Proceedings of IYSC, (2022), vol. 11

Journal homepage: http://journals.sdu.edu.kz/index.php/iysw

2022 International Young Scholars' Conference

CYRILLIC HANDWRITTEN OPTICAL CHARACTER RECOGNITION: A REVIEW OF VARIOUS RECOGNITION METHODS

M. Kalken[1], R. Jantayev[2]

[1,2]*Suleyman Demirel University, Kaskelen, Kazakhstan*

Abstract

In the age of digital technologies, to simplify the search and storage of information, the translation of handwritten documents into electronic format is an urgent task. Optical character recognition makes it possible to recognize characters from images and scans of documents with subsequent translation into a machine-readable format. At the moment, there are a lot of methods and algorithms of machine learning and computer vision that differ from each other in efficiency and method of application. In many ways, the results of the methods used differ due to the specifics of each language under study, expressed in the difference in the number and type of symbols. The purpose of this review article is to summarize the research conducted in recognizing handwritten Cyrillic characters and to conduct comparative results on methods and their results. As the data under study, we summarized and analyzed research articles on the topic of recognition of Cyrillic handwritten text.

Keywords: OCR, Handwritten Text Recognition, Machine Learning, Neural Networks, CNN.

Handwritten text recognition (HTR), due to its increasing significance and the enthusiasm of many scholars, is gaining traction in academic research. In the current digital era, HTR is in high demand in the business sector and is used to convert paper data to digital media, both online and offline[1]. Bank checks, medical paperwork, and postal documents are examples of source documents, while scanned or picture documents in image format are the major source for offline text translation into digital format [2]. All of this necessitates the development of curpnomastable HTR systems that can work with a huge number of documents in many languages. As with any business, there are issues associated with the fact that the qualities of handwritten words vary depending on the author and linguistic peculiarities such as slanted and rounded characters, diacritical marks, transverse stripes, and curved letters.

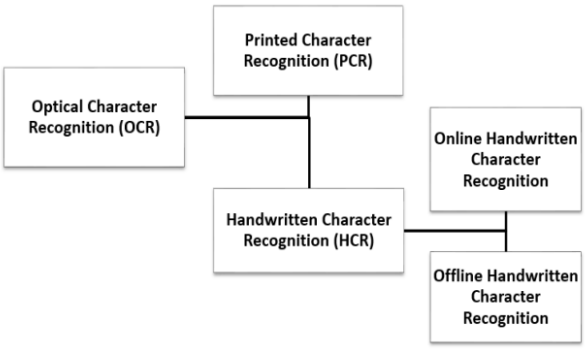

Figure 1. Optical Character Recognition Classification [3].

The success of the handwriting recognition system is determined by the precision with which characters of varying complexity are detected and, therefore, the rate at which the most suitable words are discovered. In recent years, the discipline of deep learning has achieved excellent achievements in the field of optical character recognition, and several approaches have shown to be very successful for a variety of tasks, including picture classification, object identification, and pattern recognition [4]. The study area is handwriting recognition, which has made tremendous progress and has not been abandoned. As a result of the use of Convolutional Recurrent Neural Networks (CRNN) [7]-[9], more effective optical recognition models have been found. These networks use convolutional layers responsible for object extraction from text pictures. Then, the extraction result is fed to repeated layers that propagate and decode objects using Connectionist Time Classification (CTC), which leads to the final results. The CNN contains a Long-term Short-Term Memory (STM) that is often utilized as a sequence decoder. In the future, several new techniques, such as Multidimensional LSTM (MDLSTM), have been developed to enhance the accuracy of this decoder [10], therefore increasing the possibilities of Recurrent Neural Network (RNN) architectures for multidimensional input. Despite the fact that MDLSTM [5, 7] is ineffectual owing to its computing expense and complexity, fresh research [6] propose simpler optical models. One of these models is the Bidirectional Long-Term Short-Term Memory (BLSTM) [10]. This model's findings are comparable to MDLSTM. Although the outcomes of such designs seem to be promising, optical recognition models have trouble remembering lengthy contexts. This is a result of difficulties with disappearing slopes. In addition, these models incorporate millions of trainable parameters to improve performance. Consequently, many real-world applications find it difficult to implement them [10].

