recognize handwritten Kazakh text highly efficiently. The model trained using a dataset containing various handwritten texts in Kazakh shows significant improvements over existing OCR models.

B. Problem Identification and Significance

In recent years, the low digitalization of the Kazakh language has become one of the most significant challenges, as it hinders access to educational resources, historical literature, and government services. The core issue is that the Kazakh language, unlike more widely used languages, lacks efficient tools for processing handwritten text. The problem makes it hard to digitize handwritten documents, such as essays, official statements, and historical manuscripts, which are still predominantly handwritten. As a result, the document management process, archiving, and retrieval remain slow and inefficient. Traditional methods for digitizing handwritten text, such as OCR, often fall short when applied to Kazakh due to the unique nature of its script and the scarcity of annotated datasets. The limited availability of digital resources and the absence of specialized tools for Kazakh handwriting recognition have made it hard to adapt general-purpose OCR systems to process Kazakh text effectively.

Additionally, existing systems struggle with accurately recognizing various handwriting styles, leading to higher error rates and slower processing speeds. One solution that has been explored involves using image-based approaches to enhance handwriting recognition. However, these methods often face difficulties in distinguishing the unique characters and structures of the Kazakh language. Another approach focuses on the use of machine learning techniques, such as CNNs and RNNs, to improve recognition accuracy. While these methods show promise, they often require substantial datasets, which are difficult to compile for low-resource languages like Kazakh. This paper proposes a machine learning-based solution for handwritten Kazakh text recognition. By leveraging advanced deep learning algorithms, such as CNNs and RNNs, the proposed model aims to improve text recognition accuracy significantly. This model is designed specifically to handle the unique characteristics of Kazakh handwriting, offering a scalable solution for the real-time processing of handwritten documents. By analyzing the structure and patterns in handwritten text, the model is expected to perform more efficiently than traditional methods, helping overcome the challenges of low-resource language processing.

C. Paper organization

The structure of this paper is organized as follows: Section 2 states the nature of the problem and emphasizes its importance. In Section 3, we review relevant studies, and in Section 4 we present our proposed approach. Section 5 details the experimental results and implementation features. In Section 6, we discuss and analyze the results in detail and draw conclusions.

2. Materials and methods

2.1. Related work

There are some approaches in this topic.

Table 1. Contemporary work of existing methods with features and limitations

Methods	Solutions	Advantages	Limitations
Amirgaliyev Y. et al.	17-layers CNN	Simple architecture, can be trained using unconventionally parallel method.	Language model layers weren't used. It affects to high WER score.
Amirgaliyev Y. et al.	ResNet50 CNN model as an encoder and Transformer model as a decoder	Model can predict word based on context	High computational costs
Jantayev R. et al.	CNN-LSTM model	Fast character parsing	Low accuracy
Yeleussinov A. et al.	GAN model	Trained on the larger dataset generated by GAN models	Dataset is synthetic
Bazarkulova, Aisaule et al.	CNN and Bi-directional LSTM	High accuracy	Model trained and validated on dataset of technical texts, high computational costs
Kalken, M.	Model with Connectionist Temporal Classification Loss	Relatively low computational costs	Model trained and validated on dataset of technical texts
Our solution	CNN-Transformers	Trained on a large dataset gathered from different sources	High computational costs

The model developed by Amirgaliyev Y. et al. (2020) [5] is based on 17-layer CNN. LeakyReLU used the activation function to prevent the model from vanishing gradients problem. For segmenting an image into individual characters is used algorithm segmenting a binary image of a grid of vertical and horizontal histograms. This model scores 86.47% precision and 80.65% recall.

Amirgaliyev Y. et al. (2022) [6] introduces another approach to solving this problem. This method used the ResNet50 CNN model as an encoder and the Transformer model as a decoder. The dataset used in this research consisted of mixed Russian and Kazakh words. This approach scored CER-9.46% and WER-20.18% on the KOHTD database.

