ABSTRACT

of the dissertation by Toiganbayeva Nazgul Abenovna for the degree of Doctor of Philosophy (PhD) in the educational program 6D060200 – Informatics on the topic « Development of tools and creation of a database for optical recognition of handwritten text of the Kazakh-Russian language»

Relevance of the Research Topic. Despite the widespread transition to digital data exchange across various domains, many documents—such as questionnaires, memos, examination responses, and archival records—still require manual data entry. This underscores the necessity of implementing handwriting recognition, an automated method for transcribing handwritten content using computers. Handwriting recognition (HWR) or Handwritten Text Recognition (HTR) refers to the ability to capture and interpret legible handwritten information from paper documents, images, touchscreens, and other sources.

Offline handwriting recognition is a method of converting letters or words in an image into digital text, where the input is a two-dimensional image and the output is a sequence of characters. Currently available software solutions for handwriting recognition can be divided into two main types: Online recognition, which involves recognizing text written on device screens using specialized tools (e.g., styluses); Offline recognition, which involves the automatic conversion of text from an image into character codes that can be processed and utilized by computer systems and text processing applications. The task of offline handwriting recognition has not yet been fully solved. Over the past decade, substantial progress has been achieved in the recognition of both printed and handwritten text. Despite computational complexity, a number of highly challenging problems can now be addressed through architectures developed for machine learning and deep neural networks. However, the problem of recognizing handwritten Cyrillic text in Kazakh and Russian remains insufficiently studied. Therefore, the development of algorithms for offline recognition of handwritten text in Kazakh and Russian is of critical importance and requires dedicated research. This is particularly relevant given the absence of publicly available datasets for handwritten Kazakh text. Thus, the topic is contemporary and has significant practical value.

Handwriting recognition is a complex, structured process that includes scanning paper documents containing text, separating the acquired content into textual and graphical components, selecting appropriate methods for intelligent handwriting recognition, and processing the results. Neural networks are applied to solve handwriting recognition tasks. Inspired by the biological structure of the human brain, neural networks—due to their high computational capabilities—are considered several times more effective than other learning algorithms.

In recent years, significant progress has been made in improving recognition methods and systems; however, automatic recognition of handwritten text in images remains a challenging task. Comprehensive and unconstrained datasets containing handwriting samples have become more important than ever. Handwritten text is

encountered in memos, medical records, historical documents, stylus-entered text, examination responses, and other forms of documentation. Therefore, a complete solution for the problem of handwritten text recognition from images is required. This highlights the necessity of research aimed at developing large-scale handwriting recognition systems for multiple languages. Consequently, the creation of a database and the development of efficient OCR tools for handwritten Kazakh and Russian text is a timely challenge of both scientific and high practical significance.

The aim of the dissertation is to develop tools for optical recognition of handwritten texts in Kazakh and Russian using deep learning models, as well as to create a corresponding database.

Research objectives. To achieve the stated aim, the following objectives must be accomplished:

1. Conduct a review and analysis of solutions employing deep learning methods for optical text recognition tasks.
2. Compile a dataset (corpus) of handwritten texts in Kazakh and Russian.
3. Apply a genetic algorithm to implement the segmentation process that divides text on scanned pages into lines and lines into words.
4. Perform experiments using various deep learning methods based on the handwritten text corpus in Kazakh and Russian.

Object of the research – the process of recognizing handwritten texts in Kazakh and Russian.

Subject of the research – the creation of a corpus of handwritten texts in Kazakh and Russian and the application of deep learning models for their recognition.

Research methods – a genetic algorithm is used for the segmentation of handwritten texts, while recurrent and convolutional neural networks are employed to solve the task of handwriting recognition.

Theoretical and practical Significance of the Work. The digitalization of handwritten records enables the automation of business processes in many companies and reduces manual workload. Furthermore, the recognition of handwritten texts in Kazakh and Russian can facilitate the automation of postal center operations. The Kazakh Offline Handwritten Text Dataset (KOHTD), created for the first time for the Kazakh language, provides an opportunity to evaluate and compare the performance of algorithms.

Main provisions submitted for defense:

- A fully automated system has been developed and experimentally implemented, performing the processes of collecting handwritten text data, their preprocessing, annotation, and inclusion in the corpus.

- A genetic algorithm has been adapted for accurate word segmentation in handwritten Kazakh texts, which has improved the precision of determining inter-character boundaries through its heuristic capabilities.

- An open dataset – Kazakh Offline Handwritten Text Dataset (KOHTD) – containing offline handwritten texts in Kazakh has been created.

- Neural models Flor, Puigcerver, Bluche, and Scheidl, intended for optical recognition of handwritten texts in Kazakh and Russian, have been adapted to the graphemic features of the Kazakh language and trained accordingly.