In this paper, we propose a study of classification issues in handwritten input images of

Cyrillic letters by analyzing and comparing previous research on the subject. This study is a useful resource for researchers doing study in this area. The article investigates, analyzes, and substantiates many approaches of deep learning, including those used in the area of Cyrillic handwriting recognition. The structure of the document is as follows: In Section 2, the relevant works in the area of handwritten Cyrillic OCR is presented. In order to establish an effective and accurate OCR system, Section 3 outlines and examines the existing position and future developments. Section 4 contains the conclusion.

Literature Survey

Narankhuu Natsagdorj et al. [11] proposed Mongolian handwriting recognition using CNN model LeNet-5 with 7 layers. The authors used a database that contains handwritten characters of the Mongolian language in the form of individual letters. As a result, there are 24 letters that have about 300 repetitions. To get a variety as a result, the authors also changed the size of the image from 32x32 to 20x20. Three different iteration quantities were also used, these are 300, 500 and 1000. As a result, the result estimated in the prediction error showed approximately 2.9% rate of error. The rate in error was calculated for each letter separately. The similarity of letters in writing such as "O" and "B" was emphasized.

Kostiantyn Liepieshov et al. [12] in their work "On recognition of Cyrillic Text" proposed recognition of the handwritten Ukrainian language. Three types were chosen as the database: synthetic database, Cyrillic text in the wild and handwritten text. The method proposed by Jaderberg et al. was chosen as the collection of the first base. where UberText Corpus is supposed to be used. As a second database, the data was collected by the authors from the vastness of the Internet and contains about 500 images. As a handwriting database, the data was collected from the largest mathematical Olympiad in Ukraine and contains about 82,000 images. To test all these databases, the authors used a model that consists of 8 layers of a convolution neural network with different activation functions. After training and testing the model, measurement measures such as WER and CER were selected. For first and second base, the result was 59.8 and 29.6, respectively. For third base, 21.1 and 4.8 respectively.

Nazgul Toiganbaeva et al. [13] in their work proposed the first full-fledged database of the handwritten Kazakh language. Handwritten answers to exam questions of students from Satpayev and Al-Farabi universities were used to collect data. As a result, about 3 thousand images were scanned. Then the stage of processing and preprocessing was carried out, after which the authors

switched to image segmentation to achieve results in the form of separate words in each image. Further, work was done to collect truth tables for each word, for this purpose about 600 volunteers in the form of students and teachers were involved by answering questions through the telegram messenger. As a result, the database consists of about 140 thousand images with 922 thousand characters, of which about 95% of the words of the Kazakh language and 5% of the words of the Russian language. After that, the authors conducted training and testing of recurrent neural network models such as Bluche, Flor, Puigcerver and Abdallah to check the quality of the database. The results were measured in CER, WER and SER. the best result was shown by the Flor model with indicators of 6.52%, 24.52% and 26.98%, respectively.

Daniyar Nurseitov et al. [14] also offered to work on collecting a database of Cyrillic characters, but unlike the previous work, the database mainly consists of the names of cities and words of the Russian language. The database was collected by drawing up forms to fill out and about 200 volunteers were involved. Each of them filled out the form 4-5 times, as a result, about 1400 forms were collected. As a result, the forms contain about 106 thousand words and about 715 thousand characters. After this procedure, the authors carried out labeling and segmentation with further collection of truth tables.