Jantayev R. et al. (2021) [7] introduce advanced methods for Optical Character Reading (OCR). VGG-19 model was used as a Start-of-Line (SOL) finder. The model produces five-dimensional vectors as an output, namely, the coordinates of the starting point of a line, the direction angle of the text line, the scale, and a probability of occurrence. Then customized model predicts the position of the next character based on cut images. HTR based on 18-layer CNN and Bidirectional LSTM (Long-Short Term Memory) using the proposed OCR algorithmscores CER 12%.

Another method introduced by Yeleussinov A. et al. (2023) [8] is to use a generative adversarial network (GAN). GANs can learn deep concepts without extensive use of labeled learning. In three experiments authors use TextStyleBrush, GanWriting, and ScrabbleGAN architectures as generators. ScrabbleGAN achieved the best scores in Inception Score (IS), Frechet Inception Distance (FID), and Structural Similarity Index Metric (SSIM). These metrics evaluate the similarity between generated and real images. The usage of the extended dataset with generated images helped to improve WER and CER scores 1.5 times. This method scored accuracy-78.17% and CER-17.11%.

In research provided by Bazarkulova, Aisaule et al. [9] for HTR used a model combining CNN and Bi-directional LSTM (long-short term memory) as a language model. The model includes 5 CNN layers increasing shape from 32 to 512. This approach scored WER-13.64% and CER-6.27%.

Kalken, M. (2021) introduces a model with Connectionist Temporal Classification Loss. As an encoder, this model uses 5 CNN layers and as a decoder, it uses 2 LSTM layers. Binarization, noise removal, thinning, and skeletonization were used as preprocessing methods. This model scored 85% accuracy and losses of 1.5 – 2 loss [10].

All these approaches use small language models. Their results can be improved by new NLP approaches such as Bidirectional Encoder Representations from Transformers (BERT) and Regional Proposal Networks.

2.2. Proposed Deep-Learning Model for Handwritten Text Recognition of Kazakh Language

HTR for the Kazakh language is a critical task aimed to address the challenges of low digitalization and accessibility of handwritten documents. Unlike traditional OCR systems, which primarily work with printed text, HTR focuses on identifying patterns in handwriting, which is inherently more variable and complex. Handwritten texts often serve as the medium for historical documents, educational resources, and official statements, making its digitization crucial for accessibility and preservation. This paper proposes a machine learning-based solution for HTR using a combination of advanced deep learning algorithms. This approach is designed to handle the unique characteristics of the Kazakh script and recognize diverse handwriting styles. The proposed system integrates CNNs for feature extraction from text images and RNNs for sequence learning. These components enable the model to accurately process and convert handwritten Kazakh text into digital format.

- Parameter Identification and Data Preprocessing;
- Learning Process and Performance Analysis.

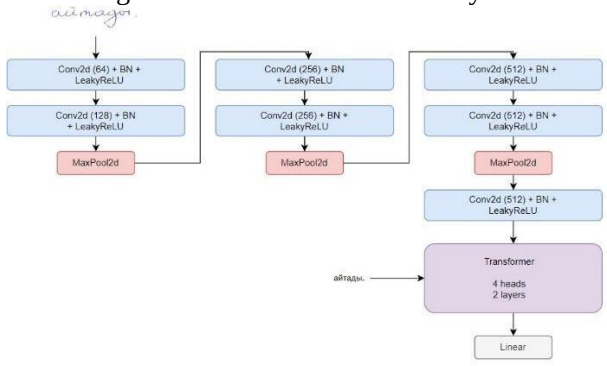

Figure 2. Architecture of Transformer-Based Handwriting Recognition Solution

A. Parameter Identification and Data Preprocessing

In order to achieve accurate recognition of handwritten Kazakh text, several crucial steps need to be taken, including data collection, feature extraction, and data preprocessing. These stages involve preparing the input data for the model, extracting meaningful features from handwritten texts, and ensuring the data is in an optimal format for training the machine learning model.

Algorithm 1 Extracting Key Features from Handwritten Text

Input: u (handwritten image).
Output: Feature set k = [x1, x2, ..., xn].
Extract features:
h(u) (handwriting), d(u) (Kazakh-specific), structural features.
Populate k:
x1: Text length.
x2: Kazakh diacritics (true/false).
x3: Line spacing consistency.
x4: Curvature matches Kazakh letters. x5:
Word segmentation clarity.
Add additional features (xn) based on script consistency and alignment.
Return: k.