Reliability and results of approbation. The results of the dissertation research have been published in 5 articles, and a certificate has been obtained for the inclusion of information in the State Register of Rights to Copyright-Protected Objects. The research findings have been presented at scientific seminars of the Department of Computer Science of Al-Farabi Kazakh National University, as well as at international conferences.

Publication in a journal indexed in the Scopus database:

1) Toiganbayeva N., Kasem M, Abdimanap G., Bostanbekov K., Abdallah A., Alimova A., Nurseitov D. KOHTD: Kazakh Offline Handwritten Text Dataset. Signal Processing: Image Communication. Elsevier, 2021, vol.108, pp. 116827. doi https://doi.org/10.1016/j.image.2022.116827 (Scopus: Q2; CiteScore-2.5; Percentile- 73%))

Articles in journals included in the list of scientific publications recommended by the Committee for Quality Assurance in the Field of Science and Higher Education of the Ministry of Science and Higher Education of the Republic of Kazakhstan for publishing the main results of the research work:

- Toiganbayeva N.A., Abdimanap G.S., Alimova A.N., Zhunusova Zh.Kh. Creation and Recognition of a Database of Handwritten Texts in the Kazakh Language. Bulletin of Almaty University of Power Engineering and Telecommunications, 3(62), 2023, pp. 105–116.
- Toiganbayeva N.A., Alimova A.N., Abdimanap G.S., Sakyipbekova M.Zh., Gusmanova F.R. Recognition of Kazakh-Russian Handwritten Texts Based on Neural Networks. Bulletin of Abai Kazakh National Pedagogical University. Series of Physical and Mathematical Sciences, Vol. 84, No. 4 (2023), pp. 183–191.
- Toiganbayeva N.A. Recognition of Kazakh-Russian Offline Handwritten Texts Using Deep Learning Models. Bulletin of the National Engineering Academy of the Republic of Kazakhstan, No. 4 (90), 2023.

International conference:

5) Toiganbayeva N.A., Toiganbayev B.A. A Review of Issues in the Application of Neural Networks for Intelligent Handwriting Recognition Tasks. Satpayev Readings – 2021, Vol. II, pp. 86–89.

Certificate of entry into the state register of rights to copyright-protected objects:

6) Nursseitov D.B., Bostanbekov K.A., Abdimanap G.S., Toiganbayeva N.A., Alimova A.N. «Kazakh Offline Handwritten Text Dataset (KOHTD)» - Certificate of Entry into the State Register of Rights to Copyright-Protected Objects for the Database of Handwritten Texts in the Kazakh Language. No. 34615, 2023.

Researcher's Personal Contribution. In accordance with the objectives set in the dissertation, the corresponding tasks were accomplished. The work included data preprocessing for handwritten text recognition and line segmentation using a genetic algorithm, as well as the creation and population of the KOHTD database. Experiments were conducted with the Flor, Puigcerver, Bluche, and Scheidl models,

and their performance was evaluated using the standard efficiency metrics CER, WER, and SER. All models were implemented using the Python, TensorFlow, Keras, OpenMP, and Matplotlib libraries.

Scope and structure of the work. The dissertation was written in Kazakh and consists of an introduction, four chapters, and a conclusion. The total volume of the work is 97 pages, including 42 figures, 10 tables, and 95 references.

The **introduction** defines the relevance of the work and justifies the need for research in the field of large-scale handwritten text recognition systems. It also outlines the aim, objectives, and methods of the dissertation, presents the scientific novelty of the results obtained, the theoretical and practical significance of the work, and provides a list of publications.

Chapter One presents a review of solutions for optical handwritten text recognition tasks using deep learning capabilities. A quantitative analysis of datasets used in the field of handwritten text recognition is conducted. The chapter provides volumetric, structural, and linguistic characteristics of the most widely used datasets in this field. The analysis serves as a basis for evaluating the efficiency and conducting comparative studies of HTR (Handwritten Text Recognition) systems.

Chapter Two describes the process of creating the first corpus of offline handwritten texts in Kazakh - *Kazakh Offline Handwritten Text Dataset (KOHTD)* – including text segmentation and annotation stages. A statistical analysis of the KOHTD dataset is also presented in this chapter.

Chapter Three presents the main types of neural networks widely used for handwritten text recognition and describes their operating principles. It examines the architectures and capabilities of artificial neural networks, recurrent neural networks (RNN), long short-term memory networks (LSTM), and convolutional neural networks (CNN). The chapter also outlines neural network training methods and explains the role of the Connectionist Temporal Classification (CTC) algorithm in handwriting recognition.

Chapter Four discusses the Gated-CRNN method proposed by Bluche, the 2-LSTM method proposed by Puigcerver, the Gated-CNN method proposed by Flor, and the Attention-Gated-CNN-BGRU method proposed by Abdallah. Furthermore, the models were trained on the KOHTD dataset and evaluated according to the standard performance metrics CER, WER, and SER.

The **conclusion** presents the main results and findings of the dissertation.