Ruslan Jantayev et al. [15] proposed works on recognizing entire pages of handwritten Kazakh using the start, follow and read method. To achieve this goal, the authors also used their own database collected by scanning handwritten entries of university students. About 40 students were involved and about 400 pages of handwriting were collected. As a result, the number of words amounted to about 70,000 Kazakh handwritten words. The base was divided into 90% and 10% for training and testing, respectively. Since recognition is performed using a full page, the authors used the method of finding the beginning of the text, moving through the text and recognizing the text at the same time. The cut-off architecture of VGG-19 began as a start method. Further, the CNN model with 7 layers was used to follow method. The CNN-LSTM HWR architecture was used to recognize the text as read method. To test the model, the test part was divided into two parts, each of which, by CER standards, showed a result of 11% and 13%, respectively.

Abdelrahman Saleh et al. [16] proposed handwriting recognition of the Russian language using the Attention-Gated-CNN-BGRU model. The model has a deep beam structure unlike many well-known models. The database described earlier in the previous work, the HKR database, was

chosen as the database. It was divided into three parts, the first for training the model, the second and the third for testing. After completion of training and testing, the model results in the form of CER, WAR and STAR in 4.13%, 18.91% and 25.72%. Which is superior in results to other models like Bluche. Also, to compare and verify the proposed model, a test was conducted using other well-known databases such as IAM, Saintgall, Washington and Bentham. As a result, the model surpassed many other well-known models in the ratio of errors per character and word.

Daniyar Nurseitov et al. [17] proposed a comparison of different models using databases with the names of cities. Deep CNN, HTR, Bluche and Puigcerver were chosen as models. The HKR database was also taken as a database. As a result, the HTR model showed a result in the form of CER 15.78%, the Bluche model in the form of 10.15% and the puigcerver model in the form of 54.75%.

Discussion

Current research indicates that the recognition of handwritten Cyrillic text is a highly relevant topic, as databases containing handwritten text are just now being brought up to certain standards. Numerous works provide their own databases and do tests with them. In addition, the effectiveness and efficiency of recognition depend not only on the quality of the collected database, but also on the neural network model used. Varieties of well-known convolution networks are utilized in the majority of each study. For a more precise characterization and comparison of the work's outcomes, the following table takes into consideration models, databases, and test outcomes.

Table 1

Datasets and results of training different models of the studied literature

Reference	Dataset used	Architecture	CER	WER	Specificity
Narankhuu Natsagdorj et al.	Mongolian handwritten characters (7,200)	LeNet-5	2.9%	-	The database contains only individual letters.
Kostiantyn Liepieshov et al.	Ukrainian Cyrillic text in the wild (505), handwritten Cyrillic text (82,061)	7 layers of CNN	27.6% 4.8%	59.8% 21.1%	Used for the development of the new system for processing responses of participants of one of the largest math competitions in Ukraine.
Nazgul Toiganbaeva et al	Kazakh offline handwritten text dataset (140,335)	Flor Puigcerver Abdallah Bluche	6.52% 8.01% 8.22% 8.36%	24.52% 26.34% 22.60% 28.95%	A very large database of 95% of the Kazakh language has been collected and used.
Daniyar Nurseitov et al.	Handwritten Kazakh Russian (64,943)	-	-	-	In the work, the main goal was to collect a database without further testing.
Ruslan Jantayev et al.	Kazakh handwritten dataset (70,000)	Start, follow and read method	11% 13%	-	The model reads and digitizes the full page.
Abdelrahman Saleh et al.	HKR (64,943)	Attention-Gated-CNN-BGRU	4.13% 6.31%	18.91% 23.69%	A deep model with many parameters is used.
Daniyar Nurseitov et al.	HKR (64,943)	Deep CNN HTR Bluche Puigcerver	- 15.78% 10.15% 54.75%	- 25.89% 37.49% 82.91%	Different models with different structures were used by comparing the results.

Table 1 illustrates the relationship between the various databases and models utilized to produce recognition results. And currently, according to the work of Abdelrahman Saleh et al. utilizing the HKR database and the Attention-Gated-CNN-BGRU model, has the best outcome indicator in the form of 4.13 percent CER and 18.91 percent WER. Each work has its own particulars and distinctions. With these accomplishments, we can observe the growing interest of scholars in this field and the continuous improvement of handwritten Cyrillic character recognition.