B. Data Collection and Image Preprocessing

The first step involves data collection and image preprocessing. A dataset of handwritten Kazakh text is gathered from various publicly available sources. The dataset contains a diverse range of handwriting styles to make the model more robust and adaptable to different writing styles. We used different sources and gathered over 140,000 photos of handwritten text in the Kazakh language for data collection. OpenCV was used to remove noise using filters, and the rembg library was applied for background removal. The Image Enhance function from the Pillow library was used to enhance the brightness of the images.

C. Feature Extraction and Segmentation

The next step after data preprocessing is feature extraction and segmentation of the text. This process is crucial for identifying the key elements of the handwritten text, such as characters and words, and analyzing their structure.

1. Character-Level Features: The model analyzes individual character contours, stroke width, curvature, and orientation.

2. Word-Level Features: Spacing between words and alignment within the text are important to understand the relationship between neighboring characters.

3. Kazakh Script-Specific Features: Given the uniqueness of the Kazakh language, features such as diacritics, special characters, and letter placement are considered to improve recognition.

4. Using the Counter class from the collection's module in Python, we count the occurrences of elements.

D. Data Augmentation

Data augmentation is used for further improvement of the model robustness. This process generates new samples by applying transformations to the original data. At the Data Augmentation stage, the Augmentor library was used to transform and expand the original dataset of handwritten text images to improve the model's training process. After training the model, we applied the edit distance library to evaluate its performance on the augmented dataset. It helped to assess

how different augmentation techniques impacted the CER and WER metrics.

Lemma: Handwritten Kazakh text exhibits specific features that distinguish it from other languages and handwriting styles, which can be identified by analyzing character structures, spacing, and diacritical marks.

Proof: Handwritten Kazakh text often includes unique script features that differ from other languages. Character Shapes: Kazakh uses specific characters, such as "ə", "ғ", "ө", "қ", "і", "ғ", "ұ" and "ң" which have unique structures and often include diacritical marks.

Corollary: If the handwritten text contains specific Kazakh diacritical marks, such as "ə", "ғ", "ө", "қ", "і", "ғ", "ұ" and "ң" it is highly likely to be Kazakh script.

Proof: Kazakh diacritical marks are not found in most other languages. For example, the letter "қ" includes a diacritical element distinguishing it from its Russian counterpart "к." If a model is trained to detect such diacritical marks, it can reliably classify the script as Kazakh.

Corollary: If the structure of handwritten text includes consistent stroke patterns and spacing, it is likely to be legible and suitable for segmentation.

Proof: The structure of Kazakh handwriting often exhibits consistent spacing between characters and words, which facilitates segmentation.

E. Learning Process and Performance Analysis

The proposed HTR model for Kazakh script was developed using a deep-learning approach with a combination of CNNs and RNNs. The learning process involves multiple stages, including data preprocessing, model training, and evaluation of performance metrics.

Algorithm 2 Training Handwritten Text Recognition Model

Input: d (dataset), e (epochs), lr (learning rate), b (batch size).
Output: m_trained.
Split d into d_train, d_val, d_test.
Initialize model m (CNN + RNN), optimizer (Adam), and loss function (CrossEntropy).
pred = m(val_x)
predicted = torch.argmax(pred, dim=1) correct += (predicted == val_y).sum().item() total += val_y.size(0)
Compute v_a = (correct / total) * 100 Check:
If v_a > 90%
Save model: torch.save(m.state_dict(), 'm_trained.pth') Else
Adjust lr: lr = lr * 0.5
optimizer = torch.optim.Adam(m.parameters(), lr=lr) Restart training
End If
Evaluate on test set:
Return: m_trained.

Learning Process: The training consisted of feeding the preprocessed dataset of handwritten Kazakh text into the proposed model. The process was divided into the following steps:

Model Initialization: The CNN component was used for feature extraction from input images, capturing essential features such as edges, contours, and diacritical marks.

The RNN component (specifically Bidirectional LSTM) was employed to learn the sequential patterns in characters and words.

Training: The model was trained using an augmented dataset to handle variations in handwriting styles. Hyperparameters were optimized, including the learning rate, batch size, and the number of layers. The Adam optimizer was utilized for efficient convergence.

Validation: A portion of the dataset (20%) was reserved for validation to monitor overfitting and adjust the training process as needed.

The performance of the model was evaluated using the following metrics: Character Error Rate (CER): Measures the percentage of incorrect characters recognized by the model.

Implementation Features: The model was implemented using Python and PyTorch libraries. Key techniques, such as data augmentation and transfer learning, were incorporated to enhance performance and adaptability. The system's scalability allows it to handle larger datasets and adapt to new handwriting styles.

Hypothesis-1: The deep learning model trained on a diverse dataset of handwritten Kazakh text will outperform simpler models (e.g., traditional OCR systems or rule-based algorithms) in terms of accuracy and Character Error Rate (CER).

Proof: Let $f_{dl}(a)$, $f_{ocr}(a)$ and $f_{rb}(a)$ represent the predicted output of the Deep Learning model, traditional OCR system, and rule-based algorithm, respectively, for an input a (where a corresponds to features extracted from handwritten text).

The accuracy of each model is given by:

$$accuracy\ f(a) = 1/m \sum_{k=0}^{m} \mathrm{II}(f(a_i) - b_i) \quad (1)$$

where m is the number of test samples, a_i is the input, b_i is the true label $\mathrm{II}(\cdot)$ is the indicator function.

The character error rate(cer) is defined as:

$$cer = \text{substitutions+ deletions+ insertions} / total\ characters$$

By comparing the performance metrics (accuracy and cer) of $f_{dl}(a)$, $f_{ocr}(a)$ and $f_{rb}(a)$ on the same test set,we can confirm that $fdl(a)$ consistently outperforms the other models,proving the hypothesis.

Lemma-1: Deep learning models (e.g., CNNs + RNNs) minimize overfitting compared to traditional OCR systems for handwritten Kazakh text recognition.

Proof: Let N_{DL} and N_{OCR} represent the loss functions for the Deep Learning model and the OCR system, respectively.

The loss function for the deep learning model is typically the CTC (Connectionist Temporal Classification) loss:

$$N_{DL} = -\frac{1}{m}\sum_{f=1}^{m} \log(H(b_i|a_i)) \quad (2)$$

where $H(b_i|a_i)$ is the probability of predicting b_i given a_i.

The OCR system uses simpler loss functions without temporal alignment.

$$N_{OCR} = -\frac{1}{m}\sum_{f=1} \log(f_{OCR}(a_i) \neq b_i) \quad (3)$$

Deep learning models reduce overfitting by incorporating techniques like dropout and batch normalization, which regulate the training process. Empirically, N_{DL} exhibits lower validation loss and better generalization compared to N_{OCR}, proving the lemma.

Corollary: The use of advanced data augmentation

techniques in deep learning models reduces variance and improves accuracy in handwritten Kazakh text recognition.

Proof: Data augmentation creates synthetic variations of training data (e.g., rotated, skewed, or scaled images), increasing diversity. Let B_{orig} be the accuracy of the model trained on the original dataset, and B_{aug} be the accuracy of the model trained with augmented data.

Empirically:

$$B_{aug} > B_{orig} \quad (4)$$

This improvement is due to the model's ability to generalize better unseen variations in handwriting styles.

Hypothesis-2: The combination of CNNs and RNNs in the model architecture improves performance in recognizing handwritten Kazakh text compared to standalone CNN or RNN architectures.

Proof: Let $f_{cnn+rnn}\ a$, $f_{cnn}\ a$ and $f_{rnn}\ a$ represent the outputs of the combined CNN-RNN architecture, standalone CNN, and standalone RNN, respectively.

The CER and Word Error Rate (WER) are defined as metrics for evaluation:

Empirical results show that:

$cer(f_{cnn+rnn}(a)) < cer(f_{cnn}(a))$, $wer(f_{cnn+rnn}(a)) < wer(f_{cnn}(a)$

The combined architecture leverages the feature extraction capabilities of CNNs and the sequential modeling of RNNs, improving overall recognition accuracy.