Conclusion

The recognition of handwritten Cyrillic text is a highly relevant topic, as databases containing handwritten text are just now being brought up to certain standards. The effectiveness and efficiency of recognition depend not only on the quality of the collected database, but also on the neural network model used. This survey reveals that the majority of researchers recommend using CNN. CNN-based models are well suited for image identification tasks and particularly outperform all other models. Additionally, the accuracy rate depends not only on CNN architecture, but also on the dataset and pre-processing techniques used. The subsequent phase would include the development of a highly accurate automated system. After developing an accurate OCR, it will assist not only the general populace but also the financial and government sectors of CIS nations. A significant effort was made to promote the Russian and Kazakh languages by collecting huge, publicly accessible databases. This will play a significant role in the promotion of machine learning models to enhance recognition outcomes. Consequently, it is now necessary to gather and standardize databases of other Cyrillic languages.

References

1. R. Reeve Ingle. et. al., A Scalable Handwritten Text Recognition System, IEEE Trans. Pattern Anal. Mountain View, CA 94043, USA, 2019.

2. Tsochatzidis L, Symeonidis S, Papazoglou A, Pratikakis I. HTR for Greek Historical Handwritten Documents. J Imaging. 2021.

3. Phillip Benjamin Ströbel. et. al., Fink G.A., Evaluation of HTR models without Ground Truth Material, University of Zurich, University of Bern), 2022.

4. Chieh-Chi Kao. et. al., R-CRNN: Region-based Convolutional Recurrent Neural Network for Audio Event Detection, Amazon Alexa, 2019.

5. Xinyu Fu et. al., CRNN: A Joint Neural Network for Redundancy Detection, Solution Architect and Engineering Asia Pacfic and Japan, Nvidia, 2017.

6. Baoguang Shi et al, An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Huazhong University of Science and Technology, 2015.

7. Alex Sherstinsky et al. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network, Physica D: Nonlinear Phenomena, 2018.

8. Has¸im Sak, et. al. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling, Google, USA, 2018.

9. Haowei Jiang et. al., Recurrent Neural Network from Adder's Perspective: Carry-lookahead RNN. Whiting School of Engineering, 2021.

10. Ralf C. Staudemeyer et. al., Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks Last accessed November, Schmalkalden University of Applied Sciences, 2019.

11. Narankhuu Natsagdorj, Mongolian Handwriting Character Recognition Based On Convolutional Neural Network(CNN), 6th International Conference on Information Engineering for Mechanics and Materials, 2016.

12. Kostiantyn Liepieshov, Oles Dobosevych, On recognition of Cyrillic Text, 2019.

13. Nazgul Toiganbayevaa, Mahmoud Kasemc, Galymzhan Abdimanap, et al. KOHTD: Kazakh Offline Handwritten Text Dataset, 2021.

14. Daniyar Nurseitov, Kairat Bostanbekov, Daniyar Kurmankhojayev, Anel Alimova, Abdelrahman Abdallah, HKR For Handwritten Kazakh & Russian Database, Multimedia Tools and Applications, 2021.

15. Ruslan Jantayev, Shirali Kadyrov, Yedilkhan Amirgaliyev, Complete Kazakh Handwritten Page Recognition Using Start, Follow And Read method, Journal of Theoretical and Applied Information Technology, 2021.

16. Abdelrahman Abdallah, Mohamed Hamada and Daniyar Nurseitov, Attention-Based Fully Gated CNN-BGRU for Russian Handwritten Text, Journal of Imaging, 2020

17. Daniyar Nurseitov, Kairat Bostanbekov, Maksat Kanatov, Anel Alimova, Abdelrahman Abdallah, Galymzhan Abdimanap, Classification Of Handwritten Names Of Cities And Handwritten Text Recognition Using Various Deep Learning Models, Advances in Science, Technology and Engineering Systems Journal, 2020.