Corollary-2: Early stopping during model training prevents overfitting and ensures better generalization in handwritten text recognition tasks.

Proof: Let the validation loss at iteration t be $N_{val}(t)$.Early stopping halts training when:

$$Nval(\mathrm{t}) - Nval(\mathrm{t}-1) < €, \forall t > Tpatience \quad (6)$$

where € is a small threshold, and $T_{patience}$ is the patience parameter. If training continues beyond $T_{patience}$, the model risks learning noise, leading to overfitting. Early stopping ensures the model generalizes better, as validated by lower test CER and WER scores.

3. Results and discussion

3.1. Experimental results

This section provides experimental setup, dataset description, and results.

A. Experimental Setup

The proposed handwritten text recognition (HTR) model for the Kazakh language was implemented using Python and the PyTorch deep learning framework. The experiments were conducted on a system with the following specifications:

Table 2. Parameters and their description for conducting the experiments

Name of the parameters	Description of the parameters
Programming Language	Python (version 3.8.1, 64 bits)
Development Environment	Google Colab

The training and evaluation were performed on a dataset consisting of diverse samples of handwritten Kazakh text.

The dataset was divided into three subsets:

Training Set: 80% of the data, augmented with transformations such as rotation, scaling, and brightness adjustment to handle variations in handwriting styles.

Test Set: 20% of the data, used for final evaluation.

The model architecture combines CNNs for feature extraction and RNNs with Bidirectional LSTM layers for sequence learning. The CTC loss function was used to handle unsegmented input-output mapping. The Adam optimizer was employed with an initial learning rate of 0.001 and a batch size of 64.

B. Dataset Description

The dataset used for training, validation, and testing of the HTR model for the Kazakh language was carefully compiled to include a diverse range of handwriting styles, document types, and script structures. This diversity ensures the robustness and adaptability of the model across various real-world scenarios.

Data Collection: The dataset was gathered from multiple sources to represent a wide variety of handwriting styles and content types:

Historical Documents: Scanned images of open manuscripts and archival materials in Kazakh.

Educational Samples: Essays, notes, and assignments written by students in Kazakh, reflecting modern handwriting styles.

Out of over 140,000 collected images, 16,500 images were selected and used for training, validation, and testing, ensuring high-quality and representative data.

Table 3. Dataset Statistics

Subset	Number of Samples	Percentage
Training Set	13,200 images	80%
Test Set	3,300 images	20%

The statistical summary the data provided: Density Distribution of Symbols:

Max length of expression: 21 characters.

The most common character: 'a' (378 occurrences) The least common character: 'Ф' (1 occurrence)

The most common expression: 'және' (11 occurrences) The least common expression: 'құруға' (1 occurrence).

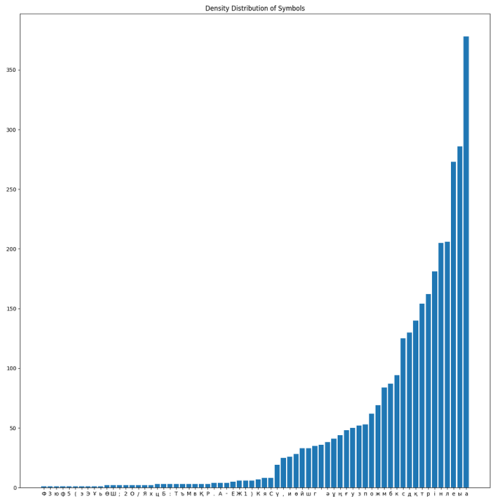

Figure 3. Density Distribution of Symbols

Predictions in the HTRProject (Kazakh Language):

In this project, the predictions refer to the model's ability to recognize handwritten Kazakh characters and words accurately. The primary goal is to predict the most likely character or word based on input data (i.e., the images of handwritten text).

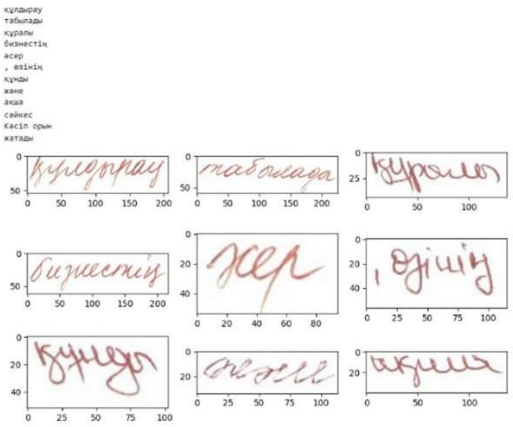

Figure 4. Predicting characters and words in Kazakh handwritten text

A. Results

The performance of the model was evaluated using the following metrics:

Character Error Rate (CER): Measures the percentage of incorrect characters.

Word Error Rate (WER): Measures the percentage of incorrect words.

Table 4. Summarizes the results of the proposed model

Metric	Value
Character Error Rate (CER)	
Word Error Rate (WER)	

The proposed system for handwritten text recognition in the Kazakh language combines CNNs for feature extraction with RNNs that include Bidirectional LSTM layers for sequence learning. The model was trained using the CTC loss function, which is ideal for handling unsegmented input-output mapping, making it particularly well-suited for the challenges posed by handwritten text recognition.

The use of CNNs for feature extraction allows the model to effectively capture spatial hierarchies in the handwritten characters, while the RNN with Bidirectional LSTM layers enables the system to learn the dependencies between characters in a word, both forward and backward, which is crucial for accurately recognizing context and sequences in the handwritten text.

The system's high adaptability allows it to handle a variety of handwriting styles, demonstrating the model's robustness. The integration of the CTC loss function has proven to be especially effective in scenarios without clear boundaries between characters, further enhancing the model's ability to handle complex, unsegmented text.

4. Conclusions

In conclusion, this research introduces a robust system for handwritten Kazakh text recognition, leveraging advanced machine learning techniques, including CNNs and RNNs with Bidirectional LSTM layers. The use of the CTC loss function allows the system to effectively handle unsegmented input, making it well-suited for real-world applications where text is not clearly separated.

This solution is poised to improve text processing in various fields such as education, archival systems, and document automation.

A. Future Work

Future research will focus on further optimizing the model's performance by experimenting with deeper architectures or alternative loss functions to improve recognition in more complex scenarios. Additionally, real- time processing capabilities could be explored to ensure the system can handle continuous input efficiently. The integration of explainable AI techniques would provide transparency and trust in the decision-making process, helping users understand how the system arrives at its predictions. Furthermore, exploring the application of this model in multilingual settings could broaden its scope and enhance its utility across different languages and scripts.

References

[1] Popova, M. & De Bot, K. (2020). Maintenance of the Russian language in Kazakhstan: activity of Russia. *Alkalmazott Nyelvtudomány, 20(2)*

[2] Tatineni, S. (2020). Deep Learning fo r Natural Language Processing in Low-Resource Languages. *International Journal of Advanced Research in Engineering and Technology (IJARET), 11(5), 1301- 1311*

[3] Turganbayeva, A. & Tukeyev, U. (2020, March). The Solution of the Problem of Unknown Words Under Neural Machine Translation of the Kazakh Language. *In Asian Conference on Intelligent Information and Database Systems, Singapore: Springer Singapore*

[4] Bogdanchikov, A. (2022). Classification of scienti fi c documents in the Kazakh language using deep neural networks and a fusion of images and text. *Big Data and Cognitive Computing, 6(4), 123*

[5] Amirgaliyev, Y. (2020). Kazakh handwritten recognition

[6] Amirgaliyev, Y. (2022). ResNet50+Transformer: kazakh offlin e handwritten text recognition

[7] Jantayev, R. (2021). Complete kazakh handwritten page recognition using start, follow and read method

[8] Yeleussinov, A. (2023). Improving OCR Accuracy for Kazakh Handwriting Recognition Using GAN Models. *Applied Sciences, (13), 5677.* https://doi.org/10.3390/app13095677

[9] Bazarkulova, A. (2023). Kazakh handwriting recognition

[10] Kalken, M. (2021)7 Handwritten optical character recognition: implementation for kazakh language

Қазақ тілінде қолжазба мәтінді тану жүйесін әзірлеу

А. Разақ*, Б. Макежанұлы, О. Әлімсеитов, Ж. Қалпеева, Ә. Аяпбергенова

Satbayev University, Алматы, Қазақстан
**Корреспонденция үшін автор: a-razaque@onu.edu*

Аңдатпа. Қазақ тілін цифрландырудың төмен дәрежесі бюрократияның тиімділігіне, қазақ тіліндегі әдебиет пен білімнің қолжетімділігіне әсер ететін проблема болып табылады. Бұл зерттеуде қазақ тіліне арналған қолжазба мәтінін (HTR) танудың заманауи тәсілі ұсынылған. Ол құжат айналымы мен мәтінді талдауды оңтайландырады, Қазақ әдебиеті мен тарихи ресурстардың қолжетімділігін арттырады, оқытушыларға оқушылардың эсселерін бағалауға, ал судьяларға шешім қабылдауға көмектеседі. Бұл шешім бизнестегі, білім берудегі және мемлекеттік қызметтердегі операциялық процестерді оңтайландырады. Мәтінді аударудың дәлдігі мен өнімділігін арттыру үшін ең заманауи Алгоритмдер біріктірілген. Қазақ тіліне арналған HTR-де қазақ әліпбиіне арнайы бейімделген HTR жүйесін құру үшін машиналық оқытудың (ML) тиімді әдістері қолданылады. Se конволюциялық нейрондық желілердің (CNN), қайталанатын нейрондық желілердің (RNN), кескіндерді үлкейту, тасымалдауды үйрену және классикалық ML әдістерінің мүмкіндіктерін пайдаланады. HTR Python бағдарламалау тілі, oxv кітапханалары, PyTorch және Scikit - learn көмегімен жүзеге асырылады. Жүйе әртүрлі тақырыптағы қазақ қолжазба мәтінінің үлкен деректер жиынтығында оқытылды.

Негізгі сөздер: *қолжазба мәтінді тану, машиналық оқыту, қазақ тілі, терең оқыту, конволюциялық нейрондық желілер, қайталанатын нейрондық желілер, таңбалардағы қателер жиілігі, сөздердегі қателер жиілігі*.

Разработка системы распознавания рукописного текста на казахском языке

А. Разақ*, Б. Макежанұлы, О. Әлімсеитов, Ж. Қалпеева, Ә. Аяпбергенова
Satbayev University, Алматы, Казахстан
**Автор для корреспонденции: a-razaque@onu.edu*

Аннотация. Низкая степень цифровизации казахского языка является проблемой, которая влияет на эффективность бюрократии, доступность литературы и образования на казахском языке. В этом исследовании представлен современный подход к распознаванию рукописного текста (HTR) для казахского языка. Оно оптимизирует документооборот и анализ текста, повышает доступность казахской литературы и исторических ресурсов, помогает преподавателям в оценке эссе учащихся, а судьям - в принятии решений. Это решение оптимизирует операционные процессы в бизнесе, образовании и государственных службах. Для повышения точности и производительности перевода текста интегрированы самые современные алгоритмы. В HTR для казахского языка используются эффективные методы машинного обучения (ML) для создания системы HTR, специально адаптированной для казахского алфавита. В se используются возможности сверточных нейронных сетей (CNN), рекуррентных нейронных сетей (RNN), увеличения изображений, обучения переносу и классических методов ML. HTR реализован с использованием языка программирования Python, библиотек OXV, PyTorch и Scikit- learn. Система была обучена на большом наборе данных казахского рукописного текста различной тематики.

Ключевые слова: *распознавание рукописного текста, машинное обучение, казахский язык, глубокое обучение, сверточные нейронные сети, рекуррентные нейронные сети, частота ошибок в символах, частота ошибок в словах*.

Received: 19 July 2024
Accepted: 16 December 2024
Available online: 31 December 